No. 20-619

IN THE SUPREME COURT OF THE UNITED STATES

◆

A.S. a 9-year old child with Autism Spectrum Disorder (ASD) entitled to Special Education and Related services per IDEA represented by his parents R.S. *Pro se* and E.S. *Pro se*

Plaintiffs-Petitioners

-v.-

Board of Education Shenendehowa Central School District,

Interim Commissioner Betty Rosa, of The University of the State of New York

Defendants-Respondents

◆

Petition for Rehearing
On Writ of Certiorari
To the U.S. Court of Appeals for the 2nd Circuit

◆

PETITION FOR REHEARING

ON SUBSTANTIAL GROUNDS NOT PREVIOUSLY PRESENTED

Petition for Rehearing for the denial of Petition of Writ of Certiorari appealing the Decision, Order and Judgment of The United States Court of Appeals of the Second Circuit by Judges Pierre N. Leval, Raymond J. Lohier, Jr. and Michael H. Park to dismiss for lack of jurisdiction the Appeal from the Memorandum-Decision and Order and Judgment of The United States District Court for the Northern District of New York by Judge Lawrence E. Khan entered February 20, 2019 and Motion to Reopen Granted on March 16, 2020 and postmarked on March 16, 2020 where FRAP suggests 14-day timeline begins on March 19, 2020 in Action No. 20-1153.

NEW QUESTIONS PRESENTED IN THIS PETITION FOR REHEARING

1. Is the Supreme Court Aware that an Autism Gene Therapy clinical trial is likely less than 5 years away as a result of the Advent of CRISPR/Cas9 based Gene Editing Technologies such as Base Editing and Prime Editing?

2. Is the Supreme Court Aware that when an Autism Gene Therapy Clinical Trial Commences it will have to hold itself to the same standard that the Lovaas UCLA Early Autism Program (Lovaas, O. I., 1987, Journal of Consulting and Clinical Psychology, 55:3-9) and High Fidelity Replications (Cohen, H., Amerine-Dickens, M., & Smith, T., 2006, Developmental and Behavioral Pediatrics, 27:S145-S155; Howard, J. S., Stanislaw, et al., 2014, Research in Developmental Disabilities.

35:3326-3344; Sallows, G. O., & Graupner, T. D., 2005, *AJMR*, 110:417-438) held themselves to because those standards most closely parallel one's ability to achieve "further education, employment and independent living". 20 U.S.C. § 1400(d)(1)(A) That is an achievement of IQ in the normal range, achievement if Vineland Adaptive Behavior Scales (VABS) Composite Score in the normal range that would be expected to be followed by a normal classroom placement? Thus, Autism Gene Therapy will need an intensive ABA framework in place to ensure that the outcomes of any program can be attributed to that program and not an IBI or intensive ABA program completed in parallel?

3. Is the Supreme Court Aware that if a viable and national effort to use proven approaches to autism is not in place that this country will soon (in about 2 decades) be stuck in permanent or long-term recession—removing our position as the world's leading power—as a result of the increasing incidence of Autism?

QUESTIONS PRESENTED IN ORIGINAL PETITION FOR WRIT OF CERTIORARI

1. Whether an appellate court may *sue sponte* dismiss an appeal which has been filed within the time limitations stated in the Federal Rules of Appellate Procedure FRAP Rule 26(c) that adds 3 days for service by mail to file an appeal for which the motion has been granted to

reopen the time to file an appeal under rule 4(a)(6) of FRAP?

2. Whether non-attorney *pro se* parents can reasonably have been expected to know of unwritten rules that lawyers take for granted that FRAP Rule 26(c) does not apply to mailed motions that are granted to reopen the time to file an appeal under rule 4(a)(6) of FRAP when that is impossible to determine when reading the Federal Rules of Appellate Procedure?

3. Whether the interpretation of FRAP is intended to be based on the stand-alone document and whether supplementary rules are required for its interpretation where such supplementary rules are referenced within FRAP to the particular application of FRAP rule 26(c) on FRAP rule 4(a)(6)?

4. Is Intensive Behavioral Intervention or its equivalent intensive Applied Behavior Analysis (ABA) required for a specific period of time for a child with autism in order for the IEP to be "reasonably calculated" for the child to make progress in light of their circumstance?

5. In light of question 4, is there any other way to raise measures by "technically sound instruments that may assess the relative contribution of cognitive and behavioral factors, in addition to physical or developmental factors." (20 U.S.C. § 1414

(b)(2)(C); 8 N.Y.C.R.R. § 200.6(6)(ii)(x)) such as IQ and Vineland Adaptive Behavior Scales (VABS) such that "further education, employment and independent living" 20 U.S.C. § 1400(d)(1)(A) is a reasonable expectation for at least half of all school aged children with autism?

6. Can a court defer to the opinion of a lower judicial body when there is an alleged bias of that lower judicial body?

7. Are the rules, regulations and laws of 8 N.Y.C.R.R. §200 et seq. and also The IDEA 20 U.S.C. §§ 1400-1482 especially as it relates to persons with autism written so that they are unconstitutionally vague and such that they cause confusion and variation in opinion in the courts, absent expensive expert testimony, and unlawfully empower school personnel, schools, school districts other Local Education Agencies (LEAs) to broadly interpret the education law themselves especially on such pertinent matters of Least Restrictive Environment (LRE) determinations and the appropriateness of a particular educational approach such that it permits the curtailing of the rights of students receiving special education and their parents and consistently results in a denial of a FAPE, a denial of access to the students LRE to the maximum extent appropriate and also results in confusion amongst the appellate courts on how to interpret the education law and render a judgment?

8. Given the nature of the common developmental delays found in nearly all autism spectrum disorder (ASD) diagnoses, if a student with a an ASD entitled to an Individualized Education Plan (IEP) and special education and related services should the three measures of 1) expressive language, 2) conversational ability (measured in the number of peer aged exchanges that a student can consistently demonstrate) with typically developing peers if in their LRE and 3) a reduction in prompt dependence be guaranteed goals on the student's IEP since these measures are necessary to the purpose of The Individuals with Disabilities Education Act (The IDEA) (20 U.S.C. §§ 1400-1482) which is "to ensure that students with disabilities have available to them a FAPE in the LRE to the maximum extent appropriate that emphasizes special education and related services designed to meet their unique needs and prepare them for further education, employment, and independent living" (20 U.S.C. §§ 1400(d)(1)(A))?

9. If Question 8 (corrected) is not answered in the affirmative does 20 U.S.C. §§ 1400(d)(1)(A)) have any meaning for a child with autism?

TABLE OF CONTENTS

Appendix 2, McEachin, J.J., Smith, T., Lovaas, O.I. (1993) Long-term outcome for children with autism who received early intensive behavioral treatment. AJMR, 97, 4, 359 – 372.

Appendix 3, Sallows, G. O., & Graupner, T. D. (2005). Intensive behavioral treatment for children with autism: Four-year outcome and predictors. AJMR, 110, 417–438.

Appendix 4, Cohen, H., Amerine-Dickens, M., & Smith, T. (2006). Early intensive behavioral treatment: Replication of the UCLA model in a community setting. Developmental and Behavioral Pediatrics, 27, S145–S155.

Appendix 5, Howard, J. S., Stanislaw, H., Green, G., Sparkman, C. R., & Cohen, H. G. (2014). Comparison of behavior analytic and eclectic early interventions for young children with autism after three years. Research in Developmental Disabilities, 35 (12), 3326-3344.

Appendix 6, Thusberg, J., Olatubosun, A. & Vihinen, M. Performance of mutation pathogenicity prediction methods on missense variants. Hum Mutat. 2011 32 (4):358-68.

Appendix 7, Gerasimavicius, L., Liu, X. & Marsh, J.A. Identification of pathogenic missense mutations using protein stability predictors. 2020, Sci Rep 10, 15387

TABLE OF AUTHORITIES FOR PETITION FOR REHEARING

STATUTES, RULES AND REGULATIONS

20 U.S.C. § 1400(d)(1)(A)...................................iii

GOVERNMENT AND NONPROFIT SOURCES

Mindspec: Informatics for Autism and Neurodevelopmental Disorders
Web Address: http://www.mindspec.org/
Specific Referenced Webpage Within: Autism Informatics Portal
Web Address:
http://autism.mindspec.org/autdb/submitsearch?selfld_0=GENES_GENE_SYMBOL&selfldv_0=&numOfFields=1&userAction=viewall&tableName=AUT_HG&submit2=View+All..1

Center for Disease Control and Prevention: Data & Statistics on Autism Spectrum Disorder
Web Address:
https://www.cdc.gov/ncbddd/autism/data.html........10

Autism Speaks: Autism Statistics and Facts
Web Address:
https://www.autismspeaks.org/autism-statistics.....11

VIDEO PRESENTATIONS

PUBLICATIONS

methods on missense variants. Hum Mutat. 2011 32 (4):358-68..2

Gerasimavicius, L., Liu, X. & Marsh, J.A. Identification of pathogenic missense mutations using protein stability predictors. 2020, Sci Rep 10, 15387..2

AUTHORITIES ORIGNALLY USED FOR PETITION FOR A WRIT OF CERTIORARI

CASES

(Application of a Student with a Disability, NYSED SRO Decision No. 17-008)

A.M. v. New York City Dep't of Educ., 845 F.3d 523, 541-45 (2d Cir. 2017)

Amanda J. ex rel. Annette J. v. Clark County School Dist. 267 F.3d 877 (9th Cir. 2001)

Bd. of Educ. v. Rowley, 458 U.S. 176 (1982)

Deal v. Hamilton County Board of Educ., 392 F.3d 840 (6th Cir. 2004)

T.*R. v. Kingwood Twp. Bd. of Educ., 205 F.3d 572, 577 (3d Cir. 2000)*

L.G. ex rel. E.G. v. Fair Lawn Bd. of Educ. 486 Fed. Appx. 967 (3rd Cir. 2012)

L.B. and J.B. on behalf of K.B. v. Nebo Sch. Dist., 379 F.3d 966 (10th Cir. 2004)

Madison Board of Education v. S. V. 2020 WL 5055149 (U.S. NJ 2020)

R.E.B. v. State of Hawaii Department of Education 870 F.3d 1025 (9d Cir. 2017)

R.E.B. v. State of Hawaii Department of Education 55770 Fed. Appx. 796 (9d Cir. 2019)

Renee J. v. Houston Independent School District 333 F.Supp.3d 674 (S.D.TX.)

R.S. v. Bd. of Educ. Shenendehowa Cent. Sch. Dist., 1:17-CV-0501 (LEK/CFH) (N.D.N.Y. Feb. 20, 2019)

Sumter County School Dist. 17 v. Heffernan ex rel. TH 642 F.3d 478 (4th Cir. 2011)

T.H. v. Board of Education of Palatine, 55 F.Supp.2d 830 (N.D. Ill. 1999)

Winkelman v. Parma City School District, 127 S. Ct. 1994 (2007)

Wittenberg v. Winston–Salem/Forsyth County Board of Education, 2008 WL 11189389 (M.D.N.C. 2008)

Z.F. v. South Harrison Community School Corp. 2005 WL 2373729 (S.D. IA. 2005)

STATUTES, RULES AND REGULATIONS

FOURTEENTH AMENDMENT SECTION II

"Void for Vagueness" Doctrine of the U.S. Constitution

IDEA

Daubert Standard

Education of the Handicapped Act Amendments of 1990, Pub. L. No. 101-476, 104 Stat. 1103 (1990)

Individuals with Disabilities Education Act Amendments of 1997, Pub. L. No. 105–17, 111 Stat. 37 (1997)

Individuals with Disabilities Education Act Amendments of Pub. L. No. 108-446, 118 Stat. 2647 (2004)

20 U.S.C. §§ 1400-1482 et seq)

20 U.S.C. § 1400(c)(1)

20 U.S.C. § 1400(c)(1) (2000 & Supp. IV 2004)

20 U.S.C. § 1400(d)(1)(A)

20 U.S.C. § 1400(d)(1)(A-B)

20 U.S.C. § 1412(a)(5)(A)

20 U.S.C. § 1414(b)(2)(C)

20 U.S.C. § 1414(d)

20 U.S.C. § 1414(d)(1)(A)(i)(IV)

28 U.S.C. § 1254

8 N.Y.C.R.R. § 200 et seq.

8 N.Y.C.R.R. § 200.4(d)(2)(v)(b)

8 N.Y.C.R.R. § 200.6(6)(ii)(x)

Fed. R. App. P. 4(a)(6)

Fed. R. App. P. 26(c)

LEGISLATIVE MATERIALS

S. Rep. No. 94–168 (1975), as reprinted in 1975 U.S.C.C.A.N. 1425

Cong. Rec. 19492 (1975)

PUBLICATIONS

Cohen, H., Amerine-Dickens, M., & Smith, T. (2006). *Early intensive behavioral treatment: Replication of the UCLA model in a community setting.* Developmental and Behavioral Pediatrics, 27, S145–S155

Howard, J. S., Stanislaw, H., Green, G., Sparkman, C. R., & Cohen, H. G. (2014). Comparison of behavior analytic and eclectic early interventions for young children with autism after three years. *Research in Developmental Disabilities,* 35 (12), 3326 - 3344

Howard, J. S., Sparkman, C. R., Cohen, H. G., Green, G., & Stanislaw, H. (2005). A comparison of intensive behavior analytic and eclectic treatments for young children with autism. *Research in Developmental Disabilities*, 26, 359–383

Koegel, R. L., Werner, G. A., Vismara, L. A., & Koegel, L. K. (2005). The effectiveness of contextually supported play date interactions between children with autism and typically developing peers. *Research and Practice for Persons with Severe Disabilities*, 30, 93–102

Lee, P. F., Thomas, R. E., & Lee, P. A. (2015). Approach to autism spectrum disorder: Using the new DSM-V diagnostic criteria and the Can MEDS-FM framework. Canadian family physician Medecin de famille canadien, 61(5), 421–424

Lovaas, O. I. (1987). Behavioral treatment and normal educational and intellectual functioning in young autistic children. *Journal of Consulting and Clinical Psychology, 55*(1), 3–9

McEachin, J.J., Smith, T., Lovaas, O.I. (1993) Long-term outcome for children with autism who received early intensive behavioral treatment. *AJMR*. 97, 359–372

Sallows, G. O., & Graupner, T. D. (2005). Intensive behavioral treatment for children with autism: Four-year outcome and predictors. *AJMR*, 110, 417–438

REASONS FOR GRANTING THE PETITION FOR REHEARING

I. THE FUTURE OF AUTISM GENE THERAPY MAY DEPEND ON THIS PETITION.

The advent of CRISPR/Cas9 based gene therapies will soon, hopefully within 5 years, enable researchers to pursue Autism Gene Therapy and Gene Therapies for related neurodevelopmental disorders. Although, the picture is complex for Autism. For example, there is nearly 1100 genes associated with autism. http://autism.mindspec.org/autdb/submitsearch?selfld_0=GENES_GENE_SYMBOL&selfldv_0=&numOfFields=1&userAction=viewall&tableName=AUT_HG&submit2=View+All (Autism Informatics Portal) For any given gene there is a large number of potential autism causing mutations where causality is not always easy to establish. There may be a mutation of a gene with possible cause of autism and in some instances the prediction that the mutation causes autism is nearly certain and in other instances not. Causal links may be easily made with nonsense mutations or protein truncating variants (that reduces the length of the protein) and frameshift (that changes virtually every amino acid—in comparison to the natural functioning form—that follows where the location of the frameshift occurs) mutations both that materially change the protein and impair its function and also often in cases involving in-frame deletions (the loss of amino acids) or insertions (the addition of amino acids). Missense mutations that change a single amino acid, can be more difficult to create a causal link to autism. Structural biology combined with Statistics and computational science including machine learning has made it possible to predict the likelihood that a missense mutation would affect protein function. (Thusberg, J., Olatubosun, A., Vihinen, M. *Hum Mutat, 2011* *32*:359-68.; Gerasimavicius, L., Liu, X. & Marsh, J.A. *Sci Rep, 2020* 10:15387) Even with these tools, in most cases of autism scientists can only predict with low likelihood that a specific missense mutation was the cause of the autism. In tens of thousands of instances there is 3 or less documented cases of autism for one specific mutation that is often a missense mutation, but multiple different mutations on the same gene.

Autism while being a spectrum disorder also has a separate spectrum for each of the 1100 genes. A single gene can have a mutation in one of a number of places. The degree that the mutation impairs the protein's function determines the severity of the autism within the gene's spectrum. Additionally, individuals that express less protein than average will be more greatly impacted from the mutation of one of typically 2[1] functioning genes, a term referred to as haploinsufficiency[2]. With exceptions to X and Y chromosomes in males where generally there is one functioning gene. Additionally, factors such as multiple functional domains on a single protein can also contribute to the broad spectrum. This web of complexity, that is the broadness of the autism spectrum for any particular gene creates an ethical

[1] Especially, when the protein encoded for by the gene plays a more essential function.

[2] In haploinsufficiency one of two copies of a gene is sufficiently nonfunctional such that there is an observable difference in the individual.

dilemma of correcting a supposed autism causing mutation before ruling out that the individual can achieve typically levels of IQ and Vineland Adaptive Behavior Scales (VABS) from intensive behavioral intervention (IBI) or intensive Applied Behavior Analysis (ABA). Why should an individual that can achieve typical levels of IQ and VABS and thus indistinguishable from their typically developing peers be subjected to gene therapy in its early stages if they can achieve the intended outcome without it?

If this court finds that an IEP that does not include IBI or intensive ABA for persons with autism cannot be "reasonably calculated" to confer educational benefit making IBI or intensive ABA a matter of right for persons with autism, then any aspect of the autism gene therapy ethical dilemma that relates to ruling out the individual can achieve typical levels from IBI or intensive ABA is in principle resolved.

II. THE ABILITY OF AUTISM GENE THERAPY TO SERVE THE GREATEST GOOD LIKELY REQUIRES THAT INTENSIVE ABA METHODOLOGY BE AN EDUCATIONAL RIGHT TO PERSONS WITH AUTISM FOR 2 TO 3 CONSECUTIVE UNINTERRPUTED YEARS.

The ethical dilemma is further complicated because it would not be ethically correct to limit a program to the mutations that leaves persons worst off—as such instances have a host of challenges that reduces the likelihood of their success in the early stages of autism gene therapy—and thus that does not serve the greatest good. One might argue that individuals with a particular autism causing gene are profoundly affected and they would clearly not recover without gene therapy. However, a program only on those individuals is less likely to succeed in a relatively short time window for reasons discussed above and thus would not serve the greatest good. What serves the greatest good early on in a program is to initially commence gene therapy those that can reasonably be expected to achieve typical levels of IQ and VABS—in a relatively short time window—with a successful correction of the autism causing gene but cannot do so without such a gene therapy program. This is consistent with a prevailing view on a limited professional staffing scenario[3] (see Tristram Smith Keynote Presentation on Evidence-Based Practices for Children with ASD. https://www.youtube.com/watch?v=tQ2fA32ysZQ (5/30/2014) at 1:44:30 – 1:45:23) in IBI or intensive ABA. We have considered the principles behind them and we agree with it. We also envision that autism gene therapy clinical trials would take place across all correctable mutations on a given gene including those less likely to quickly recover from gene therapy where there is a sizable percentage (40+%) of individuals with the mutation that could be expected to recover within 2 years from a successful correction of the underlying gene, while unable to do so solely from IBI or intensive ABA. This gene targeting has the added benefit where those less likely to quickly recover from

[3] In this scenario there is not enough trained personnel to provide intensive ABA or IBI to all persons with autism so those that are projected to benefit most are given priority over those that are expected to minimally benefit who instead receive the typical program offered by the school district.

gene therapy could participate in a gene therapy program while keeping with the principle of initially targeting those for gene therapy that can reasonably be expected to recover from gene therapy in a relatively short time window.

As Autism Gene Therapy early on will be based on a limited resource model due to limited initial investment whose early and rapid success determines the amount and speed with which further funds will be invested into such an industry. In other words, the early success of autism gene therapy in recovering persons with autism to typically developing levels will mean that an increasing amount of funds will be poured into the industry in a rather short time window thereby increasing the number of people recovered from autism as a function of time. These substantial grounds not previously presented further establishes the Supreme Court's Role in granting the Petition for a Writ of Certiorari. If the Supreme Court can within its jurisdiction hear a case whose outcome can have far reaching implications that benefit practically all members of humanity, then The Court ought to hear this case. As the greatest good is served for both those than can achieve typical levels from intensive ABA and those that will require gene therapy to do so. No person with autism will be left behind.

Establishing that a mutation cannot be corrected with IBI requires that school aged persons with autism have IBI or intensive ABA as a matter of right for 3 years followed by 2—3 years part time transitional ABA. This decision falls withing the jurisdiction of

The Supreme Court as it is the basis of the Petition and the U.S. Courts of Appeals are scattered in their treatment on. **(see Petition for Writ of Certiorari 20-619)** This Court can imagine a situation where a child receives Autism Gene Therapy in the future but in some instances the child has access to IBI or intensive ABA and in other cases not. Further, there is wide programming variation across IEPs written to support unproven eclectic intervention programs. This creates an efficacy nightmare because one may not be able to determine the cause of the improvement in IQ and VABS. Was it the child's intervention program or was it the gene therapy that brought about the improvement to typical levels?

Similarly, there is a challenge associated with the placement being a source that negatively impacts the potential gains associated with a gene therapy program. Did the environment itself e.g. the highly restrictive placement that includes settings that do not provide access to model typically developing peers bring about the less than desired outcome from a gene therapy program? The environment lacking model peers for even part of the day can equally be the source that negatively impacts the IQ and VABS outcome. If one typically developing child is by themselves placed in an educational placement that only included peers with autism would they be expected to develop normally? The answer is almost certainly no! So then if a child with autism with a self-contained placement receives autism gene therapy how are they expected to recover? This matter can be decided on by The Supreme Court because the very foundations of IBI or intensive ABA support an intervention model that takes measures to avoid detrimental self-contained placements.

It has been well established that the placement itself can make an otherwise perfectly designed intervention program ineffective. Lovaas noted this that in a setting where all intervention was provided in a self-contained environment led to results that did not allow persons with autism to recover, while the identical program in environments that did not include others with autism led to 47% achievement of typically developing levels that were sustained by all but one participant. (Smith, T., Lovaas, O.I., 1993, *AJMR. 97*:359-372)

Not hearing this case will potentially stifle an Autism Gene Therapy program that could correct the autism for those persons that cannot sufficiently benefit from IBI to reach typical levels, a number that is about 53% of persons with autism. It has been 34 years since Lovaas had reported on IBI (Lovaas, 1987). All replications and efforts to improve upon the results have not been able to improve the outcome. Across all studies ever reported no program has achieved results that exceed the Lovaas Program or its Replications (Sallows, 2005; Cohen, 2006; Howard, 2014). As further support, in the Wisconsin Early Autism Program Sallows and Graupner (Sallows, 2005) noted two types of participants in their Lovaas Program Replication, Rapid Learners and Moderate Learners. Interestingly, they cannot reliably[4] be distinguished

[4] Stronger social engagement skills at program onset were correlated with better outcomes.

from each other at program onset as their starting points are similar. *See Table 3* (pp. 426, Sallows, 2005) (Pet. Reh. App. 31). Rapid Learners average Intake IQ: 55.3 and VABS: 61.73. Moderate Learners average Intake IQ: 47.8 and VABS: 58.7. However, after the follow-up the Rapid Learners achieved a mean IQ of 103.73 and a mean VABS of 88.6 while Moderate Learners achieved a mean IQ of 50.4 and VABS of 49.1[5]. From these results is becomes readily apparent who would be potential candidates for autism gene therapy following 3 years of IBI or intensive ABA. Further, it has also been shown that eclectic intervention programs or special education as usual do not allow one to distinguish Rapid Learners for Moderate Learners except for the top 10%—20% of Rapid Learners. *See Table 3 of* (pp. 7, Lovaas, 1987) see (Pet. Reh. App. 5). Thus, 80%—90% of Rapid Learners that would not need autism gene therapy to recover from autism cannot be identified from special education as usual.

We explained to this court that in a country that spends $250 to $300 Billion a year on autism finding that an IEP that does not specify intensive ABA methodology cannot be reasonably calculated to confer educational benefit for persons with autism would result in savings of $100 Billion annually in the long term. Autism gene therapy will likely be effective on both adults and children. But adults that receive autism gene therapy will have an entirely new challenge, closing the developmental gap, that is more

[5] It should be noted that there is a broad spectrum of the moderate learners. There were many moderate learners that saw gains in IQ and VABS. But not to typical levels.

difficult to close with increasing age, and finding a way to fit into society. This may be easily surmounted for those in financially prominent families while those in families that have quite limited financial resources will find this challenge to be significant. There is also the complicated question of the underlying psychology after recovering from autism in adulthood. Thus, autism gene therapy would obviously be preferred to be completed in childhood.

III. PRIME EDITING AND BASE EDITING TECHNOLOGY PLATFORMS HAVE BEEN DEVELOPED MAKING AUTISM GENE THERAPY POSSIBLE AND THERE IS AN UNRELATED FDA APPROVED GENE THERAPY TARGETING THE BRAIN

In 2019 the FDA approved Novartis's gene therapy Zolgensma. https://www.zolgensma.com/what-is-zolgensma. Zolgensma works by delivering episomal DNA—that does not integrate into the host genomic DNA—to the brain that makes a new copy of a gene known as human Survival Motor Neuron 1 (SMN1). Persons with spinal muscular atrophy typically have two nonfunctioning copies—referred to as autosomal recessive—of the SMN1 gene.

Most cases of autism are due to haploinsufficiency. Because thought is more fine-tuned than almost any other function in the body one can imagine that the amount of protein that is needed to be expressed is based on a number of factors that our molecular machinery must be sufficiently sensitive to detect when more protein is required. There are a host of other elements in the genome that are not part of the gene that can be activated to express more of the gene in the cell when more is required or to stop expression when there is a sufficient surplus. Thus, CRISPR/Cas9 (Clustered Regularly Interspaced Short Palindromic Repeats/CRISPR associated Protein 9) based gene therapy platforms such as Prime Editing and Base Editing technologies may be necessary since these technologies correct mutations in the genomic DNA.

IV. THE INCIDENCE OF AUTISM IS INCREASING AND NOT FINDING THAT INTENSIVE ABA METHODOLOGY IS AN EDUCATIONAL RIGHT TO PERSONS WITH AUTISM WILL PUT THIS COUNTRY ON A COURSE TO PERMANENT OR LONG-TERM RECESSION.

Many wonder if the increase in the incidence of autism (now 1 in 54 births) is real. https://www.cdc.gov/ncbddd/autism/data.html If it is real, then it only means that it is going to get progressively worse. Why might this be? Well, there are a tremendous number of genes associated with the functioning of the nervous system. A benign mutation in one protein that functions in the nervous system may not be noticeable. However, multiple benign mutations may be noticeable. It is manifested by so many mutations of genes associated with brain function per generation. Because the mutations associated with autism are mostly completely random and the prevalence is increasing it raises a question as to whether the incidence of autism could suddenly take off in the next two generations. This may seem like a long time. Although, because there is such an extensive number of autism causing mutations the window of investigation and correction of every pathological autism causing mutation could take 30 years. So, if this matter is delayed much longer the consequences for mankind will be catastrophic.

Not hearing this case will send this country on a course into certain bankruptcy that will begin within 20 years because once the incidence becomes too far out of control the financial impact would be devastating. Gene therapy may not be able to bridge the gap quickly enough to avoid economic devastation without the framework being in place soon. Once things fall outside of a specific parameter chaos results. The butterfly effect parameter is probably somewhere in the area of an incidence of autism that is equal to a half the rate of unemployment that has economic consequences, such as recession, about 10% unemployment. Thus, an incidence of autism of 1 in 20 births or 1 in 10 families will result in a recession like situation. That is because a person with autism generally has one of the parents as the case manager https://www.autismspeaks.org/autism-statistics and makes it very difficult for that parent to hold a full-time job while meeting the needs of their son or daughter with autism. Consider the following: If the reported unemployment rate is 2.5% (which is good economic conditions) the incidence of autism is 5% causing 1 in 20 parents to not be able to maintain competitive nor full time employment and where a intervention provider will have to provide support to persons with autism at an effort equivalent to 50% of full time one can imagine a situation when 5%

incidence of autism results in 5% unemployment for parents on top of a 2.5% unemployment rate and 2.5% of employment to support persons with autism. That is effectively equivalent to 10% unemployment. We are likely less than one generation away from an autism incidence of 1 in 20 births.

It is also important to point out that if the economics of being a BCBA provider do not sufficiently improve or if baseless policies are put in place that reduce their numbers the manpower may not be in place in the future to provide IBI. We have a situation now that functions as the ideal situation. Sufficient manpower from BCBA providers and unprecedented advances in Prime Editing and Base Editing gene therapy technologies. For this reason, the Supreme Court must act now to hear this matter that falls within the jurisdiction of the court so that all persons with autism, their siblings and parents can benefit.

In Conclusion, this Petition for Rehearing and Writ of Certiorari Should be Granted!

Respectfully Submitted on February 5, 2021.

____

R.S. *Pro Se* on behalf of A.S.

____

E.S. *Pro Se* on behalf of A.S.

No. 20-619

IN THE

SUPREME COURT OF THE UNITED STATES

A.S. a 9-year old child with Autism Spectrum Disorder (ASD) entitled to Special Education and Related services per IDEA represented by his parents R.S. *Pro se* and E.S. *Pro se*

Plaintiffs-Petitioners

-v.-

Board of Education Shenendehowa Central School District,

Interim Commissioner Betty Rosa, of The University of the State of New York

Defendants-Respondents

Petition for Rehearing

On Writ of Certiorari

To the U.S. Court of Appeals for the 2nd Circuit

CERTIFICATION PETITION IS PRESENTED IN GOOD FAITH AND NOT FOR DELAY

As required by Supreme Court Rule 44.1, I certify that the PETITION FOR REHEARING in the above Case No. 20-619 is presented in good faith and not for delay.

Respectfully Submitted on February 5, 2021

By: ______________________________
R.S. *Pro Se* on behalf of A.S.

By: ______________________________
E.S. *Pro Se* on behalf of A.S.

Sworn to me on this 5th day of Feb. 2021, personally appeared Roger Swartz and Ekaterina Shishova, before me a notary public.

CHRISTINE M ALEO
Notary Public - State of New York
No. 01AL6307204
Qualified in Albany County
My Commission Exp. 06/30/2022

No. 20-619

IN THE

SUPREME COURT OF THE UNITED STATES

◆

A.S. a 9-year old child with Autism Spectrum Disorder (ASD) entitled to Special Education and Related services per IDEA represented by his parents R.S. *Pro se* and E.S. *Pro se*

Plaintiffs-Petitioners

-v.-

Board of Education Shenendehowa Central School District,

Interim Commissioner Betty Rosa, of The University of the State of New York

Defendants-Respondents

◆

Petition for Rehearing

On Writ of Certiorari

To the U.S. Court of Appeals for the 2nd Circuit

◆

CERTIFICATE OF COMPLIANCE

As required by Supreme Court Rule 44.2, I certify that the PETITION FOR REHEARING in the above Case

No. 20-619 is filed on Substantial Grounds Not Previously Presented.

Respectfully Submitted on February 5, 2021

By: ____________________________________
R.S. *Pro Se* on behalf of A.S.

By: ____________________________________
E.S. *Pro Se* on behalf of A.S.

Sworn to me on this 5th day of Feb. 2021, personally appeared Roger Swartz And Ekaterina Shishova, before me a notary public.

Christine M Aleo

CHRISTINE M ALEO
Notary Public - State of New York
No. 01AL6307204
Qualified in Albany County
My Commission Exp. 06/30/2022

No. 20-619

IN THE

SUPREME COURT OF THE UNITED STATES

◆

A.S. a 9-year old child with Autism Spectrum Disorder (ASD) entitled
to Special Education and Related services per IDEA
represented by his parents R.S. *Pro se* and E.S. *Pro se*

Plaintiffs-Petitioners PETITION FOR

REHEARING

-v.-

Board of Education Shenendehowa Central School District,
Interim Commissioner Betty Rosa, of The University of the State of New York

Defendants-Respondents

◆

Petition for Rehearing

On Petition for a Writ of Certiorari

To the U.S. Court of Appeals for the 2nd Circuit

◆

TABLE OF CONTENTS FOR APPENDIX TO PETITION FOR REHEARING

Pet. Reh. App.1

Journal of Consulting and Clinical Psychology
1987, Vol. 55, No. 1, 3–9

Behavioral Treatment and Normal Educational and Intellectual Functioning in Young Autistic Children

O. Ivar Lovaas
University of California, Los Angeles

Autism is a serious psychological disorder with onset in early childhood. Autistic children show minimal emotional attachment, absent or abnormal speech, retarded IQ, ritualistic behaviors, aggression, and self-injury. The prognosis is very poor, and medical therapies have not proven effective. This article reports the results of behavior modification treatment for two groups of similarly constituted, young autistic children. Follow-up data from an intensive, long-term experimental treatment group (n = 19) showed that 47% achieved normal intellectual and educational functioning, with normal-range IQ scores and successful first grade performance in public schools. Another 40% were mildly retarded and assigned to special classes for the language delayed, and only 10% were profoundly retarded and assigned to classes for the autistic/retarded. In contrast, only 2% of the control-group children (n = 40) achieved normal educational and intellectual functioning; 45% were mildly retarded and placed in language-delayed classes, and 53% were severely retarded and placed in autistic/retarded classes.

Kanner (1943) defined autistic children as children who exhibit (a) serious failure to develop relationships with other people before 30 months of age, (b) problems in development of normal language, (c) ritualistic and obsessional behaviors ("insistence on sameness"), and (d) potential for normal intelligence. A more complete behavioral definition has been provided elsewhere (Lovaas, Koegel, Simmons, & Long, 1973). The etiology of autism is not known, and the outcome is very poor. In a follow-up study on young autistic children, Rutter (1970) reported that only 1.5% of his group (n = 63) had achieved normal functioning. About 35% showed fair or good adjustment, usually required some degree of supervision, experienced some difficulties with people, had no personal friends, and showed minor oddities of behavior. The majority (more than 60%) remained severely handicapped and were living in hospitals for mentally retarded or psychotic individuals or in other protective settings. Initial IQ scores appeared stable over time. Other studies (Brown, 1969; DeMyer et al., 1973; Eisenberg, 1956; Freeman, Ritvo, Needleman, & Yokota, 1985; Havelkova, 1968) report similar data. Higher scores on IQ tests, communicative speech, and appropriate play are considered to be prognostic of better outcome (Lotter, 1967).

Medically and psychodynamically oriented therapies have not proven effective in altering outcome (DeMyer, Hingtgen, & Jackson, 1981). No abnormal environmental etiology has been identified within the children's families (Lotter, 1967). At present, the most promising treatment for autistic persons is behavior modification as derived from modern learning theory (DeMyer et al., 1981). Empirical results from behavioral intervention with autistic children have been both positive and negative. On the positive side, behavioral treatment can build complex behaviors, such as language, and can help to suppress pathological behaviors, such as aggression and self-stimulatory behavior. Clients vary widely in the amount of gains obtained but show treatment gains in proportion to the time devoted to treatment. On the negative side, treatment gains have been specific to the particular environment in which the client was treated, substantial relapse has been observed at follow-up, and no client has been reported as recovered (Lovaas et al., 1973).

The present article reports a behavioral-intervention project (begun in 1970) that sought to maximize behavioral treatment gains by treating autistic children during most of their waking hours for many years. Treatment included all significant persons in all significant environments. Furthermore, the project focused on very young autistic children (below the age of 4 years) because it was assumed that younger children would be less likely to discriminate between environments and therefore more likely to generalize and to maintain their treatment gains. Finally, it was assumed that it would be easier to successfully mainstream a very young autistic child into preschool than it would be to mainstream an older autistic child into primary school.

It may be helpful to hypothesize an outcome of the present study from a developmental or learning point of view. One may assume that normal children learn from their everyday environ-

This study was supported by Grant MH-11440 from the National Institute of Mental Health. Aspects of this study were presented at the 1982 convention of the American Psychological Association, Washington, DC, by Andrea Ackerman, Paula Firestone, Gayle Goldstein, Ronald Leaf, John McEachin, and the author. The author expresses his deep appreciation to the many undergraduate students at the University of California, Los Angeles, who served as student therapists on the project, to the many graduate students who served as clinic supervisors, and to the many parents who trusted their children to our care. Special thanks to Laura Schreibman and Robert Koegel, who collaborated in the early stages of this research project. Donald Baer, Bruce Baker, Bradley Bucher, Arthur Woodward, and Haikang Shen provided statistical advice and help in manuscript preparation. B. J. Freeman's help in arranging access to Control Group 2 data is also appreciated.

Correspondence concerning this article should be addressed to O. Ivar Lovaas, Psychology Department, University of California, 405 Hilgard Avenue, Los Angeles, California 90024.

Pet. Reh. App.2

ments most of their waking hours. Autistic children, conversely, do not learn from similar environments. We hypothesized that construction of a special, intense, and comprehensive learning environment for very young autistic children would allow some of them to catch up with their normal peers by first grade.

Method

Subjects

Subjects were enrolled for treatment if they met three criteria: (a) independent diagnosis of autism from a medical doctor or a licensed PhD psychologist, (b) chronological age (CA) less than 40 months if mute and less than 46 months if echolalic, and (c) prorated mental age (PMA) of 11 months or more at a CA of 30 months. The last criterion excluded 15% of the referrals.

The clinical diagnosis of autism emphasized emotional detachment, extreme interpersonal isolation, little if any toy or peer play, language disturbance (mutism or echolalia), excessive rituals, and onset in infancy. The diagnosis was based on a structured psychiatric interview with parents, on observations of the child's free-play behaviors, on psychological testing of intelligence, and on access to pediatric examinations. Over the 15 years of the project, the exact wording of the diagnosis changed slightly in compliance with changes in the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-III; American Psychiatric Association, 1980). During the last years, the diagnosis was made in compliance with DSM-III criteria (p. 87). In almost all cases, the diagnosis of autism had been made prior to family contact with the project. Except for one case each in the experimental group and Control Group 1, all cases were diagnosed by staff of the Department of Child Psychiatry, University of California, Los Angeles (UCLA) School of Medicine. Members of that staff have contributed to the writing of the DSM-III and to the diagnosis of autism adopted by the National Society for Children and Adults with Autism. If the diagnosis of autism was not made, the case was referred elsewhere. In other words, the project did not select its cases. More than 90% of the subjects received two or more independent diagnoses, and agreement on the diagnosis of autism was 100%. Similarly high agreement was not reached for subjects who scored within the profoundly retarded range on intellectual functioning (PMA < 11 months); these subjects were excluded from the study.

Treatment Conditions

Subjects were assigned to one of two groups: an intensive-treatment experimental group ($n = 19$) that received more than 40 hours of one-to-one treatment per week, or the minimal-treatment Control Group 1 ($n = 19$) that received 10 hours or less of one-to-one treatment per week. Control Group 1 was used to gain further information about the rate of spontaneous improvement in very young autistic children, especially those selected by the same agency that provided the diagnostic work-up for the intensive-treatment experimental group. Both treatment groups received treatment for 2 or more years. Strict random assignment (e.g., based on a coin flip) to these groups could not be used due to parent protest and ethical considerations. Instead, subjects were assigned to the experimental group unless there was an insufficient number of staff members available to render treatment (an assessment made prior to contact with the family). Two subjects were assigned to Control Group 1 because they lived further away from UCLA than a 1-hr drive, which made sufficient staffing unavailable to those clients. Because fluctuations in staff availability were not associated in any way with client characteristics, it was assumed that this assignment would produce unbiased groups. A large number of pretreatment measures were collected to test this assumption. Subjects did not change group assignment. Except for two families who left the experimental group within the first 6 months (this group began with 21 subjects), all families stayed with their groups from beginning to end.

Assessments

Pretreatment mental age (MA) scores were based on the following scales (in order of the frequency of their use): the Bayley Scales of Infant Development (Bayley, 1955), the Cattell Infant Intelligence Scale (Cattell, 1960), the Stanford-Binet Intelligence Scale (Thorndike, 1972), and the Gesell Infant Development Scale (Gesell, 1949). The first three scales were administered to 90% of the subjects, and relative usage of these scales was similar in each group. Testing was carried out by graduate students in psychology who worked under the supervision of clinical psychologists at UCLA or licensed PhD psychologists at other agencies. The examiner chose the test that would best accommodate each subject's developmental level, and this decision was reached independently of the project staff. Five subjects were judged to be untestable (3 in the experimental group and 2 in Control Group 1). Instead, the Vineland Social Maturity Scale (Doll, 1953) was used to estimate their MAs (with the mother as informant). To adjust for variations in MA scores as a function of the subject's CA at the time of test administration, PMA scores were calculated for a CA at 30 months (MA/CA × 30).

Behavioral observations were based on videotaped recordings of the subject's free-play behavior in a playroom equipped with several simple early-childhood toys. These videotaped recordings were subsequently scored for amount of (a) *self-stimulatory behaviors*, defined as prolonged ritualistic, repetitive, and stereotyped behavior such as body-rocking, prolonged gazing at lights, excessive hand-flapping, twirling the body as a top, spinning or lining of objects, and licking or smelling of objects or wall surfaces; (b) *appropriate play behaviors*, defined as those limiting the use of toys in the playroom to their intended purposes, such as pushing the truck on the floor, pushing buttons on the toy cash register, putting a record on the record player, and banging with the toy hammer; and (c) *recognizable words*, defined to include any recognizable word, independent of whether the subject used it in a meaningful context or for communicative purposes. One observer who was naive about subjects' group placement scored all tapes after being trained to agree with two experienced observers (using different training tapes from similar subjects). Interobserver reliability was scored on 20% of the tapes (randomly selected) and was computed for each category of behavior for each subject by dividing the sum of observer agreements by the sum of agreements and disagreements. These scores were then summed and averaged across subjects. The mean agreement (based both on occurrences and nonoccurrences) was 91% for self-stimulatory behavior, 85% for appropriate play behavior, and 100% for recognizable words. A more detailed description of these behavioral recordings has been provided elsewhere (Lovaas et al., 1973).

A 1-hr parent interview about the subjects' earlier history provided some diagnostic and descriptive information. Subjects received a score of 1 for each of the following variables parents reported: no recognizable words; no toy play (failed to use toys for their intended function); lack of emotional attachment (failed to respond to parents' affection); apparent sensory deficit (parents had suspected their child to be blind or deaf because the child exhibited no or minimal eye contact and showed an unusually high pain threshold); no peer play (subject did not show interactive play with peers); self-stimulatory behavior; tantrums (aggression toward family members or self); and no toilet training. These 8 measures from parents' intake interviews were summed to provide a sum pathology score. The intake interview also provided information about abnormal speech (0 = normal and meaningful language, however limited; 1 = echolalic language used meaningfully [e.g., to express needs]; 2 = echolalia; and 3 = mute); age of walking; number of siblings in the family; socioeconomic status of the father; sex; and neurological examinations (including EEGs and CAT scans) that resulted in findings of pathology. Finally, CA at first diagnosis and at the beginning of the

Pet. Reh. App.3

present treatment were recorded. This yielded a total of 20 pretreatment measures, 8 of which were collapsed into 1 measure (sum pathology).

A brief clinical description of the experimental group at intake follows (identical to that for Control Group 1): Only 2 of the 19 subjects obtained scores within the normal range of intellectual functioning; 7 scored in the moderately retarded range, and 10 scored in the severely retarded range. No subject evidenced pretend or imaginary play, only 2 evidenced *complex* (several different or heterogeneous behaviors that together formed one activity) play, and the remaining subjects showed *simple* (the same elementary but appropriate response made repeatedly) play. One subject showed minimal appropriate speech, 7 were echolalic, and 11 were mute. According to the literature that describes the developmental delays of autistic children in general, the autistic subjects in the present study constituted an average (or below average) sample of such children.

Posttreatment measures were recorded as follows: Between the ages of 6 and 7 years (when a subject would ordinarily have completed first grade), information about the subjects' first-grade placement was sought and validated; about the same time, an IQ score was obtained. Testing was carried out by examiners who were naive about the subjects' group placement. Different scales were administered to accommodate different developmental levels. For example, a subject with a regular educational placement received a Wechsler Intelligence Scale for Children–Revised (WISC–R; Wechsler, 1974) or a Stanford-Binet Intelligence Scale (Thorndike, 1972), whereas a subject in an autistic/retarded class received a nonverbal test like the Merrill-Palmer Pre-School Performance Test (Stutsman, 1948). In all instances of subjects having achieved a normal IQ score, the testing was eventually replicated by other examiners. The scales (in order of the frequency of usage) included the WISC–R (Wechsler, 1974), the Stanford-Binet (Thorndike, 1972), the Peabody Picture Vocabulary Test (Dunn, 1981), the Wechsler Pre-School Scale (Wechsler, 1967), the Bayley Scales of Infant Development (Bayley, 1955), the Cattell Infant Intelligence Scale (Cattell, 1960), and the Leiter International Performance Scale (Leiter, 1959). Subjects received a score of 3 for *normal functioning* if they received a score on the WISC–R or Stanford-Binet in the normal range, completed first grade in a normal class in a school for normal children, and were advanced to the second grade by the teacher. Subjects received a score of 2 if they were placed in first-grade in a smaller *aphasia* (language delayed, language handicapped, or learning disabled) class. Placement in the aphasia class implied a higher level of functioning than placement in classes for the autistic/retarded, but the diagnosis of autism was almost always retained. A score of 1 was given if the first-grade placement was in a class for the autistic/retarded and if the child's IQ score fell within the severely retarded range.

Treatment Procedure

Each subject in the experimental group was assigned several well trained student therapists who worked (part-time) with the subject in the subject's home, school, and community for an average of 40 hr per week for 2 or more years. The parents worked as part of the treatment team throughout the intervention; they were extensively trained in the treatment procedures so that treatment could take place for almost all of the subjects' waking hours, 365 days a year. A detailed presentation of the treatment procedure has been presented in a teaching manual (Lovaas et al., 1980). The conceptual basis of the treatment was reinforcement (operant) theory; treatment relied heavily on discrimination-learning data and methods. Various behavioral deficiencies were targeted, and separate programs were designed to accelerate development for each behavior. High rates of aggressive and self-stimulatory behaviors were reduced by being ignored; by the use of time-out; by the shaping of alternate, more socially acceptable forms of behavior; and (as a last resort) by the delivery of a loud "no" or a slap on the thigh contingent upon the presence of the undesirable behavior. Contingent physical aversives were not used in the control group because inadequate staffing in that group did not allow for adequate teaching of alternate, socially appropriate behaviors.

During the first year, treatment goals consisted of reducing self-stimulatory and aggressive behaviors, building compliance to elementary verbal requests, teaching imitation, establishing the beginnings of appropriate toy play, and promoting the extension of the treatment into the family. The second year of treatment emphasized teaching expressive and early abstract language and interactive play with peers. Treatment was also extended into the community to teach children to function within a preschool group. The third year emphasized the teaching of appropriate and varied expression of emotions; preacademic tasks like reading, writing, and arithmetic; and *observational learning* (learning by observing other children learn). Subjects were enrolled only in those preschools where the teacher helped to carry out the treatment program. Considerable effort was exercised to mainstream subjects in a normal (average and public) preschool placement and to avoid initial placement in special education classes with the detrimental effects of exposure to other autistic children. This occasionally entailed withholding the subject's diagnosis of autism. If the child became known as autistic (or as "a very difficult child") during the first year in preschool, the child was encouraged to enroll in another, unfamiliar school (to start fresh). After preschool, placement in public education classes was determined by school personnel. All children who successfully completed normal kindergarten successfully completed first grade and subsequent normal grades. Children who were observed to be experiencing educational and psychological problems received their school placement through Individualized Educational Plan (IEP) staffings (attended by educators and psychologists) in accordance with the Education For All Handicapped Children Act of 1975.

All subjects who went on to a normal first grade were reduced in treatment from the 40 hr per week characteristic of the first 2 years to 10 hr or less per week during kindergarten. After a subject had started first grade, the project maintained a minimal (at most) consultant relationship with some families. In two cases, this consultation and the subsequent correction of problem behaviors were judged to be essential in maintaining treatment gains. Subjects who did not recover in the experimental group received 40 hr or more per week of one-to-one treatment for more than 6 years (more than 14,000 hr of one-to-one treatment), with some improvement shown each year but with only 1 subject recovering.

Subjects in Control Group 1 received the same kind of treatment as those in the experimental group but with less intensity (less than 10 hr of one-to-one treatment per week) and without systematic physical aversives. In addition, these subjects received a variety of treatments from other sources in the community such as those provided by small special education classes.

Control Group 2 consisted of 21 subjects selected from a larger group ($N = 62$) of young autistic children studied by Freeman et al. (1985). These subjects came from the same agency that diagnosed 95% of our other subjects. Data from Control Group 2 helped to guard against the possibility that subjects who had been referred to us for treatment constituted a subgroup with particularly favorable or unfavorable outcomes. To provide a group of subjects similar to those in the experimental group and Control Group 1, subjects for Control Group 2 were selected if they were 42 months old or younger when first tested, had IQ scores above 40 at intake, and had follow-up testing at 6 years of age. These criteria resulted in the selection of 21 subjects. Subjects in Control Group 2 were treated like Control Group 1 subjects but were not treated by the Young Autism Project described here.

Results

Pretreatment Comparisons

Eight pretreatment variables from the experimental group and Control Group 1 (CA at first diagnosis, CA at onset of treat-

Table 1
Means and F Ratios From Comparisons Between Groups on Intake Variables

Group	Diagnosis CA	Treatment CA	PMA	Recognizable words	Toy play	Self-stimulation	Sum pathology	Abnormal speech
Experimental	32.0	34.6	18.8	.42	28.2	12.1	6.9	2.4
Control 1	35.3	40.9	17.1	.58	20.2	19.6	6.4	2.2
F[a]	1.58	4.02*	1.49	.92	2.76	3.37	.82	.36

Note. CA = chronological age; PMA = prorated mental age. Experimental group, n = 19; Control Group 1, n = 19.
[a] *df* = 1, 36.
* $p < .05$.

ment, PMA, sum pathology, abnormal speech, self-stimulatory behavior, appropriate toy play, and recognizable words) were subjected to a multivariate analysis of variance (MANOVA; Brecht & Woodward, 1984). The means and *F* ratios from this analysis are presented in Table 1. As can be seen, there were no significant differences between the groups except for CA at onset of our treatment ($p < .05$). Control subjects were 6 months older on the average than experimental subjects (mean CAs of 35 months vs. 41 months, respectively). These differences probably reflect the delay of control subjects in their initiation into the treatment project because of staff shortages; analysis will show that differential CAs are not significantly related to outcome. To ascertain whether another test would reveal a statistically significant difference between the groups on toy play, descriptions of the subjects' toy play (taken from the videotaped recordings) were typed on cards and rated for their developmental level by psychology students who were naive about the purpose of the ratings and subject group assignment. The ratings were reliable among students (r = .79, $p < .001$), and an *F* test showed no significant difference in developmental levels of toy play between the two groups.

The respective means from the experimental group and Control Group 1 on the eight variables from the parent interview were .89 and .74 for sensory deficit, .63 and .42 for adult rejection, .58 and .47 for no recognizable words, .53 and .63 for no toy play, 1.0 and 1.0 for no peer play, .95 and .89 for body self-stimulation, .89 and .79 for tantrums, and .68 and .63 for no toilet training. The experimental group and Control Group 1 were also similar in onset of walking (6 vs. 8 early walkers; 1 vs. 2 late walkers), number of siblings in the family (1.26 in each group), socioeconomic status of the father (Level 49 vs. Level 54 according to 1950 Bureau of the Census standards), boys to girls (16:3 vs. 11:8); and number of subjects referred for neurological examinations (10 vs. 15) who showed signs of damage (0 vs. 1). The numbers of favorable versus unfavorable prognostic signs (directions of differences) on the pretreatment variables divide themselves equally between the groups. In short, the two groups appear to have been comparable at intake.

Follow-Up Data

Subjects' PMA at intake, follow-up educational placement, and IQ scores were subjected to a MANOVA that contrasted the experimental group with Control Groups 1 and 2. At intake, there were no significant differences between the experimental group and the control groups. At follow-up, the experimental group was significantly higher than the control groups on educational placement ($p < .001$) and IQ ($p < .01$). The two control groups did not differ significantly at intake or at follow-up. In short, data from Control Group 2 replicate those from Control Group 1 and further validate the effectiveness of our experimental treatment program. Data are given in Table 2 that show the group means from pretreatment PMA and posttreatment educational placement and IQ scores. The table also shows the *F* ratios and significance levels of the three group comparisons.

In descriptive terms, the 19-subject experimental group shows 9 children (47%) who successfully passed through normal first grade in a public school and obtained an average or above average score on IQ tests (M = 107, range = 94–120). Eight subjects (42%) passed first grade in aphasia classes and obtained a mean IQ score within the mildly retarded range of intellectual functioning (M = 70, range = 56–95). Only two children (10%) were placed in classes for autistic/retarded children and scored in the profoundly retarded range (IQ < 30).

There were substantial increases in the subjects' levels of intellectual functioning after treatment. The experimental group subjects gained on the average of 30 IQ points over Control Group 1 subjects. Thus the number of subjects who scored within the normal range of intellectual functioning increased from 2 to 12, whereas the number of subjects within the moderate-to-severe range of intellectual retardation dropped from 10 to 3. As of 1986, the achievements of experimental group sub-

Table 2
Means and F Ratios for Measures at Pretreatment and Posttreatment

Group	Intake PMA	Follow-up	
		EDP	IQ
	Means		
Experimental	18.8	2.37	83.3
Control 1	17.1	1.42	52.2
Control 2	17.6	1.57	57.5
	F ratios[a]		
Experimental × Control 1	1.47	23.6**	14.4**
Experimental × Control 2	0.77	17.6**	10.4*
Control 1 × Control 2	0.14	0.63	0.45

Note. PMA = prorated mental age; EDP = educational placement. Experimental group, n = 19; Control Group 1, n = 19; Control Group 2, n = 21.
[a] *df* = 1, 56.
* $p < .01$. ** $p < .001$.

Pet. Reh. App.5

Table 3
Educational Placement and Mean and Range of IQ at Follow-Up

Group	Recovered	Aphasic	Autistic/Retarded
Experimental			
N	9	8	2
M IQ	107	70	30
Range	94–120	56–95	—[a]
Control Group 1			
N	0	8	11
M IQ	—	74	36
Range	—	30–102	20–73
Control Group 2			
N	1	10	10
M IQ	99	67	44
Range	—	49–81	35–54

Note. Dashes indicate no score or no entry.
[a] Both children received the same score.

jects have remained stable. Only 2 subjects have been reclassified: 1 subject (now 18 years old) was moved from an aphasia to a normal classroom after the sixth grade; 1 subject (now 13 years old) was moved from an aphasia to an autistic/retarded class placement.

The MA and IQ scores of the two control groups remained virtually unchanged between intake and follow-up, consistent with findings from other studies (Freeman et al., 1985; Rutter, 1970). The stability of the IQ scores of the young autistic children, as reported in the Freeman et al. study, is particularly relevant for the present study because it reduces the possibility of spontaneous recovery effects. In descriptive terms, the combined follow-up data from the control groups show that their subjects fared poorly: Only 1 subject (2%) achieved normal functioning as evidenced by normal first-grade placement and an IQ of 99 on the WISC-R; 18 subjects (45%) were in aphasia classes (mean IQ = 70, range = 30–101); and 21 subjects (53%) were in classes for the autistic/retarded (mean IQ = 40, range = 20–73). Table 3 provides a convenient descriptive summary of the main follow-up data from the three groups.

One final control procedure subjected 4 subjects in the experimental group (Ackerman, 1980) and 4 subjects in Control Group 1 (McEachin & Leaf, 1984) to a treatment intervention in which one component of treatment (the loud "no" and occasional slap on the thigh contingent on self-stimulatory, aggressive, and noncompliant behavior) was at first withheld and then introduced experimentally. A within-subjects replication design was used across subjects, situations, and behaviors, with baseline observations varying from 3 weeks to 2 years after treatment had started (using contingent positive reinforcement only). During baseline, when the contingent-aversive component was absent, small and unstable reductions were observed in the large amount of inappropriate behaviors, and similar small and unstable increases were observed in appropriate behaviors such as play and language. These changes were insufficient to allow for the subjects' successful mainstreaming. Introduction of contingent aversives resulted in a sudden and stable reduction in the inappropriate behaviors and a sudden and stable increase in appropriate behaviors. This experimental intervention helps to establish two points: First, at least one component in the treatment program functioned to produce change, which helps to reduce the effect of placebo variables. Second, this treatment component affected both the experimental and control groups in a similar manner, supporting the assumption that the two groups contained similar subjects.

Analyses of variance were carried out on the eight pretreatment variables to determine which variables, if any, were significantly related to outcome (gauged by educational placement and IQ) in the experimental group and Control Group 1. Prorated mental age was significantly ($p < .03$) related to outcome in both groups, a finding that is consistent with reports from other investigators (DeMyer et al., 1981). In addition, abnormal speech was significantly ($p < .01$) related to outcome in Control Group 1. Chronological age at onset of our treatment was not related to outcome, which is important because the two groups differed significantly on this variable at intake (by 6 months). The failure of CA to relate to outcome may be based on the very young age of all subjects at onset of treatment.

Conceivably, a linear combination of pretreatment variables could have predicted outcome in the experimental group. Using a discriminant analysis (Ray, 1982) with the eight variables used in the first multivariate analysis, it was possible to predict perfectly the 9 subjects who did achieve normal functioning, and no subject was predicted to achieve this outcome who did not. In this analysis, PMA was the only variable that was significantly related to outcome. Finally, when this prediction equation was applied to Control Group 1 subjects, 8 were predicted to achieve normal functioning with intensive treatment; this further verifies the similarity between the experimental group and Control Group 1 prior to treatment.

Discussion

This article reports the results of intensive behavioral treatment for young autistic children. Pretreatment measures revealed no significant differences between the intensively treated experimental group and the minimally treated control groups. At follow-up, experimental group subjects did significantly better than control group subjects. For example, 47% of the experimental group achieved normal intellectual and educational functioning in contrast to only 2% of the control group subjects.

The study incorporated certain methodological features designed to increase confidence in the effectiveness of the experimental group treatment:

1. Pretreatment differences between the experimental and control groups were minimized in four ways. First, the assignment of subjects to groups was as random as was ethically possible. The assignment apparently produced unbiased groups as evidenced by similar scores on the 20 pretreatment measures and by the prediction that an equal number of Control Group 1 and experimental group subjects would have achieved normal functioning had the former subjects received intensive treatment. Second, the experimental group was not biased by receiving subjects with a favorable diagnosis or biased IQ testing because both diagnosis and IQ tests were constant across groups. Third, the referral process did not favor the project cases because there were no significant differences between Control Groups 1 and 2 at intake or follow-up, even though Control Group 2 subjects were referred to others by the same agency.

Fourth, subjects stayed within their groups, which preserved the original (unbiased) group assignment.

2. A favorable outcome could have been caused not by the experimental treatment but by the attitudes and expectations of the staff. There are two findings that contradict this possibility of treatment agency (placebo) effects. First, because Control Group 2 subjects had no contact with the project, and because there was no difference between Control Groups 1 and 2 at follow-up, placebo effects appear implausible. Second, the within-subjects study showed that at least one treatment component contributed to the favorable outcome in the intensive treatment (experimental) group.

3. It may be argued that the treatment worked because the subjects were not truly autistic. This is counterindicated by the high reliability of the independent diagnosis and by the outcome data from the control groups, which are consistent with those reported by other investigators (Brown, 1969; DeMeyer et al., 1973; Eisenberg, 1956; Freeman et al., 1985; Havelkova, 1968; Rutter, 1970) for groups of young autistic children diagnosed by a variety of other agencies.

4. The spontaneous recovery rate among very young autistic children is unknown, and without a control group the favorable outcome in the experimental group could have been attributed to spontaneous recovery. However, the poor outcome in the similarly constituted Control Groups 1 and 2 would seem to eliminate spontaneous recovery as a contributing factor to the favorable outcome in the experimental group. The stability of the IQ test scores in the young autistic children examined by Freeman et al. (1985) attests once again to the chronicity of autistic behaviors and serves to further negate the effects of spontaneous recovery.

5. Posttreatment data showed that the effects of treatment (a) were substantial and easily detected, (b) were apparent on comprehensive, objective, and socially meaningful variables (IQ and school placement), and (c) were consistent with a very large body of prior research on the application of learning theory to the treatment and education of developmentally disabled persons and with the very extensive (100-year-old) history of psychology laboratory work on learning processes in man and animals. In short, the favorable outcome reported for the intensive-treatment experimental group can in all likelihood be attributed to treatment.

A number of measurement problems remain to be solved. For example, play, communicative speech, and IQ scores define the characteristics of autistic children and are considered predictors of outcome. Yet the measurement of these variables is no easy task. Consider play. First, play undoubtedly varies with the kinds of toys provided. Second, it is difficult to distinguish low levels of toy play (simple and repetitive play associated with young, normal children) from high levels of self-stimulatory behavior (a psychotic attribute associated with autistic children). Such problems introduce variability that needs immediate attention before research can proceed in a meaningful manner.

The term *normal functioning* has been used to describe children who successfully passed normal first grade and achieved an average IQ on the WISC–R. But questions can be asked about whether these children truly recovered from autism. On the one hand, educational placement is a particularly valuable measure of progress because it is sensitive to both educational accomplishments and social-emotional functions. Also, continual promotion from grade to grade is made not by one particular teacher but by several teachers. School personnel describe these children as indistinguishable from their normal friends. On the other hand, certain residual deficits may remain in the normal functioning group that cannot be detected by teachers and parents and can only be isolated on closer psychological assessment, particularly as these children grow older. Answers to such questions will soon be forthcoming in a more comprehensive follow-up (McEachin, 1987).

Several questions about treatment remain. It is unlikely that a therapist or investigator could replicate our treatment program for the experimental group without prior extensive theoretical and supervised practical experience in one-to-one behavioral treatment with developmentally disabled clients as described here and without demonstrated effectiveness in teaching complex behavioral repertoires as in imitative behavior and abstract language. In the within-subjects studies that were reported, contingent aversives were isolated as one significant variable. It is therefore unlikely that treatment effects could be replicated without this component. Many treatment variables are left unexplored, such as the effect of normal peers. Furthermore, the successful mainstreaming of a 2–4-year-old into a normal preschool group is much easier than the mainstreaming of an older autistic child into the primary grades. This last point underscores the importance of early intervention and places limits on the generalization of our data to older autistic children.

Historically, psychodynamic theory has maintained a strong influence on research and treatment with autistic children, offering some hope for recovery through experiential manipulations. By the mid-1960s, an increasing number of studies reported that psychodynamic practitioners were unable to deliver on that promise (Rimland, 1964). One reaction to those failures was an emphasis on organic theories of autism that offered little or no hope for major improvements through psychological and educational interventions. In a comprehensive review of research on autism, DeMyer et al. (1981) concluded that "[in the past] psychotic children were believed to be *potentially* capable of normal functioning in virtually all areas of development . . . during the decade of the 1970s it was the rare investigator who even gave lip-service to such previously held notions . . . infantile autism is a type of developmental disorder accompanied by severe and, to a large extent, permanent intellectual/behavioral deficits" (p. 432).

The following points can now be made. First, at least two distinctively different groups emerged from the follow-up data in the experimental group. Perhaps this finding implies different etiologies. If so, future theories of autism will have to identify these groups of children. Second, on the basis of testing to date, the recovered children show no permanent intellectual or behavioral deficits and their language appears normal, contrary to the position that many have postulated (Rutter, 1974; Churchill, 1978) but consistent with Kanner's (1943) position that autistic children possess potentially normal or superior intelligence. Third, at intake, all subjects evidenced deficiencies across a wide range of behaviors, and during treatment they showed a broad improvement across all observed behaviors. The kind of (hypothesized) neural damage that mediates a particular kind of behavior, such as language (Rutter, 1974), is not consistent with these data.

Although serious problems remain for exactly defining autism or identifying its etiology, one encouraging conclusion can be stated: Given a group of children who show the kinds of behavioral deficits and excesses evident in our pretreatment measures, such children will continue to manifest similar severe psychological handicaps later in life unless subjected to intensive behavioral treatment that can indeed significantly alter that outcome.

These data promise a major reduction in the emotional hardships of families with autistic children. The treatment procedures described here may also prove equally effective with other childhood disorders, such as childhood schizophrenia. Certain important, practical implications in these findings may also be noted. The treatment schedule of subjects who achieved normal functioning could be reduced from 40 hr per week to infrequent visits even after the first 2 years of treatment. The assignment of one full-time special-education teacher for 2 years would cost an estimated $40,000, in contrast to the nearly $2 million incurred (in direct costs alone) by each client requiring life-long institutionalization.

References

Ackerman, A. B. (1980). *The contribution of punishment to the treatment of preschool aged children.* Unpublished doctoral dissertation, University of California, Los Angeles.

American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.). Washington, DC: Author.

Bayley, N. (1955). On the growth of intelligence. *American Psychologist, 10,* 805–818.

Brecht, M. L., & Woodward, J. A. (1984). GANOVA: A univariate/multivariate analysis of variance program for the personal computer. *Educational and Psychological Measurement, 44,* 169–173.

Brown, J. (1969). Adolescent development of children with infantile psychosis. *Seminars in Psychiatry, 1,* 79–89.

Cattell, P. (1960). *The measurement of intelligence of infants and young children.* New York: Psychological Corporation.

Churchill, D. W. (1978). Language: The problem beyond conditioning. In M. Rutter & E. Schopler (Eds.), *Autism: A reappraisal of concepts and treatment* (pp. 71–85). New York: Plenum.

DeMyer, M. K., Barton, S., DeMyer, W. E., Norton, J. A., Allen, J., & Steele, R. (1973). Prognosis in autism: A follow-up study. *Journal of Autism and Childhood Schizophrenia, 3.* 199–246.

DeMyer, M. K., Hingtgen, J. N., & Jackson, R. K. (1981). Infantile autism reviewed: A decade of research. *Schizophrenia Bulletin, 7,* 388–451.

Doll, E. A. (1953). *The measurement of social competence.* Minneapolis, MN: Minneapolis Educational Test Bureau.

Dunn, L. M. (1981). *Peabody Picture Vocabulary Test.* Circle River, MI: American Guidance Service.

Education for All Handicapped Children Act of 1975. Washington, DC: *Congressional Record.*

Eisenberg, L. (1956). The autistic child in adolescence. *American Journal of Psychiatry, 112,* 607–612.

Freeman, B. J., Ritvo, E. R., Needleman, R., & Yokota, A. (1985). The stability of cognitive and linguistic parameters in autism: A 5-year study. *Journal of the American Academy of Child Psychiatry, 24,* 290–311.

Gesell, A. (1949). *Gesell Developmental Schedules.* New York: Psychological Corporation.

Havelkova, M. (1968). Follow-up study of 71 children diagnosed as psychotic in preschool age. *American Journal of Orthopsychiatry, 38,* 846–857.

Kanner, L. (1943). Autistic disturbances of affective contact. *Nervous Child, 2,* 217–250.

Leiter, R. G. (1959). Part I of the manual for the 1948 revision of the Leiter International Performance Scale: Evidence of the reliability and validity of the Leiter tests. *Psychology Service Center Journal, 11,* 1–72.

Lotter, V. (1967). Epidemiology of autistic conditions in young children: II. Some characteristics of the parents and children. *Social Psychiatry, 1,* 163–173.

Lovaas, O. I., Ackerman, A. B., Alexander, D., Firestone, P., Perkins, J., & Young, D. (1980). *Teaching developmentally disabled children: The me book.* Austin, TX: Pro-Ed.

Lovaas, O. I., Koegel, R. L., Simmons, J. Q., & Long, J. (1973). Some generalization and follow-up measures on autistic children in behavior therapy. *Journal of Applied Behavior Analysis, 6,* 131–166.

McEachin, J. J. (1987). *Outcome of autistic children receiving intensive behavioral treatment: Residual deficits.* Unpublished doctoral dissertation, University of California, Los Angeles.

McEachin, J. J., & Leaf, R. B. (1984, May). *The role of punishment in motivation of autistic children.* Paper presented at the convention of the Association for Behavior Analysis, Nashville, TN.

Ray, A. A. (1982). *Statistical Analysis System user's guide: Statistics, 1982 edition.* Cary, NC: SAS Institute.

Rimland, B. (1964). *Infantile autism.* New York: Appleton-Century-Crofts.

Rutter, M. (1970). Autistic children: Infancy to adulthood. *Seminars in Psychiatry, 2,* 435–450.

Rutter, M. (1974). The development of infantile autism. *Psychological Medicine, 4,* 147–163.

Stutsman, R. (1948). *Guide for administering the Merrill-Palmer Scale of Mental Tests.* New York: Harcourt, Brace & World.

Thorndike, R. L. (1972). *Manual for Stanford-Binet Intelligence Scale.* Boston: Houghton Mifflin.

Wechsler, D. (1967). *Manual for the Wechsler Pre-School and Primary Scale of Intelligence.* New York: Psychological Corporation.

Wechsler, D. (1974). *Manual for the Wechsler Intelligence Scale for Children-Revised.* New York: Psychological Corporation.

Received October 10, 1985
Revision received March 28, 1986 ■

American Journal on Mental Retardation
1993, Vol. 97, No. 4, 359–372

Long-Term Outcome for Children With Autism Who Received Early Intensive Behavioral Treatment

John J. McEachin, Tristram Smith, and O. Ivar Lovaas
University of California, Los Angeles

After a very intensive behavioral intervention, an experimental group of 19 preschool-age children with autism achieved less restrictive school placements and higher IQs than did a control group of 19 similar children by age 7 (Lovaas, 1987). The present study followed-up this finding by assessing subjects at a mean age of 11.5 years. Results showed that the experimental group preserved its gains over the control group. The 9 experimental subjects who had achieved the best outcomes at age 7 received particularly extensive evaluations indicating that 8 of them were indistinguishable from average children on tests of intelligence and adaptive behavior. Thus, behavioral treatment may produce long-lasting and significant gains for many young children with autism.

Infantile autism is a condition marked by severe impairment in intellectual, social, and emotional functioning. Its onset occurs in infancy, and the prognosis appears to be extremely poor (Lotter, 1978). For example, in the longest prospective follow-up study with a sound methodological design, Rutter (1970) found that only 1 of 64 subjects with autism (fewer than 2%) could be considered free of clinically significant problems by adulthood, as evidenced by holding a job, living independently, and maintaining an active and age-appropriate social life. The remaining subjects showed numerous dysfunctions, such as marked oddities in behavior, social isolation, and florid psychopathology. The majority of subjects required supervised living conditions.

Professionals have attempted a wide variety of interventions in an effort to help children with autism. For many years, no scientific evidence showed that any of these interventions brightened the children's long-term prognosis (DeMyer et al., 1981). How-

This study was supported by Grant No. MH-11440 from the National Institute of Mental Health. The study was based on a dissertation submitted to the University of California, Los Angeles, Department of Psychology, in partial fulfillment of the requirements for the doctoral degree. The authors express their deep appreciation to the many students at UCLA who served as therapists and helped to make this study possible. Special thanks to Bruce Baker and Duane Buhrmester, who helped in the design of this study. Requests for reprints of this article, copies of the Clinical Rating Scale, or additional information about this study should be sent to O. Ivar Lovaas, 405 Hilgard Ave., UCLA, Department of Psychology, Los Angeles, CA 90024-1563.

ever, since the 1960s, one of these interventions, behavioral treatment, has appeared promising. Behavioral treatment has been found to increase adaptive behaviors such as language and social skills, while decreasing disruptive behaviors such as aggression (DeMyer, Hingtgen, & Jackson, 1981; Newsom & Rincover, 1989; Rutter, 1985). Furthermore, behavioral treatment has been continuously refined and improved as a result of ongoing research efforts at a number of sites (Lovaas & Smith, 1988).

Some recent evidence has indicated that behavioral treatment has developed to the point that it can produce substantial improvements in the overall functioning of young children with autism (Simeonnson, Olley, & Rosenthal, 1987). Lovaas (1987) provided approximately 40 hours per week of one-on-one behavioral treatment for a period of 2 years or more to an experimental group of 19 children with autism who were under 4 years of age. This intervention also included parent training and mainstreaming into regular preschool environments. When re-evaluated at a mean age of 7 years, subjects in the experimental group had gained an average of 20 IQ points and had made major advances in educational achievement. Nine of the 19 subjects completed first grade in regular (nonspecial education) classes entirely on their own and had IQs that increased to the average range. By contrast, two control groups totalling 40 children, also diagnosed as autistic and comparable to the experimental group at intake, did not fare nearly as well. Only one of the control subjects (2.5%) attained normal levels of intellectual and educational functioning.

These data suggest that behavioral treatment is effective. However, the durability of treatment gains is uncertain. In one prior major study, Lovaas, Koegel, Simmons, and Long (1973) found that children with autism regressed following the termination of treatment. Other studies have shown that children with autism may display increased difficulties when they enter adolescence (Kanner, 1971; Waterhouse & Fein, 1984). Also, as was stated in the first follow-up (Lovaas, 1987), "Certain residual deficits may remain in the normal-functioning group that cannot be detected by teachers and parents and can only be isolated on closer psychological assessment, particularly as these children grow older" (p. 8). This possibility points to the need for a more detailed assessment and for continued follow-ups of the group over time.

The present investigation contained two parts: In the first part we examined whether several years after the evaluation at age 7, the experimental group in Lovaas's (1987) study had maintained its treatment gains. Subjects in the experimental group and one of the control groups completed standardized tests of intellectual and adaptive functioning. The groups were then contrasted with each other, and their current performance was compared to their performance on previous assessments. The second part of the investigation focused on those subjects who had achieved the best outcome at the end of first grade in the Lovaas (1987) study (i.e., the 9 subjects who were classified as normal functioning out of the 19 in the experimental group). We examined the extent to which these best-outcome subjects could be considered free of autistic symptomatology. A test battery was constructed to assess a variety of possible deficits: for example, idiosyncratic thought patterns, mannerisms, and interests; lack of close relationships with family and friends; difficulty in getting along with people; relative weaknesses in certain areas of cognitive functioning, such as abstract reasoning; not working up to ability in school; flatness of affect; absence or peculiarity in sense of humor. Possible strengths to be identified included normal intellectual functioning, good relationships with family members, ability to function independently, appropriate use of leisure time, and adequate socialization with peers. Numerous methodological precautions were taken to ensure objectivity of the follow-up examination.

Method

Subjects and Background

Characteristics of the subjects and their treatment have been described elsewhere (Lovaas, 1987) and will only be summarized here. The initial treatment study contained 38 children who, at the time of intake, were very young (less than 40 months if mute, less than 46 months if echolalic) and had received a diagnosis of autism from a licensed clinical psychologist or psychiatrist not involved in the study. These 38 subjects were divided into an experimental group and a control group. The assignment to groups was made on the basis of staff availability. At the beginning of each academic quarter, treatment teams were formed. The clinic director and staff members then determined whether any opening existed for intensive treatment. If so, the next referral received would enter the experimental group; otherwise, the subject entered the control group. The experimental group contained 19 children who received 40 or more hours per week of one-to-one behavioral treatment for 2 or more years. The control group was comprised of 19 children who received a much less intensive intervention (10 hours a week or less of one-to-one behavioral treatment in addition to a variety of treatments provided by community agencies, such as parent training or special education classes). The initial study also included a second control group, consisting of 21 children with autism who were followed over time by a nearby agency but who were never referred for this study. However, these 21 subjects were not available for the present investigation. On standardized measures of intelligence, the second control group did not differ from either the experimental group or the first control group at intake, nor did it differ from the first control group when evaluated again when the subjects were 7 years old. These findings suggest that, as measured by standardized tests, (a) the children with autism who were referred to us for treatment were comparable to children with autism seen elsewhere and (b) the minimal treatment provided to the first control group did not alter intellectual functioning.

Statistical analysis of an extensive range of pretreatment measures confirmed that the experimental group and control group were comparable at intake and closely matched on such important variables as IQ and severity of disturbance. The mean chronological age (CA) at diagnosis for subjects in the experimental group was 32 months. Their mean IQ was 53 (range 30 to 82; all IQs are given as deviation scores). The mean CA of subjects in the control group was 35 months; their mean IQ was 46 (range 30 to 80). Most of the subjects were mute, all had gross deficiencies in receptive language, none played with peers or showed age-appropriate toy play, all were emotionally withdrawn, most had severe tantrums, and all showed extensive ritualistic and stereotyped (self-stimulatory) behaviors. Thus, they appeared to be a representative sample of children with autism (Lovaas, Smith, & McEachin, 1989). A more complete presentation of the intake data was reported by Lovaas (1987).

The children in the experimental group and control group received their respective treatments from trained student therapists who worked in the child's home. The parents also worked with their child, and they received extensive instruction and supervision on appropriate treatment techniques. Whenever possible, the children were integrated into regular preschools. The treatment focused primarily on developing language, increasing social behavior, and promoting cooperative play with peers along with independent and appropriate toy play. Concurrently, substantial efforts were directed at decreasing excessive rituals, tantrums, and aggressive behavior. (For a more detailed description of the intervention program, see the treatment manual [Lovaas et al., 1980] and instructional videotapes that supplement the manual [Lovaas & Leaf, 1981].)

At the time of the present follow-up (1984–1985), the mean CA of the experimental group children was 13 years (range = 9 to 19 years). All children who had achieved normal functioning by the age of 7 years had ended treatment by that point. (*Normal functioning* was operationally defined as scoring within the normal range on standardized intelligence tests and successfully completing first grade in a regular, nonspecial education class entirely on one's own.) On the other hand, some of the children who had not achieved normal functioning at 7 years of age had, at the request of their parents, remained in treatment. The length of time that experimental subjects had been out of treatment ranged from 0 to 12 years (mean = 5), with the normal-functioning children having been out for 3 to 9 years (mean = 5).

The mean age of subjects in the control group was 10 years (range 6 to 14). The length of time that these children had been out of treatment ranged from 0 to 9 years (mean = 3). Thus, experimental subjects tended to be older and had been out of treatment longer than had control subjects. This difference in age occurred because the first referrals for the study were all assigned to the experimental group due to the fact that referrals came slowly (7 in the first 3.5 years) and therapists were available to treat all of them. (As noted earlier, subjects were assigned to the experimental group if therapists were available to treat them; otherwise, they entered the control group.)

Statistical analyses were conducted to test whether a bias resulted from the tendency for the first referrals to go into the experimental group. For example, it is conceivable that the first referrals could have been higher functioning at intake or could have had a better prognosis than subsequent referrals. If so, the subject assignment procedure could have favored the experimental group. To assess this possibility, we correlated the order of referral with intake IQ and with IQ at the first follow-up (age 7 years). Pearson correlations were computed across both groups and within each group. These analyses indicated that the order in which subjects were referred was not associated with intake IQ or outcome IQ. Conseque although the tendency for the first referra enter the experimental group created a tential bias, the data indicate that this unlikely.

Procedure

The assessment procedure inclu ascertaining school placement and admi tering three standardized tests. Informa on school placement was obtained f subjects' parents, who classified ther being in either a regular or a special edu tion class (e.g., a class for children autism or mental retardation, language lays, multihandicaps, or learning disal ties). The three standardized tests were follows:

1. *Intelligence test.* The Wechsler In ligence Scale for Children-Revised (Wechs 1974) was administered when subjects w able to provide verbal responses. This cluded all 9 best-outcome experimental s jects plus 8 of the remaining 10 experime subjects and 6 of the 19 control subjects. subjects who were not able to provide ve responses, the Leiter International Per mance Scale (Leiter, 1959) and the Peab Picture Vocabulary Test-Revised (Dunn, 19 were administered. All of these tests h been widely used for the assessment intellectual functioning in children with tism (Short & Marcus, 1986).

2. *The Vineland Adaptive Beha Scales* (Sparrow, Balla, & Cicchetti, 19 The Vineland is a structured interview ministered to parents assessing the ext to which their child exhibits behaviors t are needed to cope effectively with everyday environment.

3. *The Personality Inventory for C dren* (Wirt, Lachar, Klinedinst, & Seat, 19 This measure is a 600-item true–false qu tionnaire filled out by parents that asses the extent to which their children sh various forms of psychological disturba (e.g., anxiety, depression, hyperactivity, a psychotic behavior).

These three tests were intended to provide a comprehensive evaluation of intellectual, social, and emotional functioning. All of the tests have been standardized on average populations. Hence, they provide an objective basis for comparing subjects to children without handicaps across the various areas that they assess.

Data were obtained on all subjects except one girl in the control group, who was known to be institutionalized and functioning very poorly. The 9 best-outcome subjects (those who had been classified as normal functioning at age 7) received particularly extensive evaluations, as outlined later. Of the 28 remaining subjects, 17 were evaluated by staff members in our treatment program, and 11 received evaluations from outside agencies such as schools or psychology clinics. (In some cases, the outside agencies did not administer all of the measures in this battery.)

Evaluation of Best-Outcome Subjects. To ensure objectivity in the evaluation of the best-outcome subjects, we arranged for blind administration and scoring of all tests for these subjects as follows. A psychologist not associated with the study recruited advanced graduate students in clinical psychology to administer the tests. The examiners were not familiar with the history of the children, and the psychologist told them simply that the testing was part of a research study on assessment of children. The psychologist advised them that the nature of the study necessitated providing only certain standard background information: age, school placement and grade, and parent's name and phone number. To increase the heterogeneity of the sample and to control for any examiner bias, each examiner also tested one or more subjects who were matched in age to the experimental subjects and had no history of behavioral disturbance. The examiners were randomly assigned an approximately equal number of subjects for testing in the experimental group and the comparison group. Two experimental subjects were not living in the local area. Therefore, for each of them, the psychologist recruited a tester from the subject's hometown area as well as an age-matched control subject, and data were collected as just described. In addition, the child's examiner filled out a clinical rating scale following a structured interview that covered a list of standard topics, including friendships, family relations, and school and community activities. The interview was designed both for eliciting content and for sampling interpersonal style. The rating scale consisted of 22 items, each scored 0 (best clinical status) to 3 (marked deviance) points. The items were designed to include likely areas of difficulty for children with autism of average intelligence (e.g., compulsive or ritualistic behavior, empathy for and interest in others, a sense of humor) as well as areas of potential difficulty for the general child population (e.g., depressed mood, anxiety, hyperactivity). (The complete scale and a copy of instructions for the clinical interview can be obtained by writing to the third author).

Results

Experimental Versus Control Group

This first section examines the overall effects of treatment through comparison of the follow-up data from the 19 subjects who received the intensive (experimental) treatment to the data from those who received the minimal (control) treatment. Data were obtained from all subjects on school placement and from all but one subject in the control group on IQ. On the Vineland, scores were obtained for 18 of 19 experimental subjects and 15 of 19 control subjects. The lowest availability of follow-up scores was on the Personality Inventory for Children, with scores for 15 experimental subjects and 12 control subjects.

The subjects in the control group who had Personality Inventory for Children scores did not appear to differ from subjects who were missing these scores, as compared on

Pet. Reh. App.13

t tests for differences in intake IQ, IQ at 7 years old, or IQ in the present study.

As noted earlier, 17 of the 29 subjects who were not in the best-outcome group were evaluated by Project staff members, 11 were evaluated by outside agencies, and 1 was not evaluated. To check whether Project staff members were biased in their evaluations or in their selection of which subjects to evaluate, we used *t* tests to compare subjects they evaluated to those evaluated by outside agencies on intake IQ, IQ at age 7 years, and IQ in the present study. No significant differences between subjects evaluated by Project staff members and those evaluated by outside agencies were found.

School Placement. In the experimental group, 1 of the 9 subjects from the best-outcome group who had attended a regular class at age 7 (J. L.) was now in a special education class. However, 1 of the other 10 subjects had gone from a special education class to a regular class and was enrolled in a junior college at the time of this follow-up. The remaining experimental subjects had not changed their classification. Overall, then, the proportion of experimental subjects in regular classes did not change from the age 7 evaluation (9 of 19, or 47%). In the control group, none of the 19 children were in a regular class, as had been true at the age 7 evaluation. The difference in classroom placement between the experimental group and the control group was statistically significant, χ^2 (1, N = 38) = 19.05, $p < .05$.

Intellectual Functioning. The test scores for the experimental group and control group on intellectual functioning, adaptive and maladaptive behaviors, and personality functioning are summarized in Table 1. As can be seen in the table, the experimental group at follow-up had a significantly higher mean IQ than did the control group. This difference was significant, $t(35) = 2.97$, $p < .01$. Eleven subjects (58%) in the experimental group obtained Full-Scale IQs of at least 80; only 3 subjects (17%) in the control group did as well. The scores were similar to those obtained by the experimental group and control group at age 7 (mean IQs of 83 and
respectively), indicating that the experime
tal group had maintained its gains in intelle
tual functioning between age 7 and the ti
of the current evaluation.

Table 1
Mean Scores and SDs by Group and Measure a
Follow-Up

	Group			
	Experimental		Control	
Measure	Mean	*SD*	Mean	S
IQ	84.5	32.4		
Vineland[a]				
Communication	5.1	28.4		
Daily Living Skills	73.1	26.9		
Socialization	75.5	26.8		
Adaptive Behavior Composite	71.6	26.8		
Maladaptive Behavior	10.6	8.2		
PIC[b] Scales				
Mean elevation	61.8	10.2		
Scales > 70	4.0	3.9		

[a]Vineland Adaptive Behavior Scale. [b]Personality Inven
for Children.

Adaptive and Maladaptive Behav
On the Vineland, the mean overall or Co
posite score was 72 in the experimen
group and 48 in the control group. (T
average score for the general population o
this test is 100, with a standard deviation [SD
of 15.) On the three subscales—Communic
tion, Daily Living, and Socialization—ea
score closely paralleled the Composite sco
The interaction between the groups and th
subscales was not significant, indicating th
across the three subscales, the experimenta
group consistently scored higher than di
the control group. As can be seen in Table
Maladaptive Behavior was significantly highe
in the control group, $t(31) = 2.39$, $p < .05$. Th
mean score for the control group was in th
clinically significant range whereas that o
the experimental group was not. (Scores o
13 and above are considered to be indicativ
of clinically significant levels of maladaptiv
behavior at ages 6 to 9 years; 12 or above, a
12 to 13 years; and 10 or above, at 14 year
and older.) Thus, the findings indicate th
the experimental group showed more adap
tive behaviors and fewer maladaptive beha

Pet. Reh. App.14

iors than did the control group.

Personality Functioning. Scores for the experimental group and control group did not differ on overall scale elevation, with mean *t* scores of 62 and 65, respectively. (On this test, the mean *t* score for the general population is approximately 50 [*SD* = 10].) *T* scores above 60 are considered indicative of possible or mild deviance, whereas *t* scores above 70 are viewed as suggesting a clinically significant problem, namely, one that may require professional attention. There was a significant interaction between the groups and the individual scales on this test, $F(15, 390) = 2.36$, $p < .01$. Results of the Tukey test indicated that the most reliable difference between groups occurred on the Psychosis scale, on which the experimental subjects had a mean of 78 and the control subjects had a mean of 104, $F(1, 26) = 8.53$, $p < .01$. Seven subjects in the experimental group scored in the clinically preferred range (below 70), whereas no subjects in the control group scored that low. Only one other scale showed a significant difference, Somatic Concerns, $F(1, 26) = 4.60$, $p < .05$. The control subjects tended to display a below average level of somatic complaints (mean of 45 as compared to 54 for the experimental subjects).

Best-Outcome Versus Nonclinical Comparison Group

A *t* test indicated no significant difference in age between the best-outcome group and the comparison group of children without a history of clinically significant behavioral disturbance. Subjects in the best-outcome group had a mean age of 12.42 years (range 10.0 to 16.25) versus 12.92 years (range 9.0 to 15.17) for the nonclinical comparison group. Scores on the WISC-R and clinical rating scale were obtained for all subjects; 1 experimental subject and 2 nonclinical comparison subjects were missing Vineland scores, and 2 experimental subjects and 1 nonclinical comparison subject were missing Personality Inventory for Children scores. Both the Vineland and Personality Inventory for Children were completed by parents. In cases where these scores were not obtained, the parents had declined to participate.

On the measures that provide standardized scores, the functioning of the best-outcome subjects was measured most precisely by comparing the best-outcome group against the test norms. Therefore, this analysis is of primary interest. Data for the nonclinical comparison group are mainly useful in confirming that the assessment procedures were valid and in providing a contrast group for the one measure without norms, the Clinical Rating Scale. For the nonclinical comparison group, it will suffice to summarize the results as follows: On the WISC-R this group had mean IQs of 116 Verbal, 118 Performance, and 119 Full-Scale. On the Vineland the group obtained mean standard scores of 102 Communication, 100 Daily Living Skills, 102 Socialization, and 101 Composite. The mean scale score on the Personality Inventory for Children was 49. Thus, the nonclinical comparison group displayed above-average or average functioning across all areas that were assessed.

The next section is focused on the functioning of the best-outcome group on IQ, adaptive and maladaptive behavior, and personality measures and contrasts the best-outcome subjects with the comparison subjects on the Clinical Rating Scale.

Intellectual Functioning. Table 2 presents the IQ data for each subject in the best-outcome group and the mean scores for the group. This table shows that, as a whole, the 9 best-outcome subjects performed well on the WISC-R. Their IQs placed them in the high end of the normal range, about two thirds of an *SD* above the mean. Their Full-Scale IQs ranged from 99 to 136.

Subjects' scores were evenly distributed across a range from 80 to 125 on Verbal IQ and from 88 to 138 on Performance IQ. The subjects averaged 3 points higher on Performance IQ than Verbal IQ. Two of them (J. L. and A. G.) had at least a 20-point difference

Pet. Reh. App.15

Table 2
WISC-R Scores of the Best-Outcome Subjects

	Verbal					Performance					WISC-R IQ		
Subject	Infrm	Simil	Arith	Vocab	Compr	PicC	PicA	BlkD	ObjA	Cod	VIQ	PIQ	Full
R.S.	12	12	13	9	11	10	9	13	12	11	106	108	106
M.C.	17	19	11	14	10	12	16	19	19	11	125	138	136
M.M.	14	13	10	14	11	12	11	11	11	8	114	102	109
L.B.	12	16	11	13	15	7	12	17	17	19	119	131	128
J.L.	6	9	7	4	8	18	11	16	14	7	80	123	100
D.E.	9	17	8	10	15	13	9	12	9	17	98	114	105
A.G.	7	14	12	11	13	8	4	9	11	10	108	88	99
B.W.	12	11	10	10	9	7	10	9	11	10	102	95	99
B.R.	11	14	11	13	16	12	10	12	11	10	118	106	114
Mean	11.1	13.9	10.5	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	111

Note. Infrm = Information, Simil = Similarities, Arith = Arithmetic, Vocab = Vocabulary, Compr = Comprehension, PicC = Picture Completion, PicA = Picture Arrangement, BlkD = Block Design, ObjA = Object Assembly, Cod = Coding, VIQ = Verbal IQ, PIQ = Performance IQ, and Full = Full-Scale IQ.

between Verbal and Performance IQ.

On each subtest of the WISC-R, the mean for the general population is 10 (*SD* = 3). It can be seen from Table 2 that the best-outcome subjects scored highest on Similarities, Block Design, and Object Assembly. They scored lowest on Picture Arrangement and Arithmetic. Thus, the subjects consistently scored at or above average.

Adaptive and Maladaptive Behavior. Table 3 presents the data for the best-outcome group on the Vineland Adaptive Behavior Scales. It can be seen that the best-outcome group scored about average on the Composite Scale and on the subscales for Communication, Daily Living, and Socialization. However, Table 3 shows that some of the best-outcome subjects had marginal scores, including J. L., B. W., and M. M. Even so, all of the best-outcome subjects had Composite scores within the normal range.

As can be seen in Table 3, on the Maladaptive Behavior Scale (Parts I and II), the mean score for the best-outcome group indicated that, on average, these subjects did not display clinically significant levels of maladaptive behavior. Three of them scored in the clinically significant range versus one subject in the nonclinical comparison group, which had a mean of 7.7 on this scale.

Personality Functioning. The results of the Personality Inventory for Children are summarized in Table 4. The best-outcome subjects obtained valid profiles on the Personality Inventory for Children, as measured by the three validity scales (Lie, Frequency, and Defensiveness). As can be seen from the table, the subjects scored in the normal range across all scales. They tended to score highest on Intellectual-Screening, Psychosis, and Frequency. Intellectual-Screening assesses slow intellectual development, and Psychosis and Frequency assess unusual or strange behaviors. Only Intellectual-Screening was above the normal range, and this scale is affected by subjects' early history. For example, the scale contains statements such as "My child first talked before he (she) was two years old," which would be false for the best-outcome subjects regardless of their current level of functioning.

As Table 4 indicates, 4 best-outcome subjects had a single scale elevated beyond

Table 3
Scores on the Vineland Adaptive Behavior Scale for the Best-Outcome Subjects

	Adaptive behavior				Maladaptive
Subject	Com	DLS	Soc	Comp	behavior
R.S.	83	98	102	92	6
M.C.	119	93	86	98	16
M.M.	119	79	114	105	2
L.B.	107	108	112	108	4
J.L.	77	103	94	88	13
D.E.	93	81	82	80	15
A.G.	101	97	99	98	5
B.W.	83	74	105	83	9
B.R.	—	—	—	—	—
Mean	98	92	99	94	8.8

Note. Com = Communication, DLS = Daily Living Skills, Soc = Socialization, Comp = Adaptive Behavior Composite.

Pet. Reh. App.16

Table 4
T Scores on the Personality Inventory for Children for the Best-Outcome Subjects

	T score																	
Subject	Mean	<70	L	F	Def	Adj	Ach	I-S	Dvl	Som	Dep	Fam	Dlq	Wdr	Anx	Psy	Hyp	Soc
R.S.	56	1	49	54	43	61	53	76	49	44	69	47	46	69	60	65	46	64
M.C.	52	1	48	63	37	43	39	54	38	64	55	54	46	65	51	75	40	55
M.M.	49	0	42	54	43	50	42	64	48	58	48	55	46	47	53	46	54	36
L.B.	51	1	60	50	49	49	37	70	39	55	49	48	51	45	60	51	48	51
J.L.	70	9	42	84	37	85	77	94	65	78	88	65	61	69	78	76	52	72
D.E.	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
A.G.	51	0	38	45	49	57	48	39	53	51	49	69	40	55	55	55	49	63
B.W.	54	1	45	63	50	59	64	48	55	47	44	57	90	44	45	46	62	44
B.A.	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Mean	55	2	46	56	44	58	51	64	49	57	57	58	54	56	57	59	50	55

Note. Mean = mean elevation across all scales. L = Lie scale, F = Frequency, Def = Defensiveness, Adj = Adjustment, Ach = Achievement, I-S = Intellectual-Screening, Dvl = Development, Som = Somatic Concern, Dep = Depression, Fam = Family Relations, Dlq = Delinquency, Wdr = Withdrawal, Anx = Anxiety, Psy = Psychosis, Hyp = Hyperactivity, Soc = Social Skills.

the clinically significant range and a 5th (J. L.) had nine scales elevated, including the highest scores in the best-outcome group on Intellectual-Screening, Psychosis, and Frequency. Thus, this subject appeared to account for much of the elevation in scores on these scales. By comparison, there were 3 subjects in the nonclinical comparison group with at least one scale elevated.

Clinical Rating Scale. On this scale, 8 of the best-outcome subjects scored between 0 and 10, and the 9th (J. L.) scored 42. The mean was 8.8, with a standard deviation of 12.9. The nonclinical comparison subjects all scored between 0 and 5 (mean = 1.7, *SD* = 2.1). Because these *SD*s are unequal, we used a nonparametric statistic, a Mann-Whitney *U* test, revealing a significant difference between groups, $U = 19, p < .05$. Thus, the best-outcome subjects displayed more deviance than did the comparison subjects, but most of the deviance appeared to come from one subject, J. L.

Discussion

This study is a later and more extensive follow-up of two groups of young subjects with autism who were previously studied by Lovaas (1987): (a) an experimental group (n = 19) that had received very intensive behavioral treatment and (b) a control group (n = 19) that had received minimal behavioral treatment. In the present study we have reported data on these children at a mean age of 13 years for subjects in the experimental group and 10 years for those in the control group. The data were obtained from a comprehensive assessment battery.

The main findings from the test battery were as follows: First, subjects in the experimental group had maintained their level of intellectual functioning between their previous assessment at age 7 and the present evaluation at a mean age of 13, as measured by standardized intelligence tests. Their mean IQ was about 30 points higher than that of control subjects. Second, experimental subjects also displayed significantly higher levels of functioning than did control subjects on measures of adaptive behavior and personality. Third, in a particularly rigorous evaluation of the 9 subjects in the experimental group who had been classified as best-outcome (normal-functioning) in the earlier study (Lovaas, 1987), the test results consistently indicated that the subjects exhibited average intelligence and average levels of adaptive functioning. Some deviance from average was found on the personality test and the clinical ratings. However, this deviance appeared to derive from the extreme scores of one subject, J. L. (see Table 2, 3, and 4). This subject also had been removed from nonspecial education classes and placed in a class for children with language delays, and he obtained relatively low scores (about 80) on the Verbal section of the intelligence test and the Communication section of the measure of adaptive behavior. Thus, he no longer appeared to be normal-functioning. However, the remaining 8 subjects who had previously been classified as normal-functioning demonstrated average IQ, with intellectual performance evenly distributed across subtests, were able to hold their own in regular classes, did not show signs of emotional disturbance, and demonstrated adequate development of adaptive and social skills within the normal range. In addition, subjective clinical impressions of blind examiners did nòt discriminate them from children with no history of behavioral disturbance. These 8 subjects (42% of the experimental group) may be judged to have made major and enduring gains and may be described as "normal-functioning." By contrast, none of the control group subjects achieved such a favorable outcome, consistent with the poor prognosis for children with autism reported by other investigators (Freeman, Ritvo, Needleman, & Yokota, 1985).

In order to evaluate this outcome, we must pay close attention to whether or not our methodology was sound. The adequacy of our methodology is crucial because the outcome in the present study represents a major improvement over outcomes obtained in previous experimental studies on the treatment of children with autism (Rutter, 1985). The only reports of comparable outcomes have come from uncontrolled case studies (e.g., Bettelheim, 1967), and subsequent investigations have indicated that these case studies grossly overestimated the outcomes obtainable with the treatment that was provided. Similarly, reports of major gains in other populations, such as large IQ increases in children from impoverished backgrounds, also have been based on highly questionable evidence (Kamin, 1974; Spitz, 1986). Such reports have the potential to cause a great deal of harm by misleading consumers and professionals.

A detailed description of all the methodological safeguards that should be built into a treatment study is beyond the scop
the present report (see Kazdin, 1980; Ken
& Norton-Ford, 1982; Spitz, 1986). Howe
we note that we incorporated a large num
of methodological safeguards in both
original study (Lovaas, 1987) and the pres
investigation:

1. The experimental group and
control group received equivalent ass
ment batteries at intake and were found tc
very similar on a multitude of import
variables. Moreover, the number of con
group subjects who were predicted to achi
normal functioning, had they received int
sive treatment, was approximately equal
the number of experimental subjects w
actually did achieve normal functioning w
intensive treatment (Lovaas & Smith, 198
Thus, the subject assignment proced
yielded groups that were comparable pr
to treatment. This provided a strong indi
tion that the superior functioning of
experimental group after treatment wa
result of the treatment itself rather thar
biased procedure for assigning subjects
the experimental group.

2. All subjects remained in the groups
which they were assigned at intake. Onl
subjects dropped out, and they were
replaced. Therefore, the original compo
tion of the groups was essentially preserv

3. All subjects were independently
agnosed as autistic by PhD or MD clinicia
and there was high agreement on the dia
nosis between the independent clinicia
This provided evidence that subjects m
criteria for a diagnosis of autism.

4. Prior to treatment, these subje
appeared to be comparable to those dia
nosed as having autism in other resear
investigations. Evidence for this comes fro
the second control group that was incorp
rated into the initial treatment study. Th
group was evaluated by another resear
team (independent of ours), had similar I
at intake based on the same measures
intelligence that we used, yet showed simil
outcome data to those reported by oth
investigators. Additional evidence can

derived from the similarity of our intake data to data reported by other investigators (Lovaas et al., 1989). For example, although Schopler and his associates (Schopler, Short, & Mesibov, 1989) suggested that our sample had a higher mean IQ than did other samples of children with autism, their own data do not appear to differ from ours (Lord & Schopler, 1989). Thus, there is evidence that our subjects were a typical group of preschool-age children with autism rather than a select group of high-level children with autism who would have been expected to achieve normal functioning with little or no treatment.

5. The first control group, which received up to 10 hours a week of one-to-one behavioral treatment, did not differ at posttreatment from the second control group, which received no treatment from us. Both groups achieved substantially less favorable outcomes than did the experimental group. Because all groups were similar at pretreatment, this result confirms that our subjects had problems that responded only to intensive treatment rather than problems such as being noncompliant or holding back (masking an underlying, essentially average intellectual functioning that would respond to smaller-scale interventions).

6. Subjects' families ranged from high to low socioeconomic status, and, on average, they did not differ from the general population (Lovaas, 1987). Thus, although our treatment required extensive family participation, a diverse group of families was apparently able to meet this requirement.

7. The treatment has been described in detail (Lovaas et al., 1980; Lovaas & Leaf, 1981), and the effectiveness of many components of the treatment has been demonstrated experimentally by a large number of investigators over the past 30 years (cf. Newsom & Rincover, 1989). Hence, our treatment may be replicable, a point that is discussed in greater detail later.

8. The results of the present follow-up, which extended several years beyond discharge from treatment for most subjects, are an encouraging sign that treatment gains have been maintained for an extended period of time.

9. A wide range of measures was administered, avoiding overreliance on intelligence tests, which have limitations if used in isolation (e.g., bias resulting from teaching to the test, selecting a test that would yield especially favorable results, failing to assess other aspects of functioning such as social competence or school performance) (Spitz, 1986; Zigler & Trickett, 1978).

10. The use at follow-up of a normal comparison group, standardized testing, and blind rating allowed for an objective, detailed, and quantifiable assessment of treatment effectiveness. A particularly rigorous assessment was given to those subjects who showed the most improvement.

Taken together, these safeguards provide considerable assurance that the favorable outcome of the experimental subjects can be attributed to the treatment they received rather than to extraneous factors such as improvement that would have occurred regardless of treatment, biased procedures for selecting subjects or assigning them to groups, or narrow or inappropriate assessment batteries.

Despite the numerous precautions that we have taken, several concerns may be raised about the validity of the results. Perhaps the most important is that the assignment to the experimental or control group was made on the basis of therapist availability rather than a more arbitrary procedure such as alternating referrals (assigning the first referral to the experimental group, the second to the control group, the third to the experimental group, and so forth). However, it seems unlikely that the assignment was biased in view of the pretreatment data we have presented on the similarity between the experimental and control groups. On the other hand, we do not know as yet whether there exists a pretreatment variable that does predict outcome but was not among the 19 we chose, yet could have discriminated between groups. In an earlier publication (Lovaas et al., 1989), we responded in some detail to the concern about subject assignment as well as other possible problems associated with the original study. There are certain additional questions that may be raised by this follow-up investigation:

1. The experimental group was older than the control group at the time of this follow-up evaluation. We explained this finding earlier and noted that data analyses indicated that it was unlikely that this age difference reflected a bias in subject assignments.

2. The follow-up assessments for 17 of the lower functioning subjects in this study were conducted by staff members from our Project, who could have biased the test results. However, as noted previously, a check revealed no evidence of such a bias.

3. The Clinical Rating Scale, based on an interview with subjects who had been classified as normal-functioning in the original study, has no norms or data on reliability and validity. However, we regard the interview simply as an extra check on whether the examiners detected residual signs of autism or other behavior problems that were somehow overlooked in the three other (well-standardized) measures in the study and their 30 subscales. We do not regard the interview as an instrument that by itself yields conclusive results. No other interview that suited our purposes currently exists. In future investigations, we plan to use an interview that Michael Rutter and his associates are now developing for the purpose of detecting of residual signs of autism in individuals with average intelligence.

4. As in most long-term follow-up studies, we had some missing data. However, there is no evidence that the missing data would have changed the overall results.

5. In our analysis of the best-outcome group, we noted that the group averages deviated from "normal" on one subscale of the Personality Inventory for Children and on the Clinical Rating Scale. We then attributed this deviance to the extreme scores of one subject rather than to general problems within this group. We recognize that group averages are seldom interpreted this way. However, as statisticians and methodologists have pointed out (e.g., Barlow & Hersen, 1984), there are many times when group averages represent the performance of few or no subjects within the group. This was one of those times, as is clearly shown by the data on individual subjects (Tables 2, 3, and 4). Deviance was found almost exclusively in one subject, not evenly distributed across all subjects, and we have presented the results accordingly.

The most important void for research to fill at this time is replication by independent investigators who employ sound methodologies. Given the objective assessment instruments that we used and the detailed description that we have provided of the treatment (Lovaas et al., 1980), such a replication should be possible. However, the treatment is complex and to replicate it properly, an investigator probably needs to possess (a) a strong foundation in learning theory research; (b) a detailed knowledge of the treatment manual we used; (c) a supervised practicum of at least 6 months in one-to-one work with clients who have developmental delays, emphasizing discrimination learning and building complex language; and (d) a commitment to provide 40 hours of one-to-one treatment to client per week, 50 weeks per year, for at least 2 years. Our best-outcome subjects all required a minimum of 2 years of intensive treatment to achieve average levels of functioning (another indication that those subjects had pervasive disabilities and were not merely noncompliant).

A second void to fill concerns the majority of children who did not benefit to the point of achieving normal functioning with intensive treatment. Perhaps an earlier start in treatment would have been all that was needed to obtain favorable outcomes with many of these children. More pessimistically, perhaps such children require new and different interventions that have yet to be discovered and implemented. In any case, it is essential to develop more appropriate services for these children.

Finally, a rather speculative but promising area for research is to determine the extent to which early intervention alters neurological structures in young children with autism. Autism is almost certainly the result of deficits in such neurological structures (Rutter & Schopler, 1987). However, laboratory studies on animals have shown that alterations in neurological structure are quite possible as a result of changes in the environment in the first years of life (Sirevaag & Greenough, 1988), and there is reason to believe that alterations are also possible in young children. For example, children under 3 years of age overproduce neurons, dendrites, axons, and synapses. Huttenlocher (1984) hypothesized that, with appropriate stimulation from the environment, this overproduction might allow infants and preschoolers to compensate for neurological anomalies much more completely than do older children. Caution is needed in generalizing from these findings on average children to early intervention with children with autism, particularly because the exact nature of the neurological anomalies of children with autism is unclear at present (e.g., Rutter & Schopler, 1987). Nevertheless, the findings suggest that intensive early intervention could compensate for neurological anomalies in such children. Finding evidence for such compensation would help explain why the treatment in this study was effective. More generally, it might contribute to an understanding of brain–behavior relations in young children.

References

Barlow, D. H., & Hersen, M. (1984). *Single case experimental design: Strategies for studying behavior change* (2nd ed.). New York: Pergamon Press.

Bettelheim, B. (1967). *The empty fortress.* New York: The Free Press.

DeMyer, M. K., Hingtgen, J. N., & Jackson, R. K. (1981). Infantile autism reviewed: A decade of research. *Schizophrenia Bulletin, 7,* 388–451.

Dunn, L. M. (1981). *Peabody Picture Vocabulary Test-Revised.* Circle River, MN: American Guidance Service.

Freeman, B. J., Ritvo, E. R., Needleman, R., & Yokota, A. (1985). The stability of cognitive and linguistic parameters in autism: A 5-year study. *Journal of the American Academy of Child Psychiatry, 24,* 290–311.

Huttenlocher, P. R. (1984). Synapse elimination and plasticity in developing human cerebral cortex. *American Journal of Mental Deficiency, 88,* 488–496.

Kamin, L. J. (1974). *The science and politics of I.Q.* New York: Wiley.

Kanner, L. (1971). Follow-up study of 11 autistic children originally reported in 1943. *Journal of Autism and Childhood Schizophrenia, 1,* 119–145.

Kazdin, A. (1980). *Research design in clinical psychology.* New York: Harper & Row.

Kendall, P. C., & Norton-Ford, J. D. (1982). Therapy outcome research methods. In P. C. Kendall & J. N. Butcher (Eds.), *Handbook of research methods in clinical psychology* (pp. 429–460). New York: Wiley.

Leiter, R. G. (1959). Part I of the manual for the 1948 revision of the Leiter International Performance Scale: Evidence of the reliability and validity of the Leiter tests. *Psychology Service Center Journal, 11,* 1–72.

Lord, C., & Schopler, E. (1989). The role of age at assessment, developmental level, and test in the stability of intelligence scores in young autistic children. *Journal of Autism and Developmental Disorders, 19,* 483–499.

Lotter, V. (1978). Follow-up studies. In M. Rutter & E. Schopler (Eds.), *Autism: A reappraisal of concepts and treatment.* London: Plenum Press.

Lovaas, O. I. (1987). Behavioral treatment and normal educational and intellectual functioning in young autistic children. *Journal of Consulting and Clinical Psychology, 55,* 3–9.

Lovaas, O. I., Ackerman, A. B., Alexander, D., Firestone, P., Perkins, J., & Young, D. (1980). *Teaching developmentally disabled children: The me book.* Austin, TX: Pro-Ed.

Lovaas, O. I., Koegel, R. L., Simmons, J. Q., & Long, J. S. (1973). Some generalization and follow-up measures on autistic children in behavior therapy. *Journal of Applied Behavior Analysis, 6,* 131–166.

Lovaas, O. I., & Leaf, R. L. (1981). *Five video* *tapes for teaching developmentally disabled children*. Baltimore: University Park Press.

Lovaas, O. I., & Smith, T. (1988). Intensive behavioral treatment with young autistic children. In B. B. Lahey & A. E. Kazdin (Eds.), *Advances in clinical child psychology* (Vol. 11, pp. 285–324). New York: Plenum Press.

Lovaas, O. I., Smith, T., & McEachin, J. J. (1989). Clarifying comments on the young autism study: Reply to Schopler, Short and Mesibov. *Journal of Consulting and Clinical Psychology, 57*, 165–167.

McEachin, J. J. (1987). *Outcome of autistic children receiving intensive behavioral treatment: Psychological status 3 to 12 years later.* Unpublished doctoral dissertation, University of California, Los Angeles.

Newsom, C., & Rincover, A. (1989). Autism. In E. J. Mash & R. A. Barkley (Eds.), *Treatment of childhood disorders* (pp. 286–346). New York: Guilford Press.

Rutter, M. (1970). Autistic children: Infancy to adulthood. *Seminars in Psychiatry, 2*, 435–450.

Rutter, M. (1985). The treatment of autistic children. *Journal of Child Psychology & Psychiatry, 26*, 193–214.

Rutter, M., & Schopler, E. (1987). Autism and pervasive developmental disorders: Concepts and diagnostic issues. *Journal of Autism and Developmental Disorders, 17*, 159–186.

Schopler, E., Short, A., & Mesibov, G. (1989). Relation of behavioral treatment to "normal functioning": Comment on Lovaas. *Journal of Consulting and Clinical Psychology, 57*, 162–164.

Short, A., & Marcus, L. (1986). Psychoeducational evaluation of autistic children and adolescents. In S. S. Strichart & P. Lazarus (Eds.), *Psychoeducational evaluation of school-aged children with low-incidence disorders* (pp. 155–180). Orlando, FL: Grune & Stratton.

Simeonnson, R. J., Olley, J. G., & Rosenthal, S. L. (1987). Early intervention for children with autism. In M. J. Guralnick & F. C. Bennett (Eds.), *The effectiveness of early intervention for at-risk and handicapped children* (pp. 275–296). Orlando, FL: Academic Press.

Sirevaag, A. M., & Greenough, W. T. (1988). A multivariate statistical summary of synaptic plasticity measures in rats exposed to complex, social and individual environments. *Brain Research, 441*, 386–392.

Sparrow, S. S., Balla, D. A., & Cicchetti, D. V. (1984). *Interview Edition Survey Form Manual.* Circle Pines, MN: American Guidance Service.

Spitz, H. H. (1986). *The raising of intelligence.* Hillsdale, NJ: Erlbaum.

Waterhouse, L., & Fein, D. (1984). Developmental trends in cognitive skills for children diagnosed as autistic and schizophrenic. *Child Development, 55*, 236–248.

Wechsler, D. (1974). *Manual for the Wechsler Intelligence Scale for Children-Revised.* New York: Psychological Corp.

Wirt, R. D., Lachar, D., Klinedinst, J. K., & Seat, P. D. (1977). *Multidimensional descriptions of child personality: A manual for the Personality Inventory for Children.* Los Angeles: Western Psychological Services.

Zigler, E., & Trickett, P. K. (1978). IQ, social competence, and evaluation of early childhood intervention programs. *American Psychologist, 33*, 789–798.

Received: 5/15/91; first decision: 10/16/91; accepted: 1/23/92.

Intensive Behavioral Treatment for Children With Autism: Four-Year Outcome and Predictors

Glen O. Sallows and Tamlynn D. Graupner
Wisconsin Early Autism Project (Madison)

Abstract

Twenty-four children with autism were randomly assigned to a clinic-directed group, replicating the parameters of the early intensive behavioral treatment developed at UCLA, or to a parent-directed group that received intensive hours but less supervision by equally well-trained supervisors. Outcome after 4 years of treatment, including cognitive, language, adaptive, social, and academic measures, was similar for both groups. After combining groups, we found that 48% of all children showed rapid learning, achieved average posttreatment scores, and at age 7, were succeeding in regular education classrooms. Treatment outcome was best predicted by pretreatment imitation, language, and social responsiveness. These results are consistent with those reported by Lovaas and colleagues (Lovaas, 1987; McEachin, Smith, & Lovaas, 1993).

Behavioral approaches for addressing the delays and deficits common in autism have been recognized by many as the most effective treatment methods to date (Green, 1996; Maine Administrators of Service for Children With Disabilities, 2000; New York State Department of Health, 1999; Schreibman, 1988; Smith, 1993). The intervention developed at UCLA in the 1960s and 1970s is perhaps the best known and best documented (e.g., Dawson & Osterling, 1997; Green, 1996; Smith, 1993). Building on earlier research (e.g., Lovaas, Koegel, Simmons, & Long, 1973), Lovaas and staff of the UCLA Young Autism Project (1970 to 1984) began treatment with children under 4 years of age using a curriculum emphasizing language development, social interaction, and school integration skills. After 2 to 3 years of treatment, 47% of the experimental group (9 of 19 children) versus 2% of the comparison group (1 of 40 children) were reported to have achieved "normal functioning" (Lovaas, 1987; McEachin et al., 1993).

These findings demonstrated that many children with autism could make dramatic improvement, even achieve "normalcy," and many researchers now agree that intensive behavioral treatment can result in substantial gains for a large proportion of children (e.g., Harris, Handleman, Gordon, Kristoff, & Fuentes, 1991; Mundy, 1993). However, the UCLA findings also created considerable controversy, and the studies were criticized on methodological and other grounds (e.g., Gresham & MacMillan, 1998; Schopler, Short, & Mesibov, 1989). One criticism was that the UCLA group used the term *recovered* to describe children who had achieved IQ in the average range and placement in regular classrooms. Mundy (1993) suggested that children diagnosed with high functioning autism might achieve similar outcomes and pointed out that several of the recovered children in the follow-up study of the UCLA children at age 13 (McEachin et al., 1993) had clinically significant scores on some behavioral measures. The UCLA team responded by noting that (a) evaluators blind to background information had not identified the recovered children as different from neurotypical children and (b) a few elevated scores may not imply abnormality because several of the neurotypical peers had them as well (Smith, McEachin, & Lovaas, 1993). Questions were also raised regarding whether or not the UCLA results could be fully replicated without the use of aversives, which were part of the UCLA protocol, but are not acceptable in most communities (Schreibman, 1997). Some have questioned the feasibility of implementing the program without the resources of a university research center to train and supervise treatment staff (Sheinkopf & Siegel, 1998) and to help defray the cost of the program, which, due to the many hours of weekly treatment, can exceed $50,000 per year (although it has been argued that the cost of not providing treatment may be much greater over time: Jacobson, Mulick, & Green, 1998). Finally, because only about half of the children showed marked gains, the need for predictors to determine which children will benefit has been raised (Kazdin, 1993). Lovaas and his colleagues responded to these and other criticisms (Lovaas, Smith, & McEachin, 1989; Smith et al., 1993; Smith & Lovaas, 1997), but agreed with others that replication and further research were necessary.

There have now been several reports of partial replication without using aversives (Anderson, Avery, Di Pietro, Edwards, & Christian, 1987; Birnbrauer & Leach, 1993; Eikeseth, Smith, Jahr, & Eldevik, 2002; Smith, Groen, & Wynn, 2000). Most found, as did Lovaas and his colleagues, that a subset of children showed marked improvement in IQ. Although fewer children reached average levels of functioning, the treatment provided in these studies differed from the UCLA model in several ways (e.g., lower intensity and duration of treatment, different sample characteristics and curriculum, and less training and supervision of staff).

Anderson et al. (1987) provided 15 hours per week for 1 to 2 years (parents provided another 5 hours) and found that 4 of 14 children achieved an IQ over 80 and were in regular classes, but all needed some support. Birnbrauer and Leach (1993) provided 19 hours per week for 1.5 to 2 years and found that 4 of 9 children achieved an IQ over 80 (classroom placement was not reported), but all had poor play skills and self-stimulatory behaviors. The authors noted, however, that their treatment program had not addressed these areas. Smith et al. (2000) provided 25 hours per week for 33 months and reported that 4 of 15 children achieved an IQ over 85 and were in regular classes, but one had behavior problems. The authors noted that their sample had an atypically high number of mute children, 13 of 15, considerably higher than the commonly cited figure of 50% (Smith & Lovaas, 1997), and they hypothesized that this was the reason for the relatively low number of children functioning in the average range following treatment. Eikeseth et al. (2002) provided 28 hours per week for 1 year. In their sample, 7 of 13 children with pretreatment IQ over 50 achieved IQ over 85 and were in regular classes with some support. Data beyond the first year have not yet been reported.

Four groups of investigators discussed results based on behavioral treatment in classroom settings, which typically include a mix of 1:1 treatment and group activities, so that time in school may not be comparable to hours reported in home-based studies. Following 4 years of treatment, Fenske, Zalenski, Krantz, and McClannahan (1985) found that 4 of 9 children were placed in regular classes. However, neither pre–posttreatment test scores nor amount of support in school were reported. Harris et al. (1991) provided 5.5 hours per day in class and instructed parents to provide an additional 10 to 15 hours at home (no data were collected on actual hours parents provided). After 1 year of treatment, 6 of 9 children achieved IQ over 85, but were still in classes for students with learning disabilities. A later report (Harris & Handleman, 2000) found that 9 of 27 children achieved IQ over 85 and were placed in regular classes (time in treatment was not reported), but most required some support. Meyer, Taylor, Levin, and Fisher (2001) provided 30 hours of class time per week for at least 2 years and reported that 7 of 26 children were placed in public schools after 3.5 years of treatment, but 5 required support services. Pre–post IQ was not reported. Romanczyk, Lockshin, and Matey (2001) provided 30 hours of class time per week for 3.3 years and reported that 15% of the children were discharged to regular classrooms. No information on posttreatment test scores or the need for supports was provided.

In two studies researchers examined the effects of behavioral treatment for children with low pretreatment IQ. Smith, Eikeseth, Klevstrand, and Lovaas (1997) provided children who had pretreatment IQ less than 35 (M = 28) with 30 hours per week for 35 months and reported an increase in IQ of 8 points (3 of 11 children achieved increases of over 15 points) and 10 of 11 achieved single-word expressive speech. Eldevik, Eikeseth, Jahr, and Smith (in press) provided children who had an average pretreatment IQ of 41 with 22 hours per week of 1:1 treatment for 20 months and reported an increase in IQ of 8 points and an increase in language standard scores of 11 points.

In three studies researchers examined results of behavioral treatment provided by clinicians working outside university settings in what has been termed *parent-managed treatment* because parents implement treatment designed by a *workshop consultant,* who supervises less frequently (e.g., once every 2 to 4 months) than the supervision that occurs in programs supervised by a local autism treatment center (e.g., twice per week). Sheinkopf and Siegel (1998) reported results for children who received 19 hours of treatment per week for 16 months supervised by three local providers. Six of 11 children achieved IQ over 90 and 5 were in regular classes, but still had residual symptoms. However, these children may not be comparable to high achievers in other studies because intelligence tests included the Merrill-Palmer, a measure of primarily nonverbal skills, known to yield scores about 15 points higher than standard intelligence tests that include both verbal and nonverbal scales. In the second study, Bibby, Eikeseth, Martin, Mudford, and Reeves (2002) described results for children who received 30 hours of treatment per week (range = 14 to 40) for 32 months (range = 17 to 43) supervised by 25 different consultants, who saw the children several times per year (median = 4, range = 0 to 26). Ten of 66 children achieved IQ over 85, and 4 were in regular classes without help. However, as the authors noted, their sample was unlike UCLA's in several ways: 15% had a pretreatment IQ under 37, 57% were older than 48 months, many received fewer than 20 hours per week, 80% of the providers were not UCLA-trained, and no child received weekly supervision. Weiss (1999) reported the results of a study in which children did receive high hours: 40 hours of treatment per week for 2 years. She saw each child every 4 to 6 weeks, reviewed videos of their performance every 2 to 3 weeks, and spoke with parents weekly. Following treatment, 9 of 20 children achieved scores on the Vineland Applied Behavior Composite (ABC) of over 90, were placed in regular classes, and had scores on the Childhood Autism Rating Scale in the nonautistic range (under 30). No pre- or posttreatment IQ data were reported.

Several researchers have described pretreatment variables that seem to predict (are highly correlated with) later outcome. Although findings have not always been consistent, the most commonly noted predictors have been IQ (Bibby et al., 2002; Eikeseth et al., 2002; Goldstein, 2002; Lovaas, 1987; Newsom & Rincover, 1989), presence of imitation ability (Goldstein, 2002; Lovaas & Smith, 1988; Newsom & Rincover, 1989; Weiss, 1999), language (Lord & Paul, 1997; Venter, Lord, & Schopler, 1992), younger age at intervention (Bibby et al., 2002; Fenske et al., 1985; Goldstein, 2002; Harris & Handleman, 2000), severity of symptoms (Venter et al., 1992), and social responsiveness or "joint attention" (Bono, Daley, & Sigman, 2004; L. Koegel, Koegel, Shoshan, & McNerney, 1999; Lord & Paul, 1997).

Multiple regression has been used to determine combinations of pretreatment variables with strong relationships with outcome. Goldstein (2002) reported that verbal imitation plus IQ plus age resulted in an R^2 of .78 with acquisition of spoken language. Rapid learning during the first 3 or 4 months of treatment has also been associated with positive outcome (Lovaas & Smith, 1988; Newsom & Rincover, 1989; Weiss, 1999). Weiss reported that rapid acquisition of verbal imitation plus nonverbal imitation plus receptive instructions resulted in an R^2 of .71 with Vineland ABC and .73 with Childhood Autism Rating Scale scores 2 years later.

We designed the present study to examine several questions. Can a community-based program operating without the resources, support, or supervision of a university center, implement the UCLA program with a similar population of children and achieve similar results without using aversives? Do significant residual symptoms of autism remain among children who achieve posttreatment test scores in the average range? Can pretreatment variables be identified that accurately predict outcome? We also examined the comparative effectiveness of a less costly parent-directed treatment model.

Method

Participants

Researchers at the Wisconsin site worked in collaboration with and observed the guidelines set by the National Institutes of Mental Health (NIMH) for Lovaas' Multi-Site Young Autism Project. Children were recruited through local birth to three (special education) programs. All children were screened for eligibility according to the following criteria: (a) age at intake between 24 and 42 months, (b) ratio estimate (mental age

[MA] divided by chronological age [CA]) of the Mental Development Index of 35 or higher (the ratio estimate was used because almost all children scored below the lowest Mental Development Index of 50 from the Bayley Scales of Infant Development Second Edition (Bayley, 1993), (c) neurologically within "normal" limits (children with abnormal EEGs or controlled seizures were accepted) as determined by a pediatric neurologist (no children were excluded based on this criterion), and (d) a diagnosis of autism by independent child psychiatrists well known for their experience and familiarity with autism. All children also met the criteria for autism based on the Diagnostic and Statistical Manual of Mental Disorders—DSM-IV (American Psychiatric Association, 1994) and the Autism Diagnostic Interview-Revised (Lord, Rutter, & LeCouteur, 1994), both administered by a trained examiner. There were no parental criteria for involvement beyond agreeing to the conditions in the informed consent document, one of which was accepting random assignment to treatment conditions. The parents of all screened children agreed to participate, and none dropped out upon learning of their group assignment, minimizing bias in selection of participants and group composition.

Thirteen children began treatment in 1996, 11 in 1997, and 14 in 1998–1999. The last group had not completed treatment when the data from the first two groups were analyzed, and their data will be reported in a subsequent paper. The 24 children admitted during the first 2 years were 19 boys and 5 girls. One girl was placed in foster care after 1 year of treatment, and the foster parents did not wish to continue treatment for her. Her data were, therefore, excluded from the analysis. The remaining 23 children had completed 4 years of treatment (or had "graduated" earlier) at the time of this report, although 1 child switched to another provider of behavioral treatment after 1 year.

Design

In accordance with the research protocol approved by NIMH, we matched children on pretreatment IQ (Bayley MA divided by CA). They were randomly assigned by a UCLA statistician to the clinic-directed group (n = 13), replicating the parameters of the UCLA intensive behavioral treatment (Lovaas, 1987) or to the parent-directed group (n = 10), intended to be a less intensive alternative treatment.

All children received treatment based on the UCLA model. Parents in both groups were instructed to attend weekly team meetings and were encouraged to extend the impact of treatment by practicing newly learned material with their child throughout the day. Demographic information as well as hours of treatment and supervision are shown in Table 1. Children averaged 33 to 34 months of age at pretest and began treatment at 35 to 37 months. Children in the clinic-directed

Table 1. Demographic Information and Hours of Service by Group

Descriptor	Clinic-directed	Parent-directed
Boys, girls	11, 2	8, 2
One-parent families	0 of 13	1 of 10
Income		
Median ($)	62,000	59,000
(Range)	(35–100+)	(30–100+)
Education (BA)		
Mothers	9 of 12	9 of 10
Fathers	10 of 12	6 of 9
Siblings (mean)	2	2
No. nonverbal (%)	8/13 (62)	2/10 (20)
Age (months) (*SD*)		
Pretest	33.23 (3.89)	34.20 (5.06)
Treatment	35.00 (4.86)	37.10 (5.36)
Posttest	83.23 (8.92)	82.50 (6.61)
1:1 hours per week (*SD*)		
Year 1	38.60 (2.91)	31.67 (5.81)
Year 2	36.55 (3.83)	30.88 (4.04)
Senior therapist	6–10 hrs per week 3, 2- to 3-hr sessions	6 hrs per month 1, 3-hr session per 2 wks
Team meetings	1 hr per week	1 hr per 1 or 2 weeks
Progress review	1 hr per wk for 1–2 years then 1 hr per 2 months	1 hr every other month

Note. The 1:1 hours for parent-directed children excludes one child who received 14 hours per week.

group were to receive 40 hours per week of direct treatment. The actual average was 39 during Year 1 and 37 during Year 2, with gradually decreasing hours thereafter as children entered school. Parents in the parent-directed group chose the number of weekly treatment hours provided by therapists. The average was 32 hours during Year 1 and 31 during Year 2, with the exception of one family that chose to have 14 hours both years. Because the parent-directed children as a group received more intensive treatment than was provided in most previous studies, only 6 to 7 hours less than the clinic-directed group, our ability to examine the effect of differences in treatment intensity was limited.

The clinic-directed group received 6 to 10 hours per week of in-home supervision from a senior therapist and weekly consultation by the senior author or clinic supervisor. Parent-directed children received 6 hours per month of in-home supervision from a senior therapist (typically a 3-hour session every other week) and consultation every 2 months by the senior author or clinic supervisor.

Direct treatment staff, referred to as *therapists,* were hired by Wisconsin Early Autism Project staff members for both the clinic- and parent-directed groups. Funding for 35 hours of 1:1 treatment per week was provided through the Wisconsin Medical Assistance program. Treatment hours in excess of 35 were funded through project funds.

Measures

We used the Bayley Scales of Infant Development, Second Edition, to determine pretreatment IQ. In addition we used the Merrill-Palmer Scale of Mental Tests (Stutsman, 1948), an older test of intelligence recommended for use with nonverbal children (Howlin, 1998), as a measure of nonverbal intelligence but not pre- or posttreatment IQ. We employed the Reynell Developmental Language Scales (Reynell & Gruber, 1990) to assess language ability, because of its extensive psychometric data for preschool-age children, and the Vineland Adaptive Behavior Scales (Sparrow, Balla, & Cicchetti, 1984) to measure adaptive functioning. Subscales of the Vineland assess Communication in Daily Life, Daily Living Skills, and Social Skills. Information regarding developmental history (including loss of language and other skills), use of supplemental treatments and pretreatment presence of functional speech was gathered from parent interviews, reports from other professionals, and direct observation.

Follow-up testing was administered annually for 4 years. As children grew older or became too advanced for the norms of pretreatment tests, we used other age-appropriate tests. Cognitive functioning of older children was assessed using Wechsler tests for 20 children—Wechsler Preschool and Primary Scale of Intelligence-Revised-WPPSI (Wechsler, 1989); Wechsler Intelligence Scale for Children—WISC-III (Wechsler, 1991)—and the Bayley II for 3 children. Although we assessed nonverbal cognitive functioning, it was not used as a measure of posttreatment IQ; we employed the Leiter-R for 11 children (Roid & Miller, 1995, 1997) and the Merrill-Palmer for 12 children. Language was measured using the Clinical Evaluation of Language Fundamentals, Third Edition—CELF III (Semel, Wiig, & Secord, 1995) for 11 children and the Reynell for 12 children. We administered the Vineland to all children for assessment of adaptive functioning.

To assess posttreatment social functioning, we readministered the Autism Diagnostic Interview-Revised and used the Personality Inventory for Children (Wirt, Lachar, Klinedinst, & Seat, 1977), which was completed by parents of all 23 children after 3 years of treatment. After 4 years of treatment, when the children were approximately 7 years old, parents and teachers completed the Child Behavior Checklist (Achenbach, 1991a, 1991b) and Vineland for all 23 children. Bierman and Welsh (1997) noted that "teacher ratings are superior to those of other informants and provide information regarding peer interaction and group acceptance that are closest to those of peers" (p. 348). Information was obtained from teachers on classroom placement (regular, regular with modified curriculum, partial special education [e.g., pullout/resource room or full special education], and supportive/therapeutic services [e.g., classroom aide, speech or occupational therapy]) when the children were 7 years old. We used the Woodcock-Johnson III Tests of Achievement (Woodcock, McGrew, & Mather, 2001) to measure academic skills of children placed in regular education classes at age 7.

The second author administered the pretreatment assessment battery prior to children being assigned to treatment groups. She received training in assessment at UCLA and met criterion for satisfactory intertester reliability. One fourth of the children in the current study were tested prior to treatment by unaffiliated community psychologists. These children earned a ratio IQ of 50.3 on the Bayley administered by the independent psychologists and 47.3 from the Wisconsin Project evaluator. The mean absolute difference was three points, r = .83, indicating absence of bias by the Wisconsin Project evaluator. Children who achieved IQs of 85 or higher at annual follow-up testing were thereafter referred for assessment by psychologists who had extensive experience testing children with autism at hospital-based assessment clinics that were not affiliated with the Wisconsin Project. These psychologists, who were unaware of group assignment or length of time in treatment, used the tests listed above. Follow-up testing of most children whose IQ remained delayed was conducted by the second author to reduce cost.

One experimental assessment procedure, the Early Learning Measure developed at UCLA (Smith, Buch, & Gamby, 2000) was administered to measure the rate of acquisition of skills during the first several months of treatment. Every 3 weeks for 3 months leading up to the beginning of treatment and for 6 months after treatment started, the same list of 40 items (10 each of verbal imitation, nonverbal imitation, following verbal instructions, and expressive object labeling), which was known only to the experimenter, was presented to the children. Two sets of scores were obtained from the Early Learning Measure. The first was the number of items the child performed correctly prior to the onset of treatment. The second set of scores was the number of weeks required for the child to learn 90% of the verbal imitation items once treatment had begun, thereby providing a measure of the child's rate of acquisition. This criterion was selected based on earlier research with the Early Learning Measure, which suggested the predictive validity of rapid acquisition of verbal imitation (Lovaas & Smith, 1988).

Treatment Procedure

The treatment procedure and curriculum were those initially described by Lovaas (Lovaas et al., 1981), except that no aversives were used, with the addition of procedures supported by subsequent research (e.g., R. Koegel & Koegel, 1995), which have been widely disseminated (e.g., Maurice, Green, & Luce, 1996). Positive interactions were built by engaging in favorite activities and responding to the gestures used by each child to indicate desires. Anticipation of success and motivation to attend were increased by employing brief, standard instructions and tasks requiring only visual attending (e.g., matching), using familiar materials (e.g., the child's own ring stacker), prompting success (physically assisting him or her to place a ring on the pole if a demonstration was not sufficient), presenting only two or three trials at a time, and reinforcing each response immediately with powerful reinforcers (e.g., edibles, physical play, or enthusiastic proclamations of success (such as "Fantastic!"). Between these brief (initially 30 seconds long) learning periods, staff members played with the children to keep the process more like play than work, generalize learned material into more natural settings, and continue to build social responsiveness.

Receptive language was generally targeted before expressive language. We used familiar instructions where success was easily prompted, such as "sit down" or "come here." Expressive language began with imitation training, first nonverbal then vocal imitation, beginning with single sounds and gradually progressing to words. Requesting was taught as early as possible, initially using nonverbal strategies if necessary (e.g., gesturing, signing, or the Picture Exchange Communication System—PECS (Bondy & Frost, 1994), in order to reduce frustration (Carr & Durand, 1985) and increase the child's frequency of communicative initiations (Hart & Risley, 1975). Children who showed more modest gains in treatment, referred to as *visual learners* by the UCLA group, denoting difficulty in processing language, took longer to acquire verbal imitation and language.

Having learned many labels, children were taught more complex concepts and skills, such as categorization and speaking in full sentences. Social interaction and cooperative play were taught as part of the in-home program, expanding from playing with staff, to playing with siblings, and then peers for up to 2 hours per day (this was more successful with the subgroup of rapidly learning children). As the children acquired social skills, they began mainstream (as opposed to special education) preschool, usually for just 1 or 2 half-days (2.5 hours each) per week. A trained *shadow* (one of the home treatment team members) initially accompanied the child to assist with attending to the teacher's instructions, joining others on the playground, and noting social errors to be addressed in 1:1 sessions at home.

Those children who progressed at a rapid pace were taught the beginnings of inferential thought (e.g., "Why does he feel sad?"). Social and conversation skills, such as topic maintenance and asking appropriate questions, were taught using role-playing (e.g., Jahr, Eldevik, & Eikeseth, 2000), video modeling (Charlop & Milstein, 1989), social stories (Gray, 1994), straightforward discussion of social rules and etiquette, and in-vivo prompting.

Academic skills were also targeted, raising the level of proficiency of rapidly learning children to first grade levels. Common classroom rules and school "survival skills" (e.g., responding to group instructions and raising one's hand to be called on–Dawson & Osterling, 1997) were taught through "mock school" exercises with several peers at home.

Staff training. Therapists were at least 18 years old, had completed a minimum of 1 year of college, and were screened for prior police contacts. Therapists received 30 hours of training, which included a minimum of 10 hours of one-to-one training and feedback while working with their assigned child. Each therapist worked at least 6 hours per week (usually three 2-hour shifts) and attended weekly or bi-weekly team meetings. Senior therapists had at least a 4-year college degree and experience consisting of 1 year as a therapist with at least two children, followed by an intensive 16-week internship program modeled after that at UCLA, for a total of 2,000 hours.

Treatment fidelity. Senior therapists and clinic-directed therapists were required to meet quality control criteria set at UCLA. This involved passing two tests. The first was a written test designed to assess knowledge of basic behavioral principles and treatment procedures described in *The Me Book* (Lovaas et al., 1981). Second, they were required to pass a videotaped review of their work (conducted by Tristram Smith, research director of the Multi-Site Project, who used the protocol described by R. Koegel, Russo, and Rincover, 1977). All senior therapists also received weekly supervision by the senior author.

Progress reviews, which the child, parents, and senior therapist attended, were held weekly for clinic-directed children and every 2 months for parent-directed children. At these reviews, the senior author or the UCLA-trained clinic supervisor observed the child's performance and recommended appropriate changes in the program. Both the senior author and clinic supervisor had met the UCLA criteria for Level Two Therapist, denoting sufficient experience and expertise in program implementation to work independent of supervision. The senior author had directed a behaviorally oriented inpatient unit for children with autism for 14 years and had trained at UCLA for 6 months. The clinic supervisor had a BA in psychology, 1 year of experience as a therapist, 2 years of full-time experience as a senior therapist, and had completed a 9-month internship at UCLA.

Data Analysis

Data analysis was carried out by a fourth year graduate student from the University of Wisconsin Department of Statistics, with consultation from a university research psychologist. We conducted an ANOVA with a least squares solution for unequal group size, used to examine treatment effects. To compare the clinic-directed and parent-directed groups, we used 2 × 2 ANOVAS (Clinic-Directed vs. Parent-Directed × Pre- vs. Posttest scores as repeated measures). An initial examination of pre–post IQ data showed that the distribution of scores was bimodal. As can be seen in Figure 1, children showed either rapid progress or more moderate progress, with no overlap between outcome distributions. This is consistent with earlier research (Birnbrauer & Leach, 1993; Howard, Sparkman, Cohen, Green, & Stanislaw, 2005; O. I. Lovaas, personal communication, August 27, 2003). Consequently, changes in scores for rapid learners and moderate learners were analyzed separately.

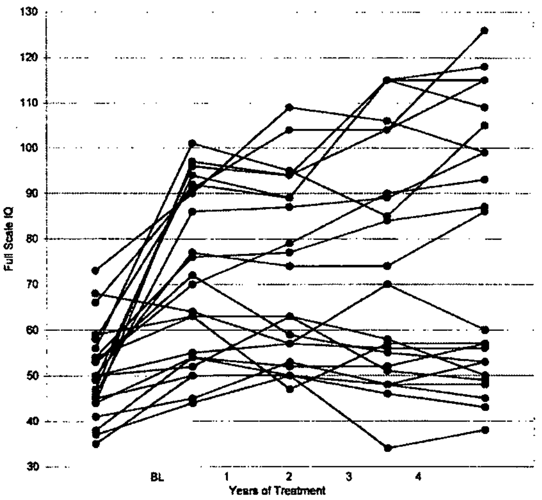

Figure 1. Changes in Full Scale IQ during 4 years of behavioral treatment.

In examining pretreatment scores of children who would later be identified as rapid learners, we found that those in the clinic-directed group had higher mean IQ (60.40, standard deviation [*SD*] = 8.31 compared to those in the parent-directed group (51.00, *SD* = 7.02), $t(9) = 1.84, p < .05$ (one tailed), Vineland scores (clinic-directed = 64.8, *SD* = 2.32; parent-directed = 59.83, *SD* =3.34), $t(9) = 2.31, p < .05$ (one tailed), and Verbal Imitation (clinic-directed = 3.88; parent-directed = 1.67), $W(4, 6) = 31, p = .03$ (Wilcoxon test). Because these pretreatment differences would interfere with clear interpretation of posttreatment differences between subgroups (e.g., clinic-directed vs. parent-directed rapid learners), these comparisons were omitted. We used linear and logistic regression (best subset selection approach–Hosmer, Jovanovic, & Lemeshow, 1989) to develop prediction models using pretreatment measures as predictors of 3-year outcome.

Results

The average Full Scale IQ for all 23 children increased from 51 to 76, a 25-point increase. Eight of the children achieved IQs of 85 or higher after 1 year of treatment (5 clinic-directed and 3 parent-directed), and 3 more reached this level after 3 to 4 years (3 parent-directed) for a total of 11, or 48%, of the 23 children. Children with higher pretreatment IQs were more likely to reach 4-year IQs in the average range (75% of children with IQs between 55 and 64 versus 17%, 1 of 6 children with IQs between 35 and 44).

As shown in Table 2, there were no significant differences between groups at pre- or posttest. Combining children in both groups, we found that pretest to posttest gains were significant for Full Scale IQ, $F(1, 21) = 18.77, p < .01$, Verbal IQ, $F(1, 18) = 13.39, p < .01$, Performance IQ, $F(1, 18) = 46.79, p < .01$, receptive language, $F(1, 21) = 9.18, p < .01$, Vineland Communication, $F(1, 21) = 7.57, p < .05$, Vineland Socialization, $F(1, 21) = 10.30, p < .01$, Autism Diagnostic Interview-Revised Social Skills, $F(1, 18) = 19.15, p < .01$, and Communication, $F(1, 18) = 41.19, p < .01$.

Rapid and Moderate Learners

A group of rapid learners showed much larger improvements than did moderate learners (analogous to the terms *best outcome* and *non-best outcome* used in UCLA reports). Figure 1 shows Full Scale IQs prior to treatment and over the next 4 years for all 23 children. Eleven of them (5 clinic-directed and 6 parent-directed) showed a large increase in IQ, from a mean of 55 prior to treatment to 104 after 4 years. These rapid learners represented 48% of all 23 children. The IQ of the remaining 12 children (8 clinic-directed and 4 parent-directed) did not show a significant increase, consistent with earlier UCLA reports (e.g., Smith et al., 2000).

Pre- and posttreatment scores of rapid and moderate learners are shown in Table 3. Rapid learners showed significant gains in all areas measured (i.e., Full Scale IQ, $F(1, 21) = 143.19, p < .01$, Verbal IQ, $F(1, 18) = 70.76, p < .01$, Performance IQ, $F(1, 18) = 165.27, p < .01$, Nonverbal IQ, $F(1, 19) = 16.69, p < .01$, Receptive Language, $F(1, 20) = 217.76, p < .01$, Expressive Language, $F(1, 20) = 77.76, p < .01$, and all Vineland subscales: Communication, $F(1, 21) = 147.07, p < .01$, Daily Living Skills ($F(1,21) = 20.50$, p < .01), Socialization, $F(1, 21) = 42.89, p < .01$, and Applied Behavior Composite, $F(1, 21) = 54.17, p < .01$). However, the rate of increase over time, skill areas, and children was not uniform. As can be seen in Figure 2, during the first year, Performance IQ of rapid learners rose to the average range (a 40-point increase, WPPSI-R), whereas Verbal IQ and Vineland Socialization scores rose to around 80 (a 25-point increase) and language scores (Reynell and Clinical Evaluation of Language Fundamentals) rose only to the 60s. Changes during the second year of treatment were comparatively modest, perhaps reflecting the effect of having acquired speech during the first year but still lacking more complex language. The rate of improvement increased again during the third and fourth years, and all scores increased to the average range.

The gradual decrease in the slope of the graphs in Years 3 and 4 is largely an artifact of increasing age and does not reflect a decrease in rate of MA growth, which, except for the large increase during Year 1, averaged 18 months per year throughout the study. This rate of growth in skills is necessary for children with pretreatment scores below 60 to "catch up" to peers. Although some writers have noted a rate of growth among treated children of 10 to 12 months per year, this is not enough for them to reach scores in the average range within just a few years (Howard et al., 2005), and the longer that children are delayed, the more skills they must learn to catch up.

Pet. Reh. App.30

Table 2. Pretreatment and Outcome Scores of Clinic- (CD) and Parent-Directed (PD) Groups

Measure/	Pretreatment		Posttreatment		ANOVA, combined
Group	Mean	*SD*	Mean	*SD*	groups, pre- vs. posttreatment (*df*)
Full Scale IQ					
CD	50.85	10.57	73.08	33.08	18.77 (1,21)**
PD	52.10	8.98	79.60	21.80	
Verbal IQ					
CD	—	—	78.00	33.48	13.39 (1,18)**
PD	—	—	76.30	26.66	
Perform IQ					
CD	—	—	84.90	25.86	46.79 (1,18)**
PD	—	—	90.70	20.72	
Nonverbal IQ					
CD	70.58	16.54	77.58	25.24	2.07 (1,21)
PD	82.67	14.94	89.44	18.35	
Rec Language					
CD	38.85	6.09	55.85	36.23	9.18 (1,21)**
PD	38.78	6.44	65.78	25.81	
Exp Language					
CD	47.92	6.17	53.38	31.91	1.30 (1,20)
PD	48.44	6.96	59.22	25.13	
Vineland					
Com					
CD	57.46	4.97	73.69	32.32	7.57 (1,21)*
PD	63.20	5.58	81.40	24.33	
DLS[a]					
CD	63.92	5.53	66.23	25.95	.11 (1,21)
PD	64.20	3.68	64.20	12.42	
Soc					
CD	58.38	6.17	73.92	23.49	10.30 (1,21)**
PD	60.30	5.76	68.90	10.11	
ABC[b]					
CD	59.54	5.31	69.00	28.04	2.81 (1,21)
PD	60.90	5.94	66.70	14.68	
ADI-R[c]					
Social					
CD	17.54	3.73	12.33	10.58	19.15 (1,18)**
PD	18.90	1.14	13.10	9.42	
Com					
CD	12.85	2.44	8.08	6.91	41.19 (1,18)**
PD	12.90	1.22	8.80	7.43	
Ritual					
CD	5.38	1.69	5.08	3.75	1.72 (1,18)
PD	6.40	1.11	5.60	3.50	

Note. CD $n = 13$; PD $n = 10$ except for Verbal IQ and Performance IQ, where n was 10 for both groups because 3 CD children had only Bayley tests. Neither the main effect of groups (CD vs. PD) nor the interaction of groups by time was significant for any variable. Full scale IQs at pretreatment are Bayley scores.

[a]Daily living skills. [b]Adaptive Behavior Composite. [c]Autism Diagnostic Interview-Revised.

*$p < .05$. **$p < .01$.

Pet. Reh. App.31

Table 3. Pretreatment and Outcome Scores of Rapid (R) and Moderate (M) Learners

Measure/	Pretreatment		Posttreatment		ANOVA Pre–Post
Group	Mean	*SD*	Mean	*SD*	comparisons
Full Scale IQ					
R	55.27	8.96	103.73	13.35	143.19 (1,21)**
M	47.83	9.37	50.42	6.98	0.45 (1,21)
Verbal IQ					
R	—	—	101.45	18.72	70.76 (1,18)**
M	—	—	47.44	2.06	.02 (1,18)
Perform IQ					
R	—	—	107.55	9.44	165.27 (1,18)**
M	—	—	63.67	8.43	11.81 (1,18)**
Nonverbal IQ					
R	83.56	14.84	108.78	10.96	16.69 (1,19)**
M	69.83	15.93	67.70	12.35	0.19 (1,19)
Rec Language					
R	39.30	6.91	93.60	12.64	217.76 (1,20)**
M	38.42	5.59	31.83	9.87	3.84 (1,20)
Exp Language					
R	49.90	7.75	85.70	15.07	77.76 (1,20)**
M	47.50	6.54	30.83	5.89	20.24 (1,20)**
Vineland					
Com					
R	60.82	4.02	105.09	12.83	147.07 (1,21)**
M	59.17	7.22	51.33	10.94	5.07 (1,21)*
DLS[a]					
R	66.45	4.25	82.27	16.34	20.50 (1,21)**
M	61.83	4.20	49.83	10.61	12.87 (1,21)**
Soc					
R	61.55	6.58	87.73	14.94	42.89 (1,21)**
M	57.08	4.63	57.08	6.40	0.00 (1,21)
ABC[b]					
R	61.73	4.59	88.64	15.68	54.17 (1,21)**
M	58.67	6.09	49.08	7.76	7.51 (1,21)*
ADI-R[c]					
Social					
R	16.45	3.26	4.18	4.37	46.89 (1,21)**
M	19.67	1.55	21.18	6.28	0.43 (1,21)
Com					
R	11.00	3.54	2.00	2.73	52.04 (1,21)**
M	13.75	0.60	14.81	3.59	1.26 (1,21)
Ritual					
R	5.91	1.62	2.73	2.67	16.46 (1,21)**
M	5.92	1.44	7.91	2.47	4.87 (1,21)*

Note. R $n = 11$; M $n = 12$. Posttreatment language scores for moderate learners are Reynell ratio scores (AE/CA), which are about 10 points lower than standard scores. Effect size expressed as proportion of variance was .88 for Full Scale IQ, .90 for receptive language, .84 for expressive language, and .73 for Vineland ABC, all quite large (Cohen, 1988). Full Scale IQs at pretreatment are Bayley scores.

[a] Daily living skills. [b] Adaptive Behavior Composite. [c] Autism Diagnostic Interview-Revised.

$^{*}p < .05$. $^{**}p < .01$.

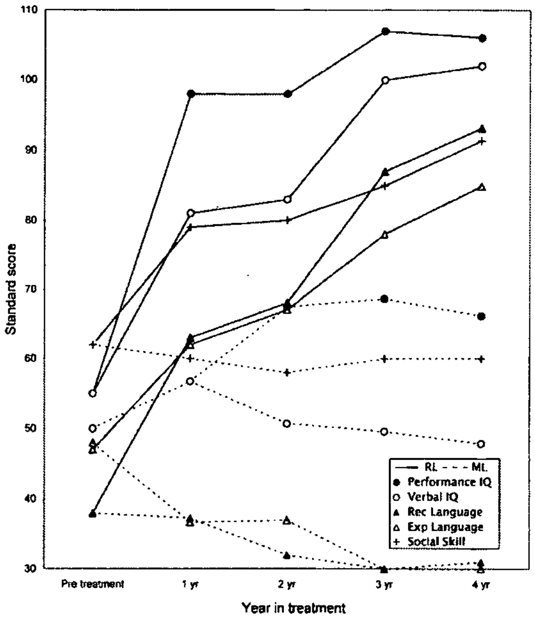

Figure 2. Mean IQ, language, and socialization scores during treatment for rapid (RL) and moderate (ML) learners. Initial IQ and language scores are ratio scores as are all language scores of moderate learners.

Most parents waited until their children were 6 years old to enter kindergarten, per our recommendation, in order to allow them more time to acquire social interaction skills. At an average age of 7.67, the 11 rapidly learning children were succeeding in regular first or second grade classes following the regular curriculum. On the Woodcock Johnson III Tests of Achievement, Oral Expression averaged 102 (*SD* = 11.9, 1 scored below 85), Listening Comprehension averaged 101 (*SD* = 15.27, 2 scored below 85), Broad Reading averaged 105 (*SD* = 11.9, all scored over 85), Broad Math averaged 104 (*SD* = 18.4, one scored below 85), Spelling averaged 112 (*SD* = 18.83, all scored over 85) and general Academic Knowledge averaged 98 (*SD* = 18.1, 2 scored below 85). Three children had aides because of inattentiveness and 3 received speech therapy, although all spoke fluently.

The 12 moderate learners showed a significant improvement in Performance IQ, $F(1, 18) = 11.81, p < .01$, as shown in Table 3, but the post-treatment mean score (63.67) was over two *SD*s below the average range. Although these children did not "catch up" to peers, they did show increases in developmental age equivalents. Cognitive skills increased from 16 to 44 months; adaptive skills, from 16 to 37 months; language skills, from less than 12 months to 27 months; and social skills, from 10 to 31 months. At the end of the study, these children were continuing to gain skills at a rate of 3.4 to 4.3 months per year in expressive language and social skills, respectively. All but 2 of them acquired speech, allowing them to communicate basic needs while also reducing frustration. Two thirds learned to read simple stories (e.g., first grade level words with two sentences per page). Most acquired the ability to relate to others and to play with peers. Four of the children were in regular classes with an aide, but all had a modified curriculum. Six children had a mixture of some time in regular class and some time in special education, and 2 were in full-time special education classes (one for students with cognitive disabilities and the other for those with emotional disturbances).

Assessment of Residual Symptoms in Rapidly Learning Children

Parents completed the Personality Inventory for Children for all 23 children. As shown in Table 4, rapidly learning children as a group scored in the average range on all factor scales, although 2 scored in the clinically significant range on Factor III (they tended to worry). Moderate learners were rated as having more tantrums (Factor I), more difficulty interacting with others (Factor II), and more learning problems (Factor IV).

Parents and teachers completed the Child Behavior Checklist for all 23 children. Results were analyzed using 2 × 2 ANOVAS (Rapid Learners vs. Moderate Learners × Parent vs. Teacher as repeated measures). As shown in Tables 4 and 5, rapid learners as a group scored in the nonclinically significant range on all scales, although they did score less normally than did moderate learners on Scale 3 (they worried more). Moderate learners were rated as less interactive (Scale 1), more preoccupied (Scale 5), less attentive (Scale 6), and more easily frustrated (Scale 8).

The differences in Child Behavior Checklist ratings between parents and teachers were small, reaching significance on two scales (1 and 8). However, these results largely reflected differences within the average range. Parents did not rate any children in the clinically significant range on either scale, and teachers rated only 2 children on

Table 4. Mean Scores of Rapid and Moderate Learners on Posttreatment Only Tests of Residual Symptoms: Parent Ratings

	PIC[a] factor				Child Behavior Checklist[b] scale					
Learner	I	II	III	IV	1	3	4	5	6	8
Rapid (R)										
(*n* = 11)	53.45	62.36	55.27	64.18	59.09	55.40	57.82	65.64	62.64	52.91
(*SD*)	(9.38)	(8.34)	(13.90)	(13.65)	(6.26)	(6.14)	(7.49)	(9.87)	(9.12)	(4.98)
Moderate (M)										
(*n* = 12)	66.83	79.25	49.73	97.55	58.83	51.75	61.92	70.42	67.67	53.33
(*SD*)	(12.93)	(9.42)	(8.77)	(18.77)	(6.27)	(3.06)	(7.35)	(7.92)	(8.17)	(4.62)
R vs. M[c]	3.43**	4.86**	1.06	5.13**	0.01	1.80*	1.61	1.64	1.73*	0.08

[a] Personality Inventory for Children and Child Behavior Checklist scores ≥70 are clinically significant and scores ≥67 are borderline. Scores below those levels are not reliably different from "normal" (Achenbach, 1991b; Lacher, 1982). Factor I = Undisciplined/Poor Self Control, II = Social Incompetence, III = Internalizing/Somatic Symptoms, IV = Cognitive Development. [b] Scale I = Withdrawn, 3 = Anxious/Depressed, 4 = Social Problems, 5 = Thought Problems, 6 = Attention Problems, 8 = Aggression. [c] *t* tests are one-tailed, with a *df* of 19.
**p* < .05. ***p* < .01.

Scale 1 (both moderate learners) and 3 on Scale 8 in the significant range (1 rapid and 2 moderate learners).

Whereas checklists such as the Personality Inventory for Children and the Child Behavior Checklist can be used to assess the presence of problems, the Classroom Edition of the Vineland is used to assess the presence of skills (e.g., "initiates conversation," "responds to hints or indirect cues in conversation"). Teachers completed this measure for all 23 children except the 2 who were among the highest functioning. As shown in Table 5, teacher ratings of Communication and Socialization for the remaining 9 rapid learners were in the average range. Moderate learners were rated as having deficiencies in both areas.

We examined changes in behavior related to diagnosis by comparing the Autism Diagnostic Interview-Revised administered prior to and after 3 years of treatment using 2 × 2 ANOVAS (Rapid Learners vs. Moderate Learners × Pretreatment vs. Posttreatment as repeated measures). As shown in Table 3, rapid learners as a group showed significant improvements on all three Autism Diagnostic Interview scales: Communication, $F(1, 21) = 52.04$, $p < .01$, Reciprocal Interaction, $F(1, 21) = 46.89$, $p < .01$, and stereotyped behaviors, $F(1, 21) = 16.46$, $p < .01$. Eight of 11 rapid learners scored in the nonautistic range in all three areas, and many had their diagnoses removed by the referring child psychiatrists. Of the rapid learners who had remaining problems, 1 still had some lan-

Table 5. Mean Scores of Rapid and Moderate Learners on Posttreatment Only Tests of Residual Symptoms: Teacher Ratings

	Vineland		Child Behavior Checklist scales[a]					
Learners	Comm.	Social	1	3	4	5	6	8
Rapid (R)	94.44	89.89	57.00	55.90	56.73	65.55	59.36	57.60
n = 11 (*SD*)	(13.97)	(18.36)	(7.34)	(6.93)	(6.30)	(11.37)	(12.33)	(6.11)
Moderate (M)	58.58	61.58	64.33	55.17	58.00	72.58	63.25	61.25
n = 12 (*SD*)	(7.90)	(6.02)	(6.03)	(6.56)	(5.57)	(7.06)	(7.94)	(7.45)
R vs. M[b]	6.84**	4.60**	2.93**	0.36	0.37	2.41*	1.33	2.86**

[a] Child Behavior Checklist scores ≥67 are borderline. Scores below these levels are not reliably different from "normal" (Achenbach, 1991b; Lacher, 1982). *t* tests are one-tailed. Scale 1 = Withdrawn, 3 = Anxious/Depressed, 4 = Social Problems, 5 = Thought Problems, 6 = Attention Problems, 8 = Aggression. [b] *t* tests are one-tailed, with a *df* of 19.
**p* < .05. ***p* < .01.

Pet. Reh. App.34

Table 6. Combined Parents' and Teachers' Ratings of Residual Symptoms of Rapid Learners

	Social Skills					
	VABS[b]	Isolates	Not liked	Anxious	Inattntn	
Child[a]	Com, Soc	PIC[c] 1&2	CBC 1,4	CBC 3, PIC 3	CBC 5,6	Moody CBC 8
CD						
1	104	50	50	47.7	50	50
2	115.5	50	50	48.3	50	50
3	115	51	50	51.3	55	50
4	101.3	57.5	68.3	52	79.5	65.5
5	95.5	51	56.3	60	62.5	53
PD						
1	107.5	59	55.3	68.3	54	54.5
2	79.5	54.5	57.3	46.3	67.5	54.5
3	77.5	67.5	60	51.3	64.8	61.5
4	77.5	69	63.8	63.7	70.8	58
5	86.5	67	61.3	67.0	67.8	51
6	99.5	64	62.3	51.3	65	55.5

[a]CD = clinic directed, PD = parent directed. [b]Vineland Adaptive behavior Scales (VABS) scores below 85 are moderately low and 116–130, moderately high. [c]Personality Inventory for Children (PIC) and Child Behavior Checklist (CBC) scores ≥70 are clinically significant; and ≥67, borderline; below these levels, are not reliably different from "normal" (Achenbach, 1991b; Lacher, 1982).

guage delays, 1 was rigid in play, and 1 was elevated in all three areas. The latter child had received treatment from a non-UCLA affiliated provider after the first year.

Combined measures of residual symptoms are shown in Table 6. Eight of 11 rapid learners showed increases in social skills to the adequate range (above 85), although 3 had some borderline problems, including 1 who had significant problems with Preoccupation/Inattention. The remaining 3 rapid learners showed moderately low social skills (below 85), and 2 had problems with Preoccupation/Inattention, one of which was clinically significant. All 3 of these latter children were in the parent-directed group and took longer than 2 years to achieve IQ in the average range. These results are similar to those described in UCLA reports, where 3 of 8 best outcome children scored below 85 on Vineland Communication, 3 were elevated on the Vineland Maladaptive Behavior scale, and 5 had at least one significant elevation on the Personality Inventory for Children. In interpreting these results, McEachin et al. (1993) noted that 3 of their nonclinical children also had significant Personality Inventory elevations.

Predicting Outcome

Early learning measure. Performance of rapid and moderate learners on each of the four subscales of the Early Learning Measure is shown in Figure 3. As can be seen, the difference in their rates of learning was evident early in treatment. Thirteen of 23 children passed the Early Learning Measure (90% correct on verbal imitation). All 11 who later achieved scores in the average range passed by 16 weeks of treatment (9 children) or before reaching 42 months of age (2 children).

Pretreatment variables. Table 7 shows correlations between pretreatment variables and three outcome variables following 3 years of treatment: (a) Full Scale IQ; (b) Language, defined as the mean of three measures—Vineland Communication scores from parents and teachers representing language usage at home and school and language scores from the Reynell or Clinical Evaluation of Language Fundamentals; (c) Social Skills, defined as the mean of three measures—Vineland Socialization scores from parents and teachers and Factor II (Social Incompetence) from the Personality Inventory for Children.

The ability to imitate on the Early Learning Measure was highly correlated with outcome in all three areas. Seven children were able to imitate 3 of 20 sounds prior to treatment (mean total sounds imitated during the first three Early Learning Measures was 2.43, range = 0 to 15, *SD* = 4.04), and all went on to achieve IQs in the average range.

Pet. Reh. App.35

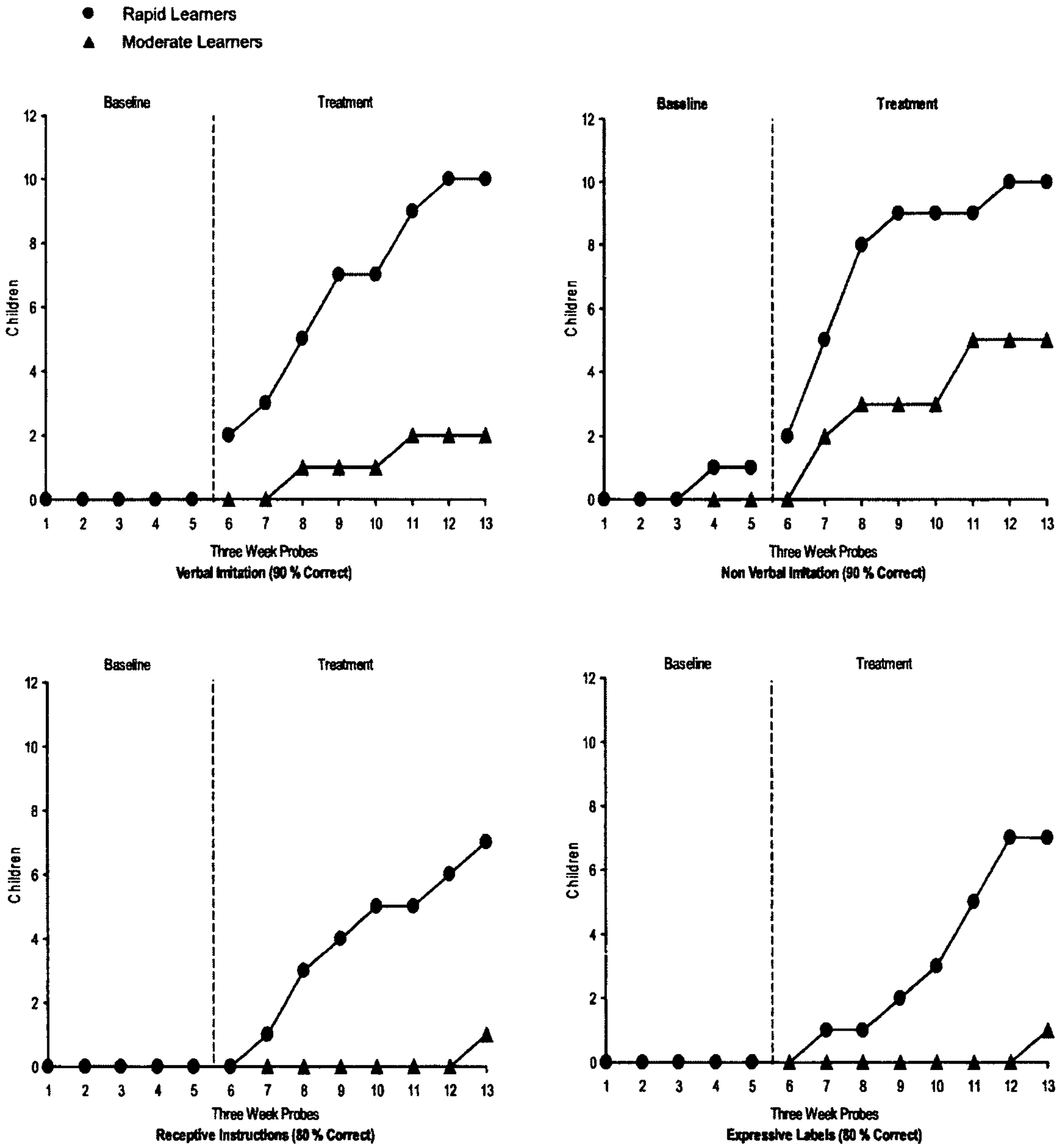

Figure 3. Performance of rapid (RL) and moderate (ML) learners on the Early Learning Measure.

We used linear regression using the best subset approach (Hosmer et al., 1989) to select the most powerful predictors for each outcome area. Based on previous research, potential predictor variables included IQ, imitation, language, social relatedness, and severity of symptoms. Posttreatment IQ was best predicted by the subset of variables including pretreatment Early Learning Measure (receptive language, nonverbal imitation, and verbal imitation), pretreatment IQ, Autism Diagnostic Interview Impairment in Social Interaction (low social interest, unresponsive to others' approaches, lack of shared attention), and Autism Diagnostic Interview Communication scores. This set of variables yielded a correlation of .83 with posttreatment IQ, which is a strong relationship.

Pet. Reh. App.36

Table 7. Correlations Between Pretreatment and Posttreatment Measures

	Follow-up				
	One year		Three year		
Pretreatment measure[a]	IQ	IQ change	IQ	Language	Social
Reynell					
Expressive	.46*	.37	.35	.41	.45*
Comprehension	.30	.19	.24	.27	.31
ELM					
Nonverbal Imitation	.59**	.41	.71**	.69**	.81**
Exp. Labeling	.48*	.54*	.46*	.56**	.65**
Rec. Instructions	.47*	.27	.56**	.56**	.67**
Verbal Imitation	.62**	.59**	.65**	.69**	.80**
VABS					
Communication	.49*	.35	.33	.44*	.41
DLS[b]	.57*	.40	.57**	.60**	.63**
Motor	.36	.16	.17	.22	.27
Socialization	.44*	.31	.41*	.43*	.47*
Composite	.56*	.32	.37	.43*	.46*
Merrill-Palmer IQ	.20	−.01	.08	.06	−.07
Bayley Ratio IQ	.51*	−.01	.45*	.34	.28
ADI-R					
Communication	−.49*	−.35	−.59**	−.52*	−.57**
Socialization	−.22	−.18	−.63**	−.50*	−.52*
Ritualistic	−.12	−.17	−.12	−.10	−.10
First year IQ change	.86**	—	.87**	.92**	.82**
IQ at one year	—	.86**	.75**	.84**	.75**

[a] Reynell = Reynell Developmental Language Scales, ELM = Early Learning Measure, VABS = Vineland Adaptive Behavior Scales, ADI-R = Autism Diagnostic Interview-Revised. [b] Daily Living Skills.
*$p < .05$. **$p < .01$.

The amount of variation in posttreatment IQ explained by this subset of pretreatment variables was 70%.

Social skill acquisition was also predicted by the pretreatment ability to imitate. The subset of variables, including pretreatment Early Learning Measure scores (receptive language, nonverbal imitation, and verbal imitation) and Autism Diagnostic Interview Communication yielded a correlation of .90 with posttreatment social skill scores, a strong relationship. The amount of variance in posttreatment social skill scores explained by this subset of pretreatment variables was 82%.

Finally, language skill acquisition was also predicted by the pretreatment ability to imitate. The subset of variables including pretreatment Early Learning Measure scores (receptive language, nonverbal imitation, and verbal imitation), Vineland Daily Living Skills, and Autism Diagnostic Interview Communication yielded a correlation of .87 with posttreatment language scores, a strong relationship. The amount of variance in posttreatment language scores explained by this subset of pretreatment variables was 75%.

Parents of 6 children (26%) reported acquisition of 5 to 25 words, all of which were later lost between 15 and 26 months of age. Language regression in other studies has varied between 20% and 50% (Howlin, 1998), with a mean near 30% (Shinnar et al., 2001) and median age of 18 months (Tuchman & Rapin, 1997). Shinnar et al. reported that among those children who regained some language, only 8% achieved typical language. In the present study, loss of speech was not

Pet. Reh. App.37

related to outcome. Three rapid learners and 3 moderate learners had a clear loss, and 6 rapid learners and 2 moderate learners had no loss (Rapid Learners vs. Moderate Learners × Pre- vs. Posttreatment, χ^2 (1, N =14) = .16, ns. Three of 6 children with clear regression (50%) achieved typical language. However, having no speech at the start of treatment (age 36 months), whether from earlier loss (and not having recovered any) or never having developed speech, was associated with slower learning.

We used logistic regression to develop models to predict the probability of achieving 3-year outcome scores in the average range based on pretreatment measures. The most accurate model for the current set of data combined pretreatment Verbal Imitation from the Early Learning Measure and pretreatment Autism Diagnostic Interview Communication as follows: $p/(1\text{-}p) = e^y$, where e = (approximately) 2.718284 and y = [1.76 (total verbal imitation items correct out of 20 trials from standard set administered three times, 3 weeks apart) −2.64 (Autism Diagnostic Interview-Communication score) + 32.57]. Using a score above 0.5 to classify children as potentially "best outcome," this model correctly predicted 10 of 11 such children (sensitivity = 10/11 = .91), with one false positive and one false negative (specificity = 21/23 = .91). Predictive power was .91.

Hours of treatment. Table 8 shows the distribution of direct intervention hours for rapid learners during treatment. Most children received predominantly 1:1 intervention during the first year, and then gradually spent more time in school. Once children were able to use language, treatment was focused increasingly on building the social skills necessary to function in school and to interact with peers.

The number of weekly hours of treatment seemed less related to outcome than did pretreatment variables. Rapid learners averaged 34 hours per week during the first year (range = 25 to 40) and 31 during the second year (range = 20 to 39). Those who learned at a more moderate rate had identical averages, although they had less peer play due to limited play and language skills.

The hours shown in Table 8 do not include time spent by parents generalizing gains made in therapy, which they found quite difficult to estimate. In an effort to assess the impact of parental involvement, senior therapists rated parents on the percentage of involvement in their child's treatment during the first year. Although the correlation with outcome, r = .32, was not significant, the real impact of parental involvement may not be seen until formal treatment has ceased, when parents who were more involved all along and, therefore, acquired more skills, may be better prepared to help their child deal with new challenges.

Table 8. Average Allocation of Treatment Hours Over Time for Rapid Learners

	Years of treatment								
Staffing	.5	1	1.5	2	2.5	3	3.5	4	4.5
n	11	11	10	8	7	7	7	6	6
1:1	33	29	24	22	20	18	15	12	10
	(15–40)	(16–35)	(10–33)	(15–31)	(10–27)	(5–28)	(0–25)	(4–25)	(0–15)
School	5	6	8	8	12	13	18	28	33
	(0–12)	(0–12)	(0–25)	(0–16)	(8–20)	(8–25)	(8–30)	(15–35)	(25–35)
School shadow	1	1	4	5	8	11	7	5	5
	(0–5)	(0–5)	(0–15)	(0–15)	(3–15)	(6–18)	(0–18)	(0–12)	(2–15)
Peer shadow	0	3	3	6	5	4	4	3	2
	(0)	(0–5)	(0–5)	(2–9)	(0–9)	(0–8)	(2–8)	(0–6)	(0–4)
Total	34	33	31	33	33	33	26	21	17
	(25–40)	(26–40)	(20–37)	(20–39)	(25–37)	(20–40)	(7–40)	(6–31)	(12–20)

Note. Ranges are in parentheses. Total hours include school hours only when a shadow was present. Hours are for children still in treatment at each point in time. One child transferred to another provider after 1 year. Children began "graduating" from treatment after 2 years. Children who had difficulty learning complex material maintained full hours longer, but treatment focused more on 1:1 hours to teach skills and less on peer interaction due to lower social interest and language delays.

Among rapid learners, the number of hours of structured home-based peer play was significantly related to teachers' ratings of social skills at 4 years. Although most children began peer play by 48 months of age, those who were subsequently rated by teachers as being within the average range (Vineland Socialization score of at least 90, and no Child Behavior Checklist scores over 65 on Scale 1 (Withdrawn) or Scale 4 (Social Problems), had several things in common. By age 54 months, they were all receiving at least 6 (mean = 8) hours of supervised peer play per week with at least two unfamiliar peers (i.e., not siblings or cousins), and this continued for at least 6 months (M = 13), p = .008 (Fisher Exact Test).

Supplemental treatments. Of the 23 children participating, 22 received some type of supplemental treatment prior to or during the first year of treatment (19 of 23 children). These services consisted of special education (21), preschool (2), and private therapies beyond what was offered in school: speech (5), sensory integration (7), auditory integration training (2), music therapy (1), and horseback riding (1). Hours per week of supplemental treatment ranged from 0 to 14 (average = 6) prior to and 0 to 15 (average = 7) hours during the first year of treatment. Between the first and third year of treatment, biomedical management became more popular, and more parents tried them. Nine children were on Gluten-Casein free diets (for 1 month to 21 months), 10 received mega-vitamins and/or dimethylglycine–DMG (for 1 month to 3 years), 4 received Secretin (1 to 4 doses), 4 were given Nystatin (for 1 month to 12 months), and 1 received 20 doses of Intravenous Immune Globulin. However, the correlation between hours of supplemental treatment and outcome (−.335 with IQ, −.384 with language, and −.334 with socialization) and that between the use of biomedical treatments and outcome (−.050 with IQ, −.108 with language, and −.141 with socialization) were low and not significant, supporting the conclusion that the increases in skills observed in this study were not the result of these interventions.

Discussion

In the present study we demonstrated that the UCLA early intensive behavioral treatment program could be implemented in a clinical setting outside a university with a similar sample and that the earlier findings by the UCLA group regarding favorable outcome (Lovaas, 1987; McEachin et al., 1993) could in large part be replicated without aversives. Following 2 to 4 years of treatment, 11 of 23 children (48%) achieved Full Scale IQs in the average range, with IQ increases from 55 to 104, as well as increases in language and adaptive areas comparable to data from the UCLA project. At age 7, these rapid learners were succeeding in regular first or second grade classes, demonstrated generally average academic abilities, spoke fluently, and had peers with whom they played regularly.

Parent-directed children, who received 6 hours per month of supervision (usually 3 hours every other week, which is much more than "parent-managed" or "workshop" supervision), did about as well as clinic-directed children, although they received much less supervision. This was unexpected, and it may have been due in part to parent-directed parents taking on the senior therapist role, filling cancelled shifts themselves, actively targeting generalization, and pursuing teachers and neighbors to find peers for daily play dates with their children. Although many parent-directed parents initially made decisions regarding treatment that resulted in their children progressing slowly (e.g., using their treatment hours for ineffective interventions or pushing children to learn advanced skills before they were ready), resulting in frustration and occasionally "shutting down," many parents then sought input from treatment supervisors and rapidly learned to avoid making the same mistake twice, becoming quite skillful after a few months.

Several measures were used to assess residual symptoms of autism among rapid learners, and while generally not clinically significant, some were found, particularly among those children who achieved average IQ after several years of treatment. About one third of the rapid learners were seen as having mild delays in social skills. Seeming preoccupied was also a common problem for which 3 children were assigned classroom aides because they "needed reminders to stay on task." Lovaas (1987) did not mention that aides were assigned to any of his "best outcome" children, and it is possible that our children were not as "normal." However, McEachin et al. (1993) found that in spite of scoring in the clinically significant range in one or two areas, children were able to maintain their skills, scoring in the average range on standardized tests of cognitive, emotional, and social variables and to succeed in regular classes at follow-up 6 years after treatment was stopped.

The strongest pretreatment predictors of outcome were imitation, language, daily living skills, and socialization. Rapid acquisition of new material as measured by the Early Learning Measure, first year IQ, and change in IQ after 1 year were also strong predictors. These findings are consistent with previous research. A model with 91% accuracy was derived for predicting whether a child in the present sample would be a rapid or moderate learner. The usefulness of the model must await validation with other similar samples. We note that one of the two predictors in the model was pretreatment verbal imitation, which is not widespread among untreated 3-year-old children with autism. However, the model may not discriminate among children above some as yet undetermined age because they often acquire imitation by school age (Charman et al., 1997).

Because we used the Bayley to determine pretreatment IQ and Wechsler tests at follow-up, there was a possibility that the observed increases in IQ may have reflected the use of different tests instead of treatment effects. To examine this, we compared changes in scores over time from Bayley at Time 1 to Bayley at Time 2, with changes from Bayley at Time 1 to Wechsler test at Time 2. One rapid learner was tested using the Bayley at pretreatment and again after 1 year of treatment because he was still only 3 years old. His score increased from 44 to 97, similar to increases seen in rapid learners tested with the Bayley at pretreatment and the WPPSI-R at 1 year. Ten moderate learners were tested using the Bayley at pretreatment and again after 1 year of treatment, and with Wechsler tests thereafter. For these children, Bayley to Bayley IQs increased from 47.2 to 54.3. Bayley to Wechsler IQs increased from 53.7 to 54.6. Therefore, there did not seem to be an effect on IQs attributable to using different tests.

Another possible confound was that most pre- and posttesting of moderate learners was done by the second author, perhaps introducing bias. However, the correlation between scores obtained by the second author and unaffiliated community psychologists was high, and the finding of little improvement over time on standardized tests for children in this subgroup is consistent with previous findings. A related question is whether the positive findings among rapid learners were due to treatment or maturation. Arguing against the maturation hypothesis is the negligible improvement of children receiving community services found in several longitudinal studies (Eikeseth et al., 2002; Lord & Schopler, 1989; Lovaas, 1987; Sheinkopf & Siegel, 1998).

Although we matched on age and IQ and employed random assignment, this was not sufficient to ensure equal samples. Other pretreatment variables, such as imitation, correlated even more strongly with outcome and were not equal in the two groups. As a result, we were unable to interpret treatment effects among subgroups of rapid learners. Further, the small number of children in the study limited the power of statistical tests to detect differences, and the many tests on such a small sample increased the likelihood of spurious findings, thereby limiting the implications of results for the larger population of children with autism. However, because some treatment effects were so large and have been found in other studies (e.g., that a subset of the children do well), the current results can be seen as supporting an existing body of research.

We found two interesting correlations that deserve further study. First, ratings of parental involvement were weakly related to outcome, suggesting that more overt efforts to increase parents feeling capable of contributing to treatment planning may enhance treatment effects (Ramey et al., 1992). Second, acquisition of social skills was positively related to amount and duration of supervised peer play. Some parents were uncomfortable approaching other parents to set up play dates, and problems doing so may provide a partial explanation for the lower social skills scores of their children. Even so, amount and duration of supervised peer play are surely just a few of the variables that affect acquisition of social skills. Although we do have several powerful interventions, including incidental teaching, role playing, and video modeling, to teach a curriculum of social conversation, cooperative play, and understanding the nonverbal communication of others, building typical social skills remains a work in progress (McConnell, 2002).

Hours of treatment in this study came closer than any previous replication to the intensity of hours provided in the UCLA study (Lovaas, 1987), averaging 38 hours per week for 2 years in the clinic-directed group, and the results were also the most comparable. Forty-eight percent of the children showed dramatic increases in cognitive and social skills and were able to succeed in regular education classes. However, high hours and intensive supervision were not sufficient to make up for low levels of pretreatment skills. Consistent with previous studies, low IQ (below 44) and absence of language (no words at 36 months) predicted limited progress, whereas rate of learning, imitation. and social relatedness predicted favorable outcomes (Lord, 1995). Although starting at a disadvantage, children learning at a moderate rate were still acquiring new skills after 4 years. We intend to follow all of the children for several more years to determine their outcome in adolescence and adulthood.

References

Achenbach, T. M. (1991a). *Child Behavior Checklist.* Burlington: University of Vermont Department of Psychiatry. (Available from ASEBA, 1 S. Prospect St., Burlington, VT O5401-3456 and online at http://checklist.uvm.edu)

Achenbach, T. M. (1991b). *Manual for the Teacher's Report Form and 1991 Profile.* Burlington: University of Vermont Department of Psychiatry. (Available from ASEBA, 1 S. Prospect St., Burlington, VT O5401-3456, and online at http://checklist.uvm.edu)

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: American Psychiatric Association.

Anderson, S. R., Avery, D. L., DiPietro, E. K., Edwards, G. L., & Christian, W. P. (1987). Intensive home-based intervention with autistic children. *Education and Treatment of Children, 10,* 352–366.

Bayley, N. (1993). *Bayley Scales of Infant Development* (2nd ed.). San Antonio: Psychological Corp.

Bibby, P., Eikeseth, S., Martin, N. T., Mudford, O. C., & Reeves, D. (2002). Progress and outcomes for children with autism receiving parent-managed intensive interventions. *Research in Developmental Disabilities, 23,* 81–104.

Bierman, K. L., & Welsh, J. A. (1997). Social relationship deficits. In E. J. Mash & L. G. Terdal (Eds.), *Assessment of childhood disorders* (3rd ed., pp. 328–365). New York: Guilford Press.

Birnbrauer, J. S., & Leach, D. J. (1993). The Murdoch Early Intervention Program after 2 years. *Behavior Change, 10,* 63–74.

Bondy, A., & Frost, L. (1994). The Picture-Exchange Communication System. *Focus on Autistic Behavior, 9,* 1–19.

Bono, M. A., Daley, T., & Sigman, M. (2004). Relations among joint attention, amount of intervention and language gain in autism. *Journal of Autism and Developmental Disorders, 34,* 495–505.

Carr, E. G., & Durand, V. M. (1985). Reducing behavior problems through functional communication training. *Journal of Applied Behavior Analysis, 18,* 111–126.

Charlop, M. H., & Milstein, J. P. (1989). Teaching autistic children conversational speech using video modeling. *Journal of Applied Behavior Analysis, 22,* 245–285.

Charman, T., Swettenham, J., Baron-Cohen, S., Cox, A., Baird, G., & Drew, A. (1997). Infants with autism: An investigation of empathy, pretend play, joint attention, and imitation. *Developmental Psychology, 33,* 781–789.

Dawson, G., & Osterling, J. (1997). Early intervention in autism. In M. Guralnick (Ed.), *The effectiveness of early intervention.* Baltimore: Brookes.

Eikeseth, S., Smith, T., Jahr, E., & Eldevik, S. (2002). Intensive behavioral treatment at school for 4- to 7-year-old children with autism: A one-year comparison controlled study. *Behavior Modification, 26,* 49–68.

Eldevik, S., Eikeseth, S., Jahr, E., & Smith, T. (in press). Effects of low-intensity behavioral treatment for children with autism and mental retardation. *Journal of Autism and Developmental Disorders.*

Fenske, B. C., Zalenski, S., Krantz, P. J., & McClannahan, L. E. (1985). Age at intervention and treatment outcome for autistic children in a comprehensive intervention program. *Analysis and Intervention in Developmental Disabilities, 5,* 49–58.

Gray, C. (1994). *The social story book.* Arlington, TX: Future Horizons.

Green, G. (1996). Early behavioral intervention for autism: What does research tell us? In C. Maurice, G. Green, & S. C. Luce (Eds.), *Behavioral intervention for young children with autism* (pp. 29–44). Austin, TX: Pro-Ed.

Gresham, F. M., & MacMillan, D. L. (1998). Early intervention project: Can its claims be substantiated and its effects replicated? *Journal of Autism and Developmental Disorders, 28,* 5–13.

Harris, S. L., & Handleman, J. S. (2000). Age and IQ at intake as predictors of placement for young children with autism: A four- to six-year follow up. *Journal of Autism and Developmental Disorders, 30,* 137–142.

Harris, S. L., Handleman, J. S., Gordon, R., Kristoff, B., & Fuentes, F. (1991). Changes in cognitive and language functioning of preschool children with autism. *Journal of Autism and Developmental Disorders, 21,* 281–290.

Hart, B., & Risley, T. R. (1975). Incidental teaching of language in the preschool. *Journal of Applied Behavior Analysis, 8,* 411–420.

Hosmer, D. W., Jovanovic, B., & Lemeshow, S. (1989). Best subset logistic regression. *Biometrics, 45,* 1265–1270.

Howard, J. S., Sparkman, C. R., Cohen, H. G., Green, G., & Stanislaw, H. (2005). A comparison of intensive behavior analytic and eclectic treatments for young children with autism. *Research in Developmental Disabilities, 26,* 359–383.

Howlin, P. (1998). *Children with autism and Asperger syndrome: A guide for practitioners and carers.* Chichester, West Sussex, England: Wiley.

Jacobson, J. W., Mulick, J. A., & Green, G. (1998). Cost-benefit estimates for early intensive behavioral intervention for young children with autism: General models and single state case. *Behavioral Interventions, 13,* 201–226.

Jahr, E., Eldevik, S., & Eikeseth, S. (2000). Teaching autistic children to initiate and sustain cooperative play. *Research in Developmental Disabilities, 21,* 151–169.

Koegel, L. K., Koegel, R. L., Shoshan, Y., & McNerney, E. (1999). Pivotal response intervention II: Preliminary long-term outcomes data. *Journal of the Association for Persons with Severe Handicaps, 24,* 186–198.

Koegel, R. L., & Koegel, L. K. (1995). *Teaching children with autism: Strategies for initiating positive interactions and improving learning opportunities.* Baltimore: Brookes.

Koegel, R. L., Russo, D. C., & Rincover, A. (1977). Assessing and training teachers in the generalized use of behavioral modification with autistic children. *Journal of Applied Behavior Analysis, 10,* 197–205.

Lachar, D. (1982). *Personality Inventory for Children (PIC): Revised format manual supplement.* Los Angeles: Western Psychological Services.

Lord, C. (1995). Follow-up of two-year-olds referred for possible autism. *Journal of Child Psychology and Psychiatry, 36,* 1365–1382.

Lord, C., & Paul, R. (1997). Language and communication in autism. In D. L. Cohen & F. R. Volkmar (Eds.), *Handbook of autism and pervasive developmental disorders* (2nd ed., pp. 195–225). New York: Wiley.

Lord, C., Rutter, M., & LeCouteur, A. (1994). Autism Diagnostic Interview–Revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of Autism and Developmental Disorders, 23,* 659–685.

Lord, C., & Schopler, E. (1989). The role of age at assessment, developmental level, and test in the stability of intelligence scores in young autistic children. *Journal of Autism and Developmental Disorders, 19,* 483–499.

Lovaas, O. I. (1987). Behavioral treatment and normal educational and intellectual functioning in young autistic children. *Journal of Consulting and Clinical Psychology, 55,* 3–9.

Lovaas, O. I., Ackerman, A. B., Alexander, D., Firestone, P., Perkins, J., & Young, D. (1981). *Teaching developmentally disabled children: The me book.* Austin, TX: Pro-Ed.

Lovaas, O. I., Koegel, R. L., Simmons, J. Q., & Long, J. S. (1973). Some generalization and follow-up measures on autistic children in behavior therapy. *Journal of Applied Behavior Analysis, 6,* 131–166.

Lovaas, O. I., & Smith, T. (1988). Intensive behavioral treatment for young children with autism. In B. B. Lahey & A. E. Kazdin (Eds.), *Advances in clinical child psychology* (Vol. 11, pp. 285–324). New York: Plenum.

Lovaas, O. I., Smith, T., & McEachin, J. J. (1989). Clarifying comments on the young autism study: Reply to Schopler, Short and Mesibov. *Journal of Consulting and Clinical Psychology, 57,* 165–167.

Maine Administrators of Service for Children with Disabilities. (2000). *Report of the MADSEC autism task force.* Manchester, ME: Author. (Available online at http://www.madex.org)

Maurice, C., Green, G., & Luce, S. C. (Eds.). (1996). *Behavioral intervention for young children with autism.* Austin, TX: Pro–Ed.

McConnell, S. R. (2002). Interventions to facilitate social interaction for young children with autism: Review of available research and recommendations for educational intervention and future research. *Journal of Autism and Developmental Disorders, 32,* 351–372.

McEachin, J. J., Smith, T., & Lovaas, O. I. (1993).

Long-term outcome for children with autism who received early intensive behavioral treatment. *American Journal on Mental Retardation, 97,* 359–372.

Meyer, L. S., Taylor, B. A., Levin, L., & Fisher, J. R. (2001). Alpine Learning Group. In J. S. Handleman & S. L. Harris (Eds.), *Preschool education programs for children with autism* (2nd ed., pp. 135–155). Austin, TX: Pro-Ed.

Mundy, P. (1993). Normal versus high-functioning status in children with autism. *American Journal on Mental Retardation, 97,* 381–384.

Newsom, C., & Rincover, A. (1989). Autism. In E. J. Mash & R. A. Barkley (Eds.), *Treatment of childhood disorders* (pp. 286–346). New York: Guilford.

New York State Department of Health, Early Intervention Program. (1999, May). *Clinical practice guidelines: Autism/pervasive developmental disorders, assessment and intervention for young children (ages 0–3 years).* Albany: Author.

Ramey, C. T., Bryant, D. M., Wasik, B. H., Sparling, J. J., Fendt, K. H., & LaVange, L. M. (1992). Infant health and development program for low birth weight, premature infants: Program elements, family participation, and child intelligence. *Pediatrics, 3,* 454–465.

Reynell, J. K., & Gruber, G. P. (1990). *Reynell Developmental Language Scales.* Los Angeles: Western Psychological Services.

Roid, G. H., & Miller, L. J. (1995, 1997). *Leiter International Performance Scale–Revised.* Wood Dale, IL: Stoelting.

Romanczyk, R. G., Lockshin, S. B., & Matey, L. (2001). In J. S. Handleman & S. L. Harris (Eds.), *Preschool education programs for children with autism* (2nd ed., pp. 49–94). Austin, TX: Pro-Ed.

Schopler, E., Short, A., & Mesibov, G. (1989). Relation of behavioral treatment to normal functioning: Comment on Lovaas. *Journal of Consulting and Clinical Psychology, 57,* 162–164.

Schreibman, L. (1997). Theoretical perspectives on behavioral intervention for individuals with autism. In D. L. Cohen & F. R. Volkmar (Eds.), *Handbook of autism and pervasive developmental disorders* (2nd ed., pp. 920–933). New York: Wiley.

Schreibman, L. (1988). *Autism.* Newbury Park, CA: Sage.

Semel, E., Wiig, E. H., & Secord, W. A. (1995). *Clinical evaluation of language fundamentals* (3rd ed.). San Antonio: Psychological Corp.

Sheinkopf, S. J., & Siegel, B. (1998). Home-based behavioral treatment of young children with autism. *Journal of Autism and Developmental Disorders, 28,* 15–23.

Shinnar, S., Rapin, I., Arnold, S., Tuchman, R. F., Shulman, L., Ballaban-Gil, K., Maw, M., Deuel, R. K., & Volkmar, F. R. (2001). Language regression in childhood. *Pediatric Neurology, 24,* 183–189.

Smith, T. (1993). Autism. In T. R. Giles (Ed.), *Handbook of effective psychotherapy* (pp. 107–133). New York: Plenum.

Smith, T., Buch, G. A., & Gamby, T. E. (2000). Parent-directed, intensive early intervention for children with pervasive developmental disorder. *Research in Developmental Disabilities, 21,* 297–309.

Smith, T., Eikeseth, S., Klevstrand, M., & Lovaas, O. I. (1997). Intensive behavioral treatment for preschoolers with severe mental retardation and pervasive developmental disorder. *American Journal on Mental Retardation, 102,* 238–249.

Smith, T., Groen, A., & Wynn, J. (2000). Randomized trial of intensive early intervention for children with pervasive developmental disorder. *American Journal on Mental Retardation, 105,* 269–285.

Smith, T., & Lovaas, O. I. (1997). The UCLA Young Autism Project: A reply to Gresham and McMillan. *Behavioral Disorders, 22,* 202–218.

Smith, T., McEachin, J. J., & Lovaas, O. I. (1993). Comments on replication and evaluation of outcome. *American Journal on Mental Retardation, 97,* 385–391.

Sparrow, S. S., Balla, D. A., & Cicchetti, D. V. (1984). *Vineland Adaptive Behavior Scales* (Interview Ed.). Circle Pines, MN: American Guidance Service.

Stutsman, R. (1948). *Merrill Palmer Scale of Mental Tests.* Wood Dale, IL: Stoelting.

Tuchman, R. F., & Rapin, I. (1997). Regression in pervasive developmental disorders: Seizures and epileptiform electroencephalogram correlates. *Pediatrics, 99,* 560–566.

Venter, A., Lord, C., & Schopler, E. (1992). A follow-up study of high-functioning autistic children. *Journal of Child Psychology and Psychiatry, 33,* 489–507.

Wechsler, D. (1989). *Wechsler Preschool and Primary Scale of Intelligence–Revised.* San Antonio, TX: Psychological Corp.

Pet. Reh. App.43

Wechsler, D. (1991). *Manual for the Wechsler Intelligence Scale for Children: Third Edition.* San Antonio: Psychological Corp.

Weiss, M. J. (1999). Differential rates of skill acquisition and outcomes of early intensive behavioral intervention for autism. *Behavioral Interventions, 14,* 3–22.

Wirt, R. D., Lachar, D., Klinedinst, J. K., & Seat, P. D. (1977). *Multidimensional descriptions of child personality: A manual for the Personality Inventory for Children.* Los Angeles: Western Psychological Services.

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

Received 7/22/04, accepted 7/20/05.
Editor-in-charge: William E. MacLean, Jr.

This research was supported in part by National Institute of Mental Health Grant MH48863-01A3, Multi-Site Young Autism Project. We thank Crystal (Burns) Held, the UCLA-trained clinic supervisor, for her efforts in carrying out this study; Tristram Smith for reviewing the manuscript multiple times; Robyn Dawes and Shuangge Ma for consultation on data analysis; and O. Ivar Lovaas for the opportunity to train at UCLA, his mentorship, and support. Requests for reprints should be sent to either author at the Wisconsin Early Autism Project, 6402 Odana Rd., Madison, WI 53719. E-mail: weap@wiautism.com

Errata

Several errors occurred in the article "Support Needs and Adaptive Behaviors," by Julia Harries, Roma Guscia, Neil Kirby, Ted Nettelbeck, and John Taplin (Vol. 110, No. 5, 393–404). On page 395, in last line under *Participants,* the *SD* should be 3.2 years not 3.2 months.

In Table 4 on page 400, there should not be a superscript *a* next to the ICAP heading. Also, in this table the coefficient for SIS Health and Safety subscale in Factor 3 should be −.16 not .16.

In the reference list, there should be reference to two versions of the Supports Intensity Scale (one unpublished version and one published version) as follows:

Thompson, J. R., Bryant, B., Campbell, E. M., Craig, E. M., Hughes, C., Rotholz, D. A., Schalock, R. L., Silverman, W., Tassé, M. J., & Wehmeyer, M. (2002). *Supports Intensity Scale: Standardization and users manual.* Unpublished assessment scale, American Association on Mental Retardation.

Thompson, J. R., Bryant, B., Campbell, E. M., Craig, E. M., Hughes, C., Rotholz, D. A., Schalock, R. L., Silverman, W., Tassé, M. J., & Wehmeyer, M. (2004). *Supports Intensity Scale: Users manual.* Washington, DC: American Association on Mental Retardation.

0196-206X/06/2702-0145
Developmental and Behavioral Pediatrics
Copyright © 2006 by Lippincott Williams & Wilkins, Inc.

Vol. 27, No. 2, April 2006
Printed in U.S.A.

Treatment

Early Intensive Behavioral Treatment: Replication of the UCLA Model in a Community Setting

HOWARD COHEN, Ph.D.
Valley Mountain Regional Center, Stockton, CA

MILA AMERINE-DICKENS, M.S.
Central Valley Autism Project, Modesto, CA

TRISTRAM SMITH, Ph.D.
Department of Pediatrics, University of Rochester Medical Center, Rochester, NY

ABSTRACT. Although previous studies have shown favorable results with early intensive behavioral treatment (EIBT) for children with autism, it remains important to replicate these findings, particularly in community settings. The authors conducted a 3-year prospective outcome study that compared 2 groups: (1) 21 children who received 35 to 40 hours per week of EIBT from a community agency that replicated Lovaas' model of EIBT and (2) 21 age- and IQ-matched children in special education classes at local public schools. A quasi-experimental design was used, with assignment to groups based on parental preference. Assessments were conducted by independent examiners for IQ (Bayley Scales of Infant Development or Wechsler Preschool and Primary Scales of Intelligence), language (Reynell Developmental Language Scales), nonverbal skill (Merrill-Palmer Scale of Mental Tests), and adaptive behavior (Vineland Adaptive Behavior Scales). Analyses of covariance, with baseline scores as covariates and Year 1-3 assessments as repeated measures, revealed that, with treatment, the EIBT group obtained significantly higher IQ (F = 5.21, p = .03) and adaptive behavior scores (F = 7.84, p = .01) than did the comparison group. No difference between groups was found in either language comprehension (F = 3.82, p = .06) or nonverbal skill. Six of the 21 EIBT children were fully included into regular education without assistance at Year 3, and 11 others were included with support; in contrast, only 1 comparison child was placed primarily in regular education. Although the study was limited by the nonrandom assignment to groups, it does provide evidence that EIBT can be successfully implemented in a community setting. *J Dev Behav Pediatr 27:145–155, 2006.* Index terms: *autism, early intervention, applied behavior analysis, behavioral treatment.*

The design and implementation of methodologically rigorous treatment studies are daunting tasks and, in the area of treatment for autism spectrum disorders, often emotionally charged and publicly vetted as well. Matching groups on a variety of important measures, including severity of disability, individual characteristics of the child, multiple important socio-familial and environmental factors, as well as controlling multiple treatment issues such as fidelity, intensity and length of treatment and pre-determining appropriate outcome measures are all challenging (and expensive). Moving treatment studies from the laboratory setting into the community presents additional hurtles, yet this is ultimately the setting in which the efficacy of treatment models needs to be evaluated. Cohen and colleagues are to be commended for implementing a community-based treatment study with matched samples, documentation of treatment fidelity, and comprehensive 3-year follow-up. However, the setting was based in a community program that is mandated to provide treatment to families of children with autism spectrum disorders who are then free to accept a plan or not, which prohibited random assignment to treatment. This introduced potential bias in their groups, with more educated and dual parent families in the EIBT group. There are strengths as well as limitations in this study. Although it does not resolve the controversies that continue regarding the "best" treatments for young children with ASD, we include it because of the critical need for evaluation of treatment approaches. The reviewers pointed out the limitations in this community approach as well as its strengths. The reader is encouraged to look at both in reviewing this article. We hope that it will inspire others to do these vitally needed treatment effectiveness studies. —Editor

Received September 2005; accepted February 2006.

Address for reprints: Mila Amerine-Dickens, M.S., 1317 Oakdale Rd., Suite 800, Modesto, CA 95355; e-mail: mamerine-dickens@cvap.org.

In an era when Autistic Spectrum Disorder (ASD) was viewed as largely untreatable,[1] Ivar Lovaas' 1987 outcome study[2] became a pivotal event that provided optimism about behavioral interventions for ASD. Almost half (9 of 19) of the children with autism who began intensive behavioral treatment prior to the age of 4 years from the UCLA/Lovaas clinic (40 hours per week for 2 or more years) were fully included into regular education and showed significant gains in intellectual achievement. A follow-up study of the same children showed sustained gains.[3] This finding, coupled with a general trend toward earlier diagnosis of ASD (under 3 years of age)[4] and the recent exponential increase in documented cases of ASD,[5] made Lovaas' results even more influential and replication of his research more compelling.

Replication of the UCLA/Lovaas Model involves the following key elements[6]: (1) clinical internship and training on the UCLA/Lovaas Model of intervention under the direction of qualified supervisors; (2) implementation of the model for 35 to 40 hours per week throughout the year, including one-to-one instruction, peer play training sessions, inclusion into regular education classrooms, and generalization activities; (3) parent training to foster the child's acquisition and generalization of skills; and (4) annual outcome measures.

Several studies have partially replicated the UCLA/Lovaas Model. In the only randomized clinical trial, 28 children with ASD received either intensive behavioral treatment or parent training.[7] The intensive treatment group averaged 25 hours per week in the first year which faded over the next 1 to 2 years. The comparison group participated in 10 to 15 hours per week of special education classes and received 5 hours per week of parent training for 3 to 9 months. The intensive children outperformed the comparison children on intellectual, visual-spatial, and academic measures. However, gains were substantially smaller than in Lovaas' original study. For example, the between-group IQ difference at follow-up was 16 points compared to the 31 reported by Lovaas. In other partial replications of the UCLA model, children with ASD obtained 15 to 35 hours per week of treatment and obtained results similar to those reported in the randomized clinical trials[8,9]; similar results also have been reported for other early intensive behavioral treatment (EIBT) models with about 25 hours per week of treatment.[10,11]

Concerns have been expressed about the difficulty of offering treatment at this level of intensity to community samples,[12] and mixed results of EIBT in community settings have been reported. One investigation indicated a lack of significant improvements in a sample of 66 children with ASD.[13] A multiple baseline study of 6 children found clear short-term gains but equivocal long-term effects.[14] However, a third study reported that an EIBT group (n = 29) in a community agency made statistically significant gains in all areas of development except motor skills, relative to 2 comparison groups.[15] Moreover, 13 of the 29 EIBT children (45%) achieved IQs in the average to above average range. In the first replication of the UCLA Model that included all of the elements identified by Lovaas, 11 of 23 children with ASD (48%) achieved full inclusion into regular education and IQ scores greater than 85.[16] However, the study did not have a comparison group.

Although these studies generally confirm that EIBT is effective, differing results across studies and methodological limitations such as the absence of comparison groups in many reports weaken the ability to truly validate the optimism generated by the initial Lovaas study. Accordingly, the present study was an attempt to fully replicate that study in a community setting. Research questions included the following: (1) Can the Lovaas/UCLA model be replicated in a community setting? (2) What outcomes do children with ASD achieve with this intervention?

METHODS

Participants

Participants were 42 children in 2 groups: The early intensive behavioral treatment (EIBT) group (n = 21) received 35 to 40 hours of behavioral intervention, 47 weeks per year, for 3 or more years. The comparison group (n = 21) received services from local public schools. In accord with the UCLA Young Autism Project multisite research replication protocol, participation criteria for both groups included (1) primary diagnosis of autistic disorder or pervasive developmental disorder not otherwise specified based on an evaluation by an independent licensed psychologist and confirmed by the Autism Diagnostic Interview–Revised,[17] (2) pretreatment IQ above 35 on the Bayley Scales of Infant Development–Revised (BSID-R),[18] (3) chronological age between 18 and 42 months at diagnosis and under 48 months at treatment onset, (4) no severe medical limitation or illness including motor or sensory deficits that would preclude a child from participating in 30 hours per week of treatment, (5) residence within 60 km of the treatment agency, (6) no more than 400 hours of behavioral intervention prior to intake, and (7) parent's agreement to participate actively in parent training and generalization and to have an adult present during home intervention hours.

In addition to the 21 participants in each group, there were 5 dropouts who were excluded from the data analyses (3 in the EIBT group and 2 in the comparison group). One EIBT participant moved out of the area at 17 months into treatment and was unavailable for follow-up; 2 withdrew their participation, 1 at 3 months and the other at 18 months. Dropouts were similar to completers with regard to age of diagnosis (24, 36, and 22 months), baseline IQ (42, 44, and 44), and 1-year IQ (58 and 61; score unavailable for participant who dropped out after 3 months). Two comparison children were dropped because parents either declined annual testing of their child or could not be contacted. All other eligible referrals enrolled in the study, completed yearly follow-up assessments, and were included in the data analyses.

All treatment in both groups was provided at no cost to families. Funding was split between 2 public agencies: (1) the Valley Mountain Regional Center (VMRC; Stockton, CA) and (2) the child's Special Education Local Planning Area (SELPA) of residence. VMRC is contracted by the California Department of Developmental Services to identify and coordinate services for individuals with developmental disabilities; its catchment area includes San Joaquin, Stanislaus, Calaveras, Amador, and Tuolumne Counties. SELPAs are contracted by the California Department of Education to provide special education instruction.

Design

Inasmuch as VRMC and SELPA had a mandate to provide free and appropriate services, legal and ethical considerations precluded random assignment of children to groups. Therefore, a quasi-experimental design was used. A comparison group was formed by identifying children who met participation criteria for EIBT and whose parents chose other services. Specifically, for each EIBT participant, a file review was initiated at VMRC to identify a matching child who was not receiving EIBT; the first identified child was then added to the comparison group. Comparison children were followed prospectively and received the same annual assessments as EIBT children.

To ensure that choices were available to families and that families were aware of these choices, VMRC and SELPA 6, along with nonpublic educational agencies and parents, developed an ongoing collaborative program (Autism Connection).[19] The Early Autism Diagnostic Clinic (EADC) was created by the Autism Connection (1) to provide expert evaluations for autism and related disorders (or referrals to other experts in the area) and (2) to bring together local clinicians, VMRC, parents, school district representatives, and advocates to communicate directly with each other, at the EADC, rather than requiring the parents to endure separate meetings. At the time of diagnosis, an educational consultant from the EADC and a representative from the school district of residence presented the family, orally and in writing, a Matrix of Educational Options developed by the Autism Connection. This matrix delineates the service agencies in the child's area of residence and their eligibility criteria, along with the roles and responsibilities of parents, service providers, and funding agencies in implementing interventions.

Options included special education settings, Autistic Spectrum Disorder (ASD) classes, speech and language services, occupational therapy, genetic counseling, behavior intervention services, grief counseling, Early Start programs for children under 3 years old, and EIBT Programs, including the agency in this study (Central Valley Autism Project; CVAP) and other EIBT providers. During the enrollment period (1995–2000), the number of other EIBT providers ranged from 1 to 3. At times when CVAP did not have openings, the education consultant and school representative removed CVAP from the Matrix. EADC educational consultant and school representatives were otherwise independent of the study.

Treatment Procedures: EIBT Group

EIBT consisted of 35 to 40 hours per week of intervention based on Lovaas' UCLA treatment model.[2,6,20] Seventeen of the 21 participants remained in EIBT for 3 years. Four others ended EIBT prior to 3 years but completed follow-up assessments and are included in the statistical analyses; 1 completed the intervention protocol and was fully included in regular education at Year 2, whereas 3 others were transferred to other services (2 after 6 months and 1 at Year 2) because their progress did not meet specific, predetermined developmental markers for continuing intervention. Markers at 6, 12, 24, and 36 months were identified collaboratively by Autism Connection.[21] For example, at 24 months, the IEP team considered whether the child showed one or more signs of progress such as the following: (1) the child's standardized cognitive testing indicated steady growth or near-average functioning; (2) objective data collected on EIBT instruction demonstrated that the child was mastering new skills; (3) objective data revealed an increase in the child's frequency of initiating language or peer interaction; or (4) the child was included in a general education placement with similar-aged peers for systematically increasing increments of time and was acquiring age-appropriate pre-academic skills.

The EIBT agency, CVAP, met all criteria for replication of Lovaas' UCLA treatment model and participated in a multicenter study supported by the National Institute of Mental Health. The UCLA model relies exclusively on behavioral techniques such as unambiguous instruction, shaping through positive reinforcement of successive approximations, systematic prompting and fading procedures, discrimination learning, and careful task analysis. Positive reinforcers such as edibles, sensory and perceptual objects are used initially but soon replaced by social reinforcers such as praise, tickles, hugs, and kisses. Ongoing data collection is performed to monitor skill acquisition, generalization, and frequency of problem behaviors. The intervention protocol consists of 3 primary components: (1) In-home 1:1 instruction, (2) peer play training, and (3) regular education classroom inclusion. No aversive interventions were used throughout the study.

Initially, the In-Home 1:1 Intervention Component is implemented 35 to 40 hours per week for children older than 3 years, and 20 to 30 hours per week for children younger than 3 years. The focus is on establishing foundational and spontaneous communication. The main teaching format is discrete trials,[22] but generalization activities and community outings are also part of the 35 to 40 hours per week of instruction. In discrete trials, the tutor works individually with a child in a distraction-free setting and administers 3 to 8 trials in a sitting, with 1- to 2-minute breaks between sittings, for approximately 50 minutes each hour. The remaining 10 minutes of each hour are devoted to generalization activities. These activities include structured play, in which the child has opportunities to apply skills initially mastered in the 1:1 setting (e.g., labeling toys or taking turns with the tutor during a game), and incidental teaching, in which situations were arranged to encourage initiation of language (e.g., placing preferred objects in sight but out of reach). Skill mastery in discrete trials was defined as 90% accuracy across 2 days of intervention, across 2 or more tutors. Concept mastery was defined as 90% accuracy of 5 to 10 novel items probed and mastered within a concept. After mastery, skills and concepts were systematically generalized to other more naturalistic settings and maintained by available contingencies in the natural environment. To facilitate generalization, community outings occurred 3 to 5 times per week. The UCLA curriculum was used for teaching the initial foundation skills including compliance, imitation, early receptive and expressive language, visual spatial skills, and self-help.[6,20]

At approximately 1 year into the behavioral intervention, the distribution of the 35 to 40 hours per week is typically as follows: 26 to 31 of home instruction, 3 to 5 hours of peer play, and 6 to 9 hours at preschool. Thereafter, the home component gradually decreases, whereas other components gradually increase based upon the child's inclusion in the classroom.

As part of the generalization of skills and behaviors to the natural environment, the peer play component is initiated 3 to 5 sessions per week with a typically developing peer for 15 to 60 minutes per session when the child has mastered prerequisite skills: verbal response to questions, on topic statements, simple play skills, and turn taking.[2,6,20] Skills mastered in the 1:1 setting are systematically generalized to a social/play setting with a peer of similar age. A trained tutor facilitates mastered activities for the child and peer (e.g., conversation, pretend play with toys, or turn-taking games) and prompts the peer to engage the child with subtle cues such as whispers in the peer's ear, visual signals, or indirect questions. When the child is 90% accurate initiating with peers across 3 or more peers for 18 to 24 months, additional children are presented at one time to form a group play setting.

At about the time that peer play training is initiated, the child enters a teacher-directed structured regular education preschool setting.[2] Initially, trained tutors accompany the child to school to assist the teaching staff with gaining instructional control, generalizing mastered skills to the school setting, and learning classroom skills. The tutor functions as a classroom aide and not as a 1:1 aide for the child. Initial goals for inclusion center on generalizing skills to a novel, yet structured environment. As the child achieves independent responding during specific activities (e.g. circle time, center time, and so forth), as determined by data, the shadow tutor is faded. Activities requiring social skills and behaviors are always the last to fade in the process.

When children have achieved typical levels of academic functioning in the classroom and participate without the assistance of a shadow tutor during teacher-directed activities, they still may require the assistance of the shadow tutor during social opportunities throughout the school day for an additional 2 to 3 years. Thus, an intervention with reduced hours both at home and in school may extend into the early primary grades. School hours focus on generalization of social skills and friendship development. As the child's rate of independent social interaction increases, the intervention hours are successively reduced to 0. Subsequently, consultation to the family and the school setting continue 1 to 2 hours per month for up to 1 to 2 years. Home hours focus on play sessions with peers and gradually transition to typical play dates with peers without the presence of a tutor. Periodic standardized assessments continue until the child is 18 years old.

During the course of the study, there was a growing recognition that many children who made significant gains in the first 2 years of treatment required training beyond the UCLA curriculum to develop mutually satisfying social relationships, enhance their understanding of social meanings, understand and interpret other's perspectives/knowledge/cognition/beliefs, and ultimately respond appropriately to social behaviors of peers and others. To address this need, overt social behaviors were operationally defined, both verbal (e.g., conversational skills, such as responding to statements or questions asked by others, reciprocal statements, initiating conversation, inquiring about others, remaining on topic, and sustaining conversation) and nonverbal (e.g., interpreting and responding to other's facial expressions, emotional states, voice tone, or body language), and initially taught in a discrete trial format, using the same behavioral principles and methodology described above, with an emphasis on a quick transition to generalized teaching to a social context, using incidental teaching and video modeling as tools for generalization.

Staff and Parent Training. To ensure proficiency in implementing the UCLA model, 5 CVAP staff members each completed 3- to 4-month internships at UCLA, and consultants from UCLA made on-site visits 2 to 4 times per year for the first 3 years of the study period, with frequent telephone contacts between visits (typically once per week). During this period, a random sample of 12 CVAP tutors were videotaped and scored by blind raters for adherence to UCLA procedures. The level of adherence by CVAP tutors was found to be nonsignificantly higher than adherence by tutors employed at UCLA.[23]

One UCLA-trained individual served as CVAP site director, responsible for oversight of each child's intervention; she holds a master's degree in clinical psychology/applied behavior analysis and is a Board Certified Behavior Analyst. Clinic supervisors trained and provided ongoing performance feedback to tutors. Supervisors were graduate students in behavior analysis or master's level clinicians with 2 or more years of experience in providing EIBT. Tutors were recruited from the community and were the main providers of direct services. Supervisors and tutors were assigned to each EIBT participant based on openings in their schedule and geographic location.

To become a supervisor, individuals had to meet prespecified, objective criteria, including high ratings based on direct observation of their implementation of EIBT interventions, favorable evaluations from families and staff members, satisfactory performance on a test of skill at curriculum development, and oral and written demonstration of their knowledge of applied behavior analysis and ASD.[24] Tutors had to pass a rigorous behavior observation assessment of their accuracy in conducting discrete trial training (DTT) and oral tests of their knowledge of the UCLA treatment manual.

Parents were encouraged to be involved in all levels of intervention. At the beginning of treatment, all parents attended a 12- to 18-hour training workshop across 2 to 3 days on behavioral principles and intervention methods. Thereafter, they participated in weekly training sessions to generalize their child's newly established skills to the natural environment. Parents provided ongoing information regarding their child's current level of functioning both in and out of intervention sessions, and they were asked to be active participants in their child's intervention, although there was no requirement for parents to provide any direct intervention hours.

Treatment Procedures: Comparison Group

Participants in the Comparison Group received community services that their families selected from the Matrix of Educational Options. At intake, 1 comparison child, under 3 years old, received an Early Start Autism Intervention Program, which emphasized learning readiness skills with both the parent and child. This child received less than 9 hours per week of a discrete trial program in his or her home, until the age of 3. Two comparison children received a home-based developmental intervention that ranged from 1 to 4 hours a week. At age 3, these 3 children were enrolled in a public school Special Day Class (SDC). Seventeen children who were 3 and above at intake were enrolled in SDC in the public schools. No records were available for 1 child. The instructional methodology in the SDC placements was eclectic, the child/teacher ratios varied from 1:1 to 3:1, and the classes operated for 3 to 5 days per week, for up to 5 hours per day. Related services such as speech, occupational, and behavioral therapy to these children varied from approximately 0 to 5 hours per week Three of the children spent brief sessions (up to 45 minutes per day) mainstreamed in regular education. Due to the diverse interventions provided to the comparison group, it was not possible to monitor treatment fidelity for this group.

Assessment

At pretreatment, a licensed psychologist at EADC who was independent of the study administered a standardized behavior observation,[25] parent interview, and developmental tests, including the BSID-R, Merrill-Palmer Scale of Mental Tests,[26] Reynell Developmental Language Scales,[27] and Vineland Adaptive Behavior Scales.[28] The BSID-R extrapolated table was used to generate a standard score for children who obtained an IQ below 50.[29] Administration of the BSID-R began at the starting point for the child's chronological age (or at the highest starting point for the test if the child was older than 42 months). The examiner administered each successive item after the starting point to establish a basal and ceiling; if the child did not obtain a basal on these items, the examiner administered each preceding item in succession until a basal was achieved and then followed rules in the test manual for establishing the ceiling.

From the evaluation, the psychologist made a DSM-IV diagnosis of autism or Pervasive Disorder, Disorder Not Otherwise Specified (PDDNOS).[30] Subsequently, the diagnosis was confirmed by the Autism Diagnostic Interview-Revised (ADI-R),[17] administered by a certified examiner employed by CVAP. The developmental tests (but not the ADI-R) were repeated in annual follow-up evaluations. If a participant performed at the ceiling of the BSID-R, this test was replaced with the Wechsler Preschool and Primary Scales of Intelligence.[31] Follow-up evaluations were conducted by an independent, self-employed, highly-skilled, licensed, child evaluator. VMRC made the referral and funded the evaluations. The referral to the evaluator consisted only of the name of the child, birth date, parent's names, and telephone number.

Data Analysis

IQ was the main measure of treatment response in previous EIBT studies[6–16] and was designated as the primary outcome measure in the present study. Secondary outcome measures were the Merrill-Palmer Scale of Mental Tests, Reynell Language Comprehension, Reynell Expressive Language, Vineland Adaptive Behavior Scales, and classroom placement.

To test our main hypothesis that the EIBT group would differ from the comparison group on outcome measures, we performed a repeated-measures analysis of covariance (ANCOVA) for each measure, with pretreatment score as the covariate and Year 1, Year 2, and Year 3 scores as the repeated dependent measures. Consistent with standard assumptions for an ANCOVA,[32] analyses of skew and kurtosis, as well as visual inspection, were consistent with a normal distribution in our data. Hyunh-Feldt epsilon tests confirmed that the data showed compound symmetry ($\varepsilon >$ .90), unless otherwise noted in Results.

As is usual in outcome studies with repeated measures, a few participants had missing data at one or more time points. For each outcome measure, we employed the standard procedure of removing participants with missing data from the analysis.[32] This procedure is appropriate when missing data are random or unbiased. We used visual inspection to confirm that the missing data were unbiased (e.g., the data were not primarily from participants who had unfavorable outcomes or who did not complete the full 3 years of intervention), and –Results– show the number of participants retained for each analysis.

In as much as the EIBT and comparison groups differed on several demographic variables (mother education, father education, and diagnosis), we explored whether adding these variables as covariates in the ANCOVA model would change the interpretation of the results. These analyses need to be interpreted with caution because they involve a larger number of variables than is usually considered appropriate for the relatively small sample size in the present study. However, they provided some information on whether or not the groups differed when we statistically controlled for demographic variables.

When an ANCOVA revealed a between-group difference on an outcome measure, we hypothesized that the EIBT group would show an increase in scores from Year 1 to Year 2 to Year 3, whereas scores in the comparison group would remain stable. To test this hypothesis, we examined whether the ANCOVA yielded a statistically significant Group × Time interaction; if so, we performed planned comparisons to test for an increase from Year 1 to Year 3 in the EIBT group.

Pet. Reh. App.49

Table 1. Background Information for the EIBT Group (n = 21) and Comparison Group (n = 21)

	EIBT	Comparison
Demographics		
Male/Female	18:3	17:4
Diagnosis (Autism/PDDNOS)*	20:1	15:6
Age at diagnosis [(*M*(SD)]	30.2 (5.8)	33.2 (3.7)
Mother education, yr [(*M*(SD)]*	15.3 (2.9)	13.1 (1.6)
Father education, yr [(*M*(SD)]*	15.8 (2.9)	11.8 (2.3)
Two-parent household (yes/no)*	21:0	14:7
Pretreatment Test Scores [(*M*(SD)]		
IQ	61.6 (16.4)	59.4 (14.7)
Merrill-Palmer	82.4 (17.3)	73.4 (11.9)
Reynell		
Language Comprehension	51.7 (15.2)	52.7 (15.1)
Expressive Language	52.9 (14.5)	52.8 (14.4)
VABS		
Composite	69.8 (8.1)	70.6 (9.6)
Communication	69.4 (11.8)	65.0 (6.8)
Daily Living	73.2 (9.2)	72.7 (12.5)
Socialization	70.3 (10.9)	75.1 (13.0)

EIBT indicates early intensive behavioral treatment; Reynell, Reynell Developmental Language Scales; VABS, Vineland Adaptive Behavior Scales; PDDNOS, Pervasive Disorder, Disorder Not Otherwise Specified.

*Significant difference between EIBT and comparison group ($p < .05$).

To examine the clinical significance of the results, we ascertained the number of participants in each group who achieved scores in the average range at follow-up on each measure. We also sought to identify pretreatment measures that were associated with later scores in the average range. Therefore, for the EIBT group, we conducted *t*-tests to compare pretreatment scores of participants who scored in the average range across all measures to pretreatment scores of the remaining participants.

RESULTS

Pretreatment

Table 1 summarizes the demographics and pretreatment scores of the early intensive behavioral treatment (EIBT) and comparison groups. The gender make-up mirrors the 4:1 male to female ratio in Autistic Spectrum Disorder (ASD).[31] Twenty of 21 EIBT children (95%) and 15 of 21 comparison children (71%) were diagnosed with Autistic Disorder. This difference was statistically significant, $t(40) = 2.13$, $p < .05$. The remaining children were classified with Pervasive Disorder, Disorder Not Otherwise Specified (PDDNOS). Age of diagnosis was 20 to 41 months, with the EIBT group averaging 3 months younger than the comparison group (a difference that was not statistically significant). Also, as shown in Table 1, although not a requirement for participation in the EIBT program, parents had significantly more education and were significantly more likely to be married than comparison parents. IQ, Merrill-Palmer, Reynell, and Vineland scores did not differ significantly between groups; scores in both groups indicated developmental delays comparable to other samples of children with ASD.[30]

Outcome

Table 2 presents the results of the analysis of covariance (ANCOVA) tests for each outcome measure, whereas Figure 1 presents the means and 95% confidence intervals for each group at intake, Year 1, Year 2, and Year 3. As shown in Table 2, there was a significant difference between groups on the primary outcome measure, IQ. Figure 1 reveals that the mean IQ in the EIBT group increased 25 points, from 62 at pretreatment to 87 at Year 3. Interestingly, the mean IQ in the comparison group also increased, from 59 at pretreatment to 73 at Year 3.

The EIBT and comparison groups did not differ significantly on the Merrill-Palmer. Both groups displayed a mean increase of 13 points from intake to Year 3 on this measure. Figure 1 suggests that the groups may not have been matched at pretreatment, as the mean for the EIBT was 82 compared to 73 in the comparison group. A post hoc analysis indicated that this difference approached statistical significance, $t(35) = 1.87$, $p = .07$. Also, the assumption of compound symmetry was questionable for this variable, with Hyunh-Feldt $\epsilon = .85$; because the

Table 2. Analyses of Covariance Testing for Differences Between the EIBT and Comparison Groups on Outcome Measures

	N		Sums of Squares (Between Subjects)				
Measure	E	C	Group	Covariate	Error	MSE	F
IQ	21	19	4,229.91	12,046.14	30,042.41	811.96	5.21*
Merrill-Palmer	21	16	246.27	15,613.74	20,657.91	626.00	ns
Reynell							
Language Comprehension	21	19	3,750.25	17,523.60	36,312.08	981.41	3.82**
Expressive Language	20	19	3,413.57	13,590.90	52,495.66	1,458.21	ns
VABS							
Composite	20	20	3,897.52	1,589.31	18,385.69	496.91	7.84***
Communication	20	20	3,937.71	2,937.53	25,994.10	722.06	5.45*
Daily Living	20	20	2,527.14	2,229.25	14,207.49	394.65	6.40*
Socialization	20	20	1,857.84	21.66	16,130.41	460.87	4.03**

N indicates number of participants included in the analysis; E, EIBT group; C, comparison group; ns, not statistically significant; MSE, mean square of errors (between subjects); Reynell, Reynell Developmental Language Scales; VABS, Vineland Adaptive Behavior Scales.

* $p < .05$; ** $p < .10$; *** $p < .01$.

Pet. Reh. App.50

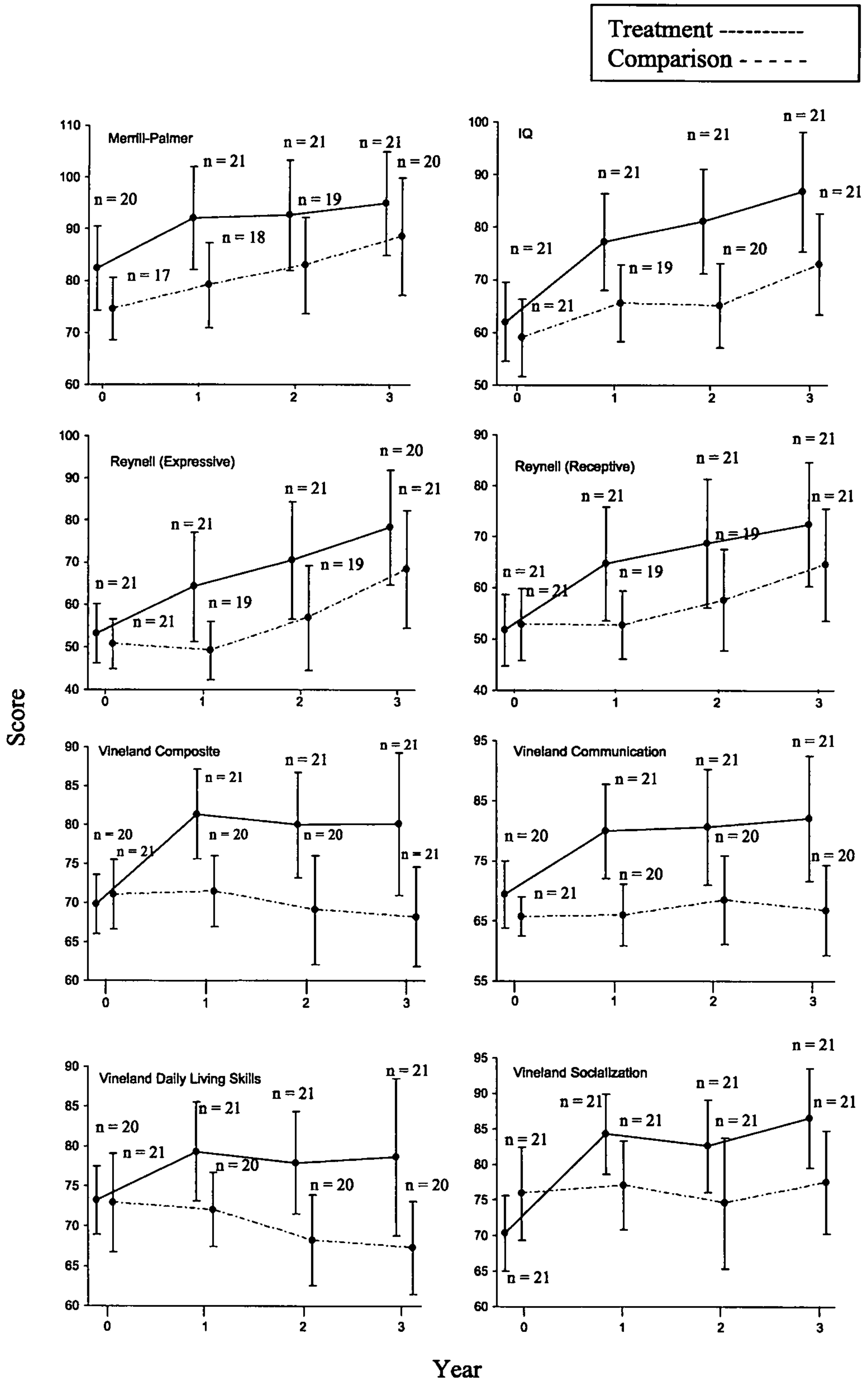

FIGURE 1. Mean and 95% confidence interval for pretreatment (Year 0) and follow-up (Years 1–3).

Table 3. Number of Children in the Average Range on each Outcome Measure for the EIBT Group (n = 21) and Comparison Group (n = 21)

Measure	EIBT	Comparison	*p*
IQ	12	7	ns
Language Comprehension[a]	8	4	ns
Expressive Language[a]	9	6	ns
VABS Composite[b]	8	3	.10
School Placement	6	0	.001

[a]Reynell Developmental Language Scales.
[b]Vineland Adaptive Behavior Scales.

ANCOVA did not approach statistical significance, alternate analyses were not attempted.

There was a trend toward a significant difference in Reynell Language Comprehension (*p* = .06). The mean score in the EIBT group increased 20 points, from 52 at pretreatment to 72 at Year 3; the mean score in the comparison group increased 9 points, from 53 at pretreatment to 62 at Year 3. The EIBT group also had a larger increase from pretreatment to Year 3 in Reynell Expressive Language (53–78, compared to 51–66), but this difference was not statistically significant (*p* = .13). The failure to find a significant difference may indicate that EIBT did not have a meaningful effect on expressive language, or it may simply reflect low statistical power to detect an effect.

The EIBT and comparison groups differed significantly in the Vineland Adaptive Behavior Scales Composite. Consistent with this finding, the EIBT group demonstrated a mean increase of 9 points compared to a 4-point decline in the comparison group, as shown in Figure 1. Inasmuch as a difference was observed in the Composite, individual scales were also analyzed. Significant differences between groups were found in Communication and Daily Living Skills, and a trend was found for Socialization (*p* = .05). Figure 1 indicates that the changes in scores from pretreatment to Year 3 for each scale were similar to the change in Composite scores. These findings support the inference that the EIBT group had more advanced adaptive behavior skills than the comparison group at the time of the outcome assessments.

An analysis of classroom placement at year 3, between the 2 groups, revealed that 17 of the 21 EIBT children and 1 of the 21 comparison children were included into regular education classroom settings. Of the 17 EIBT children, 6 were fully included without assistance, 4 were fading the shadow tutor, and 7 required full shadows.

When mother' education, father's education, or diagnosis was added as a covariate to the ANCOVA model, ANCOVA was unaltered, except in one instance: With the father's education as a covariate, the difference between groups in IQ was not statistically significant (*p* = .11). It is unclear whether this finding indicates that father's education was a confound or reflects the limited statistical power for the analysis. When mother's education, father's education, and diagnosis were all added as covariates to the ANCOVA model, IQ, Reynell Language Comprehension, and Vineland Composite continued to show a trend toward significance (*p* = .09 for all 3 outcome measures). In sum, the possibility that father's education was a confound in the analysis of IQ cannot be ruled out, but the remaining analyses indicated that reliable differences in outcome between groups remained after statistically controlling for inequalities at pretreatment.

None of the analyses for group × time interactions were statistically significant. Thus, we did not confirm our hypothesis that the EIBT group would have increasing scores from Year 1 to Year 2 to Year 3, whereas scores in the comparison group would be stable. On the contrary, Figures 1 and 2 illustrates that although the EIBT group appeared to make larger increases than the comparison group from pretreatment to Year 1, both groups exhibited stable scores from Year 1 to Year 3 in IQ, Merrill-Palmer, and Vineland. Both groups may have exhibited similar increases in scores in Reynell Language Comprehension and Expressive Language from Year 1 to Year 3.

As shown in Table 3, more EIBT participants than comparison participants achieved follow-up scores in the average range for each measure, although this difference was significant only for school placement and showed a trend toward significance for the Vineland. Ten EIBT participants scored in the average range on all measures (6 of these 10 also were included in regular education without assistance, whereas the remaining 4 continued to receive shadowing in the regular education classroom). *t*-tests did not reveal any significant differences in pretreatment test scores for these 10 participants compared to the remaining 11 participants. For example, these 10 children had a mean pretreatment IQ of 66.6 (SD = 12.4) compared to 57.7 (SD = 19.0) for the remaining 11 children, *t*(19) = 1.28, ns. However, pretreatment Reynell Language Comprehension scores showed a trend toward a difference, with a pretreatment mean of 58.1 for the participants with the most favorable outcome compared to 45.9 for the other participants, *t*(19) = 1.98, *p* = .06.

DISCUSSION

The present study suggests that the UCLA/Lovaas Model of early intensive behavioral treatment (EIBT) can be implemented in a nonuniversity community-based setting. On the primary outcome measure of IQ, the EIBT group showed a gain of 25 points, which was statistically significant compared to the gain of 14 points in the comparison group. Similar effects were found on measures of adaptive behavior. Although language comprehension showed a trend towards significance, expressive language and nonverbal cognitive skill revealed no difference between groups. The increases in test scores are similar to those reported in Lovaas' original EIBT study[2,3] and in some recent investigations.[15,16] However, the difference between the EIBT group and the comparison group on outcome measures was smaller than that in other studies, as the comparison group also made gains.

Pet. Reh. App.52

An important limitation of the study is that, because treatment was funded by public agencies that were required to offer free and appropriate services, groups could not be randomly assigned, and a quasi-experimental design was used, with parents choosing which group their child entered. Although pretreatment test scores did not differ significantly between groups, other pretreatment variables did differ. The EIBT group had more children with autism and fewer with Pervasive Disorder, Disorder Not Otherwise Specified (PDDNOS) than did the comparison group. To the extent that PDDNOS is a milder diagnosis that may have a more favorable prognosis than autism,[7] this difference may have favored the comparison group. However, the EIBT group also may have had an advantage in that it had more 2-parent families and better educated families than did the comparison group. These family variables have not been associated with outcome in previous studies,[2,7] but they might have encouraged families to select EIBT over other interventions in the present study, even though all interventions were provided at no cost to families. In addition, these variables might have given the EIBT group an advantage by making it easier for families to participate in treatment sessions and facilitate generalization of skills outside of treatment. After statistically controlling for family variables, outcome analyses continued to show improved outcomes in the EIBT group relative to the comparison group. Nevertheless, statistical controls are not a satisfactory solution for preexisting group differences, especially given the relatively small sample size in the present study. A design with random assignment would have strengthened the study and allowed for more clearcut conclusions about whether EIBT is effective or not.

Further limitations pertain to the assessment protocol in the study. As previously noted, the comparison group received such diverse interventions that a measure of treatment fidelity could not be applied. Also, outside evaluators were employed by Valley Mountain Regional Center (VMRC) for pretreatment and follow-up assessments of participants. The referrals to the evaluators did not include information on group assignment or treatment history. However, to ensure that evaluators remained unaware of this information and to allow for checks on the reliability of test administration and scoring, evaluators who were employed by the study and conducted assessments at a research site (rather than in their clinical offices) might have been preferable. Another limitation is that the assessment protocol tested developmental level more rigorously than did the features of Autistic Spectrum Disorder (ASD). The inclusion of the Autism Diagnostic Observation Schedule (ADOS),[33] in addition to the Autism Diagnostic Interview–Revised (ADI-R) and clinical diagnosis, would have increased confidence in the initial diagnosis. Including a measure such as the ADOS in follow-up assessments would have indicated whether or not children continued to display behaviors indicative of ASD. Additional measures such as the Theory of Mind Test[34] also would help address this issue; Central Valley Autism Project (CVAP) is currently involved in a study to translate this test into English and standardize it in the United States. Without such measures, the present study cannot address one of the most controversial issues raised by previous EIBT research–whether some children become indistinguishable from typically developing peers[6] or whether they continue to display characteristics of ASD. An additional follow-up evaluation of study participants with the ADOS and Theory of Mind (TOM) Test is planned to fill in some of these gaps.

In this study, advanced behaviors associated with friendship initiation and maintenance, social skills, understanding of social meaning, and response to social behaviors were identified and treated, using the same discrete trial methodology as other behaviors, which consequently increased the duration of treatment beyond 3 years for many participants (usually for 2 additional years). Although this expansion of the treatment protocol reflects the contemporary view that the defining feature of ASD is an impairment in social reciprocity, it raises the question of whether the present study truly was a replication of the UCLA model. The treatment site met all of Lovaas' criteria for replication, and the first 2 years of intervention followed the model as it has been previously described.[2] The third year also followed the model, with the addition of the training in advanced social skills. Thus, results from Years 1 and 2 are directly comparable to those of previous studies, and results from Year 3 also reflect mostly the same interventions. Research on the specific effects of the additional social skills training is warranted, as it is acknowledged that such training was not included in previous studies. Also, although discrete trial training is a common approach to teaching social skills and has some empirical support,[35,36] teaching methodologies other than discrete trials (e.g. video modeling, incidental teaching) also have empirical support and may have advantages such as generalizing more quickly to settings outside of treatment;[22] thus, the question of how best to teach such skills may be another area for research.

Interestingly, although the EIBT protocol lasted for 3 years and, in some cases, was continued beyond that time, the nonsignificant group × time interactions in the statistical analyses indicates that the EIBT group did not show reliable IQ increases relative to the comparison group after Year 1. A possible explanation is that most gains occurred in the first year of intervention. Alternatively, however, it is also possible that gains took place later in treatment but that the study measures were not sensitive to them.

Potential evidence for the latter view comes from the findings on classroom placement. A striking result was that, despite IQ gains in the comparison group, all participants but 1 remained primarily in a special education classroom setting, whereas most EIBT participants were included in regular education at least part of the day. Classroom placement is a controversial outcome measure because of concerns that it may reflect factors such as parent advocacy and school policy rather than the child's functioning.[12] However, the measure also may be an index of real-world academic and social competence.[37] If so, the differences between groups on this measure may be attributable at least in part to the social skills training that EIBT participants received. In addition, it may suggest a need for a high number of treatment hours. Dismantling studies might help address these possibilities.

The initial collaborative funding efforts by VMRC and Special Education Local Planning Areas (SELPAs) resulted in a sustainable treatment environment. Stable funding, effective guidelines and policies, and positive communication and working relationships were primary contributory variables to the feasibility of this study. Thus, this collaboration may be a useful model for other regions to employ. Other clinical strengths of this study included rigorous treatment quality control criteria, stringent staff training and evaluation standards, multiple internships at UCLA by supervising clinicians, precise programming for each individual child, advanced completion programming and skilled generalization training, yearly follow-ups by an independent evaluator using multiple outcome measures, and a centralized process and standardized protocol for diagnosing children and informing families of EIBT and other intervention options available to them. Without such standards, outcomes may differ. Nevertheless, given the methodological limitations of the present research, there is a continued need for rigorous outcome studies comparing EIBT to control conditions or other interventions.

Acknowledgments. Preparation of this manuscript was partially supported by NIMH grants R01 MH 48663 (Multi-site Young Autism Project) and U54 MH066397 (Genotype and Phenotype of Autism). The authors thank Kym Cassaretto, M.S., CVAP Clinical Director, and CVAP staff for their clinical contributions and committed efforts on behalf of children and families. The authors also thank the following individuals for their assistance with the research: Mieke San Julian, Chanti Fritzsching, M.S., BCBA, and Angela Castro from CVAP; Schelley McDonald and Marie Overmyer from VMRC; and Suzannah Ferraioli from URMC. Preliminary reports were presented at the California Association for Behavior Analysis Conference in San Francisco, CA, February 20, 2004, and at the International Meeting For Autism Research, Boston, MA, May 6, 2005.

REFERENCES

1. DeMyer MK, Hingtgen JN, Jackson RK. Infantile autism: a decade of research. *Schizophr Bull.* 1981;7:388–451.
2. Lovaas OI. Behavioral treatment and normal educational and intellectual functioning in young autistic children. *J Consult Clin Psychol.* 1987;55:3–9.
3. McEachin JJ, Smith T, Lovaas OI. Long-term outcome for children with autism who received early intensive behavioral treatment. *Am J Ment Retard.* 1993;97:359–372.
4. Zwaigenbaum L, Bryson S, Rogers T, et al. Behavioral manifestations of autism in the first year of life. *Int J Dev Neurosci.* 2005; 23:143–152.
5. California Department of Developmental Services. Department of Developmental Services Fact Book, 7th ed. California Department of Developmental Disabilities website. December, 2004. Available at www.dds.ca.gov/factsstats/factbook.cfm#pdf. Accessed August 16, 2005.
6. Lovaas OI. *Teaching Individuals with Developmental Delays: Basic Intervention Techniques.* Austin, TX: PRO-ED; 2003.
7. Smith T, Groen AD, Wynn JW. Randomized trial of intensive early intervention for children with pervasive developmental disorder. *Am J Ment Retard.* 2000;105:269–285.
8. Anderson SR, Avery DL, DiPietro EK, et al. Intensive home-based early intervention with autistic children. *Educ Treat Child.* 1987; 10:352–366.
9. Birnbrauer JS, Leach DJ. The Murdoch Early Intervention Program after two years. *Behav Change.* 1993;10:63–74.
10. Harris SL, Handleman JS. Age and IQ at intake as predictors of placement for young children with autism: a four- to six-year follow-up study. *J Autism Dev Disord.* 2000;30:137–142.
11. Weiss M. Differential rates of skill acquisition and outcomes of early intensive behavioral intervention for autism. *Behav Interv.* 1999;14:3–22.
12. Schopler E, Short A, Mesibov G. Relation of behavioral treatment to "normal functioning": comment on Lovaas. *J Consult Clin Psychol.* 1989;57:162–164.
13. Bibby P, Eikeseth S, Martin NT, et al. Progress and outcomes for children with autism receiving parent-managed intensive interventions. *Res Dev Disabil.* 2002;23:81–104.
14. Smith T, Buch GA, Gamby TE. Parent-directed, intensive early intervention for children with pervasive developmental disorder. *Res Dev Disabil.* 2000;21:297–309.
15. Howard JS, Sparkman CR, Cohen HG, Green G, Sanislaw HA. Comparison of intensive behavior analytic and eclectic treatments for young children with autism. *Res Dev Disabil.* 2005;26:359–383.
16. Sallows GO, Graupner TD. Intensive behavioral treatment for children with autism: four year outcome and predictors. *Am J Ment Retard.* 2005;110:417–438.
17. Lord C. Follow-up of two-year-olds referred for possible autism. *J Child Psychol Psychiatry.* 1995;36:1365–1382.
18. Bayley N. *Bayley Scales of Infant Development*, 2nd ed. San Antonio, TX: The Psychological Corporation; 1993.
19. Cohen HG. Pyramid building: Partnership as an alternative to litigation. In: Lovaas OI ed. *Teaching Individuals with Developmental Delays: Basic Intervention Techniques.* Austin, TX: PRO-ED, 2003:375–386.
20. Lovaas OI. *Teaching Developmentally Disabled Children: The ME Book.* Austin, TX: PRO-ED; 1981.
21. Region 6 Autism Connection. *Early Intensive Behavioral Treatment 4-way Agreement.* Stockton, CA: Region 6 Autism Connection; 2004.
22. Smith T. Discrete trial training in the treatment of autism. *Focus Autism Relat Disord.* 2000;16:86–92.
23. Mortenson S, Smith T. Quality Control in the Multisite Young Autism Project. Paper presented at: Annual Meeting of the Association for Behavior Analysis; May 1996; San Francisco, CA.
24. Davis BJ, Smith T, Donahoe P. Evaluating supervisors in the UCLA treatment model for children with autism: validation of an assessment procedure. *Behav Ther.* 2002;31:601–614.
25. California Department of Developmental Services. *Best Practice Guidelines for Screening, Diagnosis, and Assessment. Ethological Observation Schedule (ETHOS).* Sacramento, CA: California Department of Developmental Services; 2002.
26. Stutsman R. *Guide for Administering the Merrill-Palmer Scale of Mental Tests.* New York: Harcourt, Brace & World; 1948.
27. Reynell JK. *Reynell Developmental Language Scales.* Windsor, England: Nfer-Nelson; 1990.

28. Sparrow SS, Balla DA, Cicchetti DV. *Vineland Adaptive Behavior Scales*. Circle Pines, MN: American Guidance Service; 1984.
29. Robinson BF, Mervis CB. Extrapolated raw scores for the second edition of the Bayley Scales of Infant Development. *Am J Ment Retard.* 1996;100:666–671.
30. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*, 4th ed, Text Revision. Washington, DC: American Psychiatric Association; 2000.
31. Wechsler D. *Manual for the Wechsler Intelligence Scale for Children*, 3rd ed. San Antonio, TX: Psychological Corporation; 1991.
32. Nich C, Carroll K. Now you see it, now you don't: a comparison of traditional versus random-effects regression models in the analysis of longitudinal follow-up data from a clinical trial. *J Consult Clin Psychol.* 1997;65:252–261.
33. Lord C, Rutter M, DiLavore PC, et al. *Autism Diagnostic Interview Schedule*. Los Angeles: Western Psychological Services; 2001.
34. Steerneman P, Meesters C, Muris P. *TOM-Test*. Antwerpen-Appledoorn: Garant; 2003.
35. Taylor BA, Jasper S. Teaching programs to increase peer interaction. In: Maurice C, Green G, Foxx M eds. *Making a Difference: Behavioral Intervention for Autism*. Austin, TX: Pro-Ed; 2001: 97–162.
36. Weiss MJ, Harris SL. Reaching out, joining. In: *Teaching Social Skills to Young Children with Autism*. Bethesda, MD: Woodbine House; 2001.
37. Kazdin A. Replication and extension of behavioral treatment of autistic disorder. *Am J Ment Retard.* 1993;97:382–383.

Pet. Reh. App.55

Research in Developmental Disabilities 35 (2014) 3326–3344

Contents lists available at ScienceDirect

Research in Developmental Disabilities

ELSEVIER

Comparison of behavior analytic and eclectic early interventions for young children with autism after three years

Jane S. Howard [a,b,1,*], Harold Stanislaw [a], Gina Green [c], Coleen R. Sparkman [b,1], Howard G. Cohen [d]

[a] *California State University, Stanislaus, Psychology Department, 1 University Circle, Turlock, CA 95382, USA*
[b] *The Kendall Centers/Therapeutic Pathways, Modesto, CA 95354, USA*
[c] *Association of Professional Behavior Analysts, 6977 Navajo Road #176, San Diego, CA 92119, USA*
[d] *Valley Mountain Regional Center, 702 North Aurora St, Stockton, CA 95202, USA*

ARTICLE INFO

Article history:
Received 16 June 2014
Received in revised form 6 August 2014
Accepted 12 August 2014
Available online

Keywords:
Autism
Early intervention
Applied behavior analysis
Eclectic treatment
Outcomes
Longitudinal studies

ABSTRACT

In a previous study, we compared the effects of just over one year of intensive behavior analytic intervention (IBT) provided to 29 young children diagnosed with autism with two eclectic (i.e., mixed-method) interventions (Howard, Sparkman, Cohen, Green, & Stanislaw, 2005). One eclectic intervention (autism programming; AP) was designed specifically for children with autism and was intensive in that it was delivered for an average of 25–30 h per week ($n = 16$). The other eclectic intervention (generic programming; GP) was delivered to 16 children with a variety of diagnoses and needs for an average of 15–17 h per week. This paper reports outcomes for children in all three groups after two additional years of intervention. With few exceptions, the benefits of IBT documented in our first study were sustained throughout Years 2 and 3. At their final assessment, children who received IBT were more than twice as likely to score in the normal range on measures of cognitive, language, and adaptive functioning than were children who received either form of eclectic intervention. Significantly more children in the IBT group than in the other two groups had IQ, language, and adaptive behavior test scores that increased by at least one standard deviation from intake to final assessment. Although the largest improvements for children in the IBT group generally occurred during Year 1, many children in that group whose scores were below the normal range after the first year of intervention attained scores in the normal range of functioning with one or two years of additional intervention. In contrast, children in the two eclectic treatment groups were unlikely to attain scores in the normal range after the first year of intervention, and many of those who had scores in the normal range in the first year fell out of the normal range in subsequent years. There were no consistent differences in outcomes at Years 2 and 3 between the two groups who received eclectic interventions. These results provide further evidence that intensive behavior analytic intervention delivered at an early age is more likely to produce substantial improvements in young children with autism than common eclectic interventions, even when the latter are intensive.

[*] Corresponding author at: California State University, Stanislaus, Psychology Department, 1 University Circle, Turlock, CA 95382, USA. Tel.: +1 209 667 3386; fax: +1 209 993 8192.
E-mail addresses: jhoward@csustan.edu, janeshoward@mac.com (J.S. Howard), hstanislaw@csustan.edu (H. Stanislaw), ggreen@apbahome.net (G. Green), csparkman@tpathways.org (C.R. Sparkman).
[1] Address: PO Box 5157, Modesto, CA 95352, USA.

http://dx.doi.org/10.1016/j.ridd.2014.08.021

1. Introduction

The past two decades have seen increased interest in early intervention for children diagnosed with autism spectrum disorder (hereafter, "autism") among researchers, policymakers, funding sources, and consumers. Following publication of the Lovaas study in 1987, a number of researchers began evaluating the effects of intensive, comprehensive early intervention using applied behavior analysis (ABA) methods. Various ABA models for treating children with autism have been proposed, but many behavior analytic researchers agree that genuine early intensive ABA treatment programs have certain key features in common: (a) individualized, comprehensive intervention that addresses all skill domains; (b) use of multiple behavior analytic procedures (not just discrete-trial procedures or "naturalistic" techniques) to build new repertoires and reduce behaviors that interfere with skill acquisition and effective functioning; (c) direction and oversight by one or more professionals with advanced training in ABA and experience with young children with autism; (d) reliance on typical developmental sequences to guide selection of treatment goals; (e) parents and other individuals trained by behavior analysts to serve as active co-therapists; (f) intervention that is initially one-to-one, transitioning gradually to a group format as warranted; (g) intervention that often begins in homes or specialized treatment centers but is also delivered in other environments, with gradual, systematic transitions to regular schools when children develop the skills required to learn in those settings; (h) planned, structured intervention provided for a minimum of 20–30 h per week with additional hours of informal intervention provided throughout most other waking hours, year round; (i) intensive intervention beginning in the preschool years and continuing for at least 2 years (Eldevik et al., 2010; Green, Brennan, & Fein, 2002).

Substantial research has documented the effectiveness of treatments that incorporate all of the foregoing features. Eight prospective studies used comparison- or control-group designs to evaluate some variation of the Lovaas/UCLA model of early intensive ABA intervention for children with autism (Cohen, Amerine-Dickens, & Smith, 2006; Eikeseth, Smith, Jahr, & Eldevik, 2002; Eikeseth, Smith, Jahr, & Eldevik, 2007; Eldevik, Hastings, Jahr, & Hughes, 2012; Eldevik, Eikeseth, Jahr, & Smith, 2006; Lovaas, 1987; Sallows & Graupner, 2005; Smith, Groen, & Wynn, 2000). In another three studies, the ABA intervention was designed and overseen by professional behavior analysts not affiliated with Lovaas, and the ABA intervention differed somewhat from the Lovaas model (Howard, Sparkman, Cohen, Green, & Stanislaw, 2005; Remington et al., 2007; Zachor, Ben-Itzchak, Rabinovitch, & Lahat, 2007). Outcomes from those 11 studies varied and some children had larger improvements than others. In the large majority of cases, however, the mean change scores achieved by children receiving intensive ABA treatment exceeded the mean change scores for similar children in control or comparison groups who received less intensive ABA treatment, intensive or non-intensive treatment using a mixture of methods or therapies ("eclectic" treatment), or "treatment as usual" (i.e., standard early intervention or special education services). Additionally, compared to children who received other types of treatment, children who received early intensive ABA treatment were more likely to achieve post-treatment scores on one or more standardized measures that were in the normal range, and were more often placed in regular classrooms (for reviews and analyses, see Eikeseth, 2009; Eldevik et al., 2009, 2010; Green, 2011; National Autism Center, 2009; Reichow & Wolery, 2009; Rogers & Vismara, 2008).

Despite the evidence from multiple studies and meta-analyses favoring intensive ABA treatment for autism over other models of early intervention, a number of questions persist. One is whether other types of early intervention delivered with comparable intensity and individualization can produce outcomes comparable to ABA. Perhaps the most common alternative early intervention approach involves a mixture of methods drawn from ABA, speech-language pathology, occupational therapy (especially sensory integration techniques), developmental psychology, and autism-specific approaches. That model, which has been characterized as "eclectic" intervention, is widely available in the United States and elsewhere.

At least three studies have compared eclectic and ABA interventions directly. Eikeseth et al. (2002) studied children with autism who entered treatment at ages 4–7 years ($M = 5.5$ years), slightly older than children in most of the other studies of early intensive behavioral intervention. One group ($n = 13$) received Lovaas-model ABA treatment for 28 h per week, while a second group ($n = 12$) received eclectic intervention for 29 h per week. There were no significant differences between the groups when treatment began. Both forms of treatment were delivered in public school classrooms. After 1 year, the ABA treatment group had gained an average of 17 points on IQ test scores, 13 points on tests of language comprehension, 27 points on tests of expressive language, and 11 points on an adaptive behavior scale. The eclectic treatment group had average gains of only 4 points on IQ tests and 1 point on language tests, and no change in adaptive behavior. A follow-up study conducted when those children were 8 years old found that after about 3 years of treatment, the ABA treatment group had gained an average of 25 IQ points and 9–20 points on adaptive behavior scales in comparison to baseline. The eclectic intervention group had a mean gain of only 7 points on IQ tests, and declines of 6–12 points on adaptive behavior assessments (Eikeseth et al., 2007).

A study we published previously involved a comparison of intensive ABA intervention with two different eclectic intervention models (Howard et al., 2005). Twenty-nine preschool children with autism received early intensive behavior analytic intervention (IBT), 16 received intensive eclectic intervention designed for children with autism (designated the autism programming, or AP, group), and an additional 16 received typical non-intensive, eclectic early intervention services (designated the generic programming, or GP, group). All children began intervention prior to 48 months of age and received treatment for an average of 14 months. They were placed in treatment groups on the basis of parental preferences and education team decisions, and evaluated pre-treatment and annually thereafter by professionals who were neither involved in nor employed by any of the treatment programs. The three groups were shown to be similar on key variables when treatment began. After 14 months of intervention, mean scores on standardized tests of intellectual, communication, and adaptive skills were significantly higher for children in the IBT group than for children in the other two groups. Children in the IBT group had an average standard IQ score of 90, compared to 62 and 69 for children in the AP and GP groups, respectively. Developmental trajectories for most measures accelerated markedly over the 14 months of treatment for children in the IBT group, while the trajectories for children in the other two groups remained flat or declined.

For the present study, we followed children who participated in the 2005 study through an additional 2 years of treatment. We focused on four questions: (a) did Year 1 differences in the cognitive, language, and adaptive behavior scores of children in the three groups persist? (b) Did differences in the developmental trajectories of the three groups at Year 1 change during Years 2 and 3? (c) How many children in each group had standardized test scores in the normal range after 2 or 3 years of treatment? (d) To what extent were outcomes at Year 1 correlated with outcomes at Years 2 and 3?

2. Method

2.1. Participants

The same 61 children who participated in the Howard et al. (2005) study participated in this follow-up. Characteristics of the groups at intake are reported in Howard et al. (2005).[2]

Assessments were conducted 1–3 years after treatment began, but not all skill domains were assessed each year with every child. (See Section 3.1 for number of assessments available for each group at intake and at Years 1–3.) In particular, one child in the GP group and one child in the IBT group did not receive any assessments after the first year of treatment. Nonetheless, scores for all 61 children were retained for the present analyses to permit evaluations of outcomes that were not included in our 2005 publication.

2.2. Treatments

Information about the treatments participants received, school placements, and number of hours and services authorized during Years 2 and 3 was obtained through file review.

2.2.1. Intensive behavior analytic treatment (IBT)

This treatment was designed and delivered by personnel in a California non-public agency that provides ABA services to children with autism. Treatment was directed by the first author, a Board Certified Behavior Analyst-Doctoral® (BCBA-D®) and licensed psychologist, and the fourth author, a licensed speech-language pathologist. Programs were supervised by Board Certified Behavior Analysts® (BCBAs®) and other staff with master's degrees in psychology or special education and some training in ABA. They were supported by staff who were either Board Certified Assistant Behavior Analysts® (BCaBAs®) or who had bachelors degrees, most of whom were enrolled in graduate programs in ABA and related areas. Treatments were delivered to children by behavior technicians working under the supervision of the clinical staff. Behavior technicians began delivering treatment only after they had passed competency-based performance evaluations; thereafter, they were directly observed and received written or oral feedback on their implementation of behavior change protocols from their clinical supervisors an average of once or twice each week.

To varying degrees, all parents helped support treatment outside of formal treatment hours. Parent training initially focused on teaching instruction-following, promoting spontaneous language, re-directing nonfunctional repetitive behavior, managing interfering behaviors, and building skills such as toileting, dressing, and independent play. Parents were also trained to implement behavior analytic procedures that were designed to increase success in activities relevant to health and self-care, such as cooperating with medical and dental care procedures and participating in sports and other community activities.

Treatment was delivered in multiple settings, including homes, treatment centers, community settings, and regular preschool and elementary school classrooms. Treatment protocols utilized the full range of behavior analytic procedures, customized to each child's level of functioning, preferences, family circumstances, and treatment goals. Each child received an average of 35–40 h of treatment per week. The adult:child ratio during Year 1 was 1:1, but during Years 2 and 3 the ratio was gradually decreased (e.g., to 1:2 or 1:3, and then to one adult per small group of children), depending on progress and treatment targets. For further details, see Howard et al. (2005).

[2] While assembling the data for this study, we uncovered several errors in data reported in our 2005 paper. Most were minor (e.g., 1-month errors in the child's age), but the baseline scores of one child in the GP group were reported incorrectly as Year 1 scores, and the Year 1 scores of another child in the GP group were reported as baseline scores. Correcting those errors had virtually no impact on the conclusions that were drawn in the 2005 paper; all 107 of the statistical tests reported as not significant in 2005 remained non-significant, and 40 of the 43 findings that were reported as statistically significant in 2005 remained so. The three exceptions were for group differences that were only marginally significant in the 2005 publication: the difference at intake between the mean nonverbal age equivalents for the AP and GP groups changed from $p = 0.04$ to $p = 0.07$; the difference at follow-up between the mean motor standard scores of the IBT group and the two comparison groups changed from $p = 0.04$ to $p = 0.06$; and, when the mean self-help skills learning rates before and after treatment were compared, the difference between the IBT group and the two comparison groups changed from $p = 0.05$ to $p = 0.07$. Revised tables reporting all corrections are available as supplementary materials.

Pet. Reh. App.58

Data from standardized assessments as well as direct observation and measurement of target behaviors guided decision-making about the distribution of treatment hours across targets and settings. Initial treatment targets focused on foundational repertoires (e.g., attending, imitating vocal and motor sequences, following spoken directions, receptive and expressive labeling, initiating requests, tolerating change, etc.) that are often absent or at low levels in children with autism. Treatment targets during Years 2 and 3 generally focused on advanced cognitive, social, play, self-care, academic, and communication skills (for example, see Fischer, Howard, Sparkman, & Moore, 2009). More complex interactions involving peers and siblings generally occurred during Years 2 and 3 than in Year 1. On average, children in the IBT group had more than 200 goals on their annual individualized education programs (IEPs).

When children acquired the skills necessary to benefit from small group instruction (e.g., learning through observing the behavior of others, language skills close to the level of instruction, low levels of problem behaviors, independent communication of basic needs), they were placed in preschool or kindergarten programs for typically developing children for up to 15 h per week. Each child was accompanied by a behavior technician who used a variety of behavior analytic approaches, including self-management and behavioral contracting procedures, to arrange opportunities to prompt and reinforce behavior targets in order to promote skill acquisition and generalization across settings. The clinical supervisor directing the intervention also provided training and consultation to parents, teachers, and other professionals. Sample behaviors targeted in the regular classrooms included following instructions from classroom teachers and aides, engaging in classroom routines, and interacting with peers. Time spent with typically developing peers was gradually increased based on skill acquisition, maintenance and generalization of skills, and level of problem behaviors. Most children did not enter kindergarten until age 6.

2.2.2. Autism programming (AP) and generic programming (GP)

Brief descriptions of the AP and GP interventions are presented next; for details see Howard et al. (2005). The AP programs were designed specifically for children with autism. Intervention procedures were drawn from the Training and Education of Autistic and Related Communication Handicapped Children (TEACCH) approach, sensory integration therapy, commercially available programs (e.g., the Picture Exchange Communication System; Bondy & Frost, 1994), and some behavior analytic procedures, such as discrete-trial procedures. Children in this group received an average of 25–30 h of intervention per week in public school classrooms with staffing ratios of 1:1 or 1:2. Thus, the AP programs provided eclectic intervention at an intensity that was comparable to IBT.

The GP intervention was delivered in special education classrooms that served children with a variety of diagnoses and educational needs. Programming that was described as "developmentally appropriate" and "language rich" was provided for an average of 15–17 h per week, with slightly more hours as children approached age 6. Adult:child ratios averaged 1:6.

Approximately one third of the children in both the AP and GP treatment groups received "pull out" speech therapy sessions of less than 30 minutes once or twice a week during Years 1 and 2. About 20% of the children in both groups received some services in general education classrooms, which often included such activities as lunch, physical education, or recess. On average, each child in the AP and GP groups had fewer than 15 goals on his/her annual IEP.

2.2.3. Summary

All of the children in the IBT group and the majority in the two eclectic treatment groups had similar placements and programming during Years 2 and 3 as in Year 1. Some children changed from one eclectic treatment to the other after Year 1, while the Year 2 and/or Year 3 intervention was not available for a few AP and GP children. This information is summarized in Fig. 1, which is similar to a Sankey diagram. Sankey diagrams are used in engineering (see Schmidt, 2008, for an overview) and vary the width of each arrow in proportion to the number represented by that arrow. Thus, in the GP treatment group, the arrow leading from GP treatment in Year 1 to AP treatment in Year 2 (which represents $n = 3$ children) is three times as wide as the arrow leading from AP treatment in Year 2 back to GP treatment in Year 3 (representing $n = 1$ child).

2.3. Design

We utilized a between-groups design to compare performances of children in the IBT group with those of children in the two eclectic treatment groups at intake and at followup assessments about 1–3 years later. As reported in Howard et al. (2005), the three groups of children were substantially similar on most key variables at intake. The only significant differences were in mean chronological ages and parents' education, which were controlled for statistically (see Section 2.3.2 below).

2.3.1. Dependent measures

The principal dependent measures in this study were scores on full-scale IQ tests (cognitive skills), measures of language development, and adaptive behavior scales (composite scores as well as communication, self-help, and social skills scores). Scores on nonverbal IQ tests, receptive and expressive communication skills assessments, and motor skills were also analyzed. Since these latter skills were often not measured in Year 3, we report the Year 2 scores if Year 3 scores were not available.

All intake and follow-up assessments were conducted by experienced, qualified examiners who were not involved in treating any children in any of the groups. Assessments were conducted in the child's home, in the examiner's office, at a

Pet. Reh. App.59

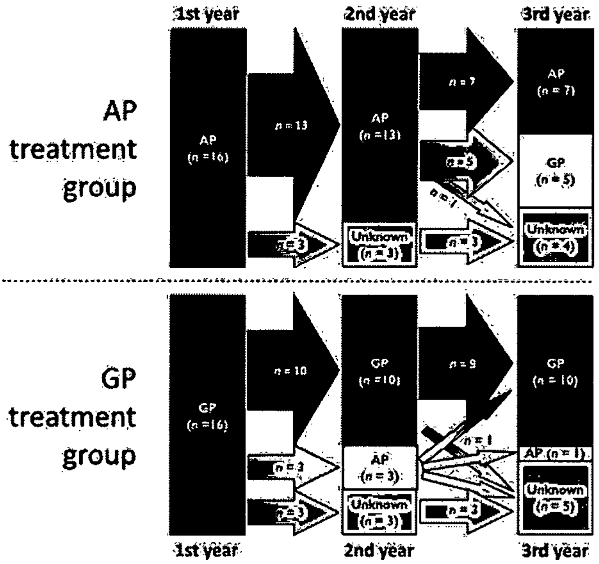

Fig. 1. Movement of children between AP and GP treatments by year. Children in the IBT treatment group had the same treatment all three years. Agency files did not report the type of treatment that was received during Years 2 and 3 for some of the children who initially received the AP or GP treatments.

Table 1
Age (in months) at each assessment, and interval between intake and each subsequent assessment.

Measure	IBT		AP		GP		IBT mean minus AP/GP mean	AP mean minus GP mean
	M	SD	*M*	SD	*M*	SD		
Age at diagnosis	30.07	5.30	39.31	5.52	34.94	5.18	−7.06**	4.38*
Age at intake testing	30.86	5.16	37.44	5.68	34.75	4.80	−5.23**	2.69
Age at Year 1 follow-up	45.24	5.84	50.69	5.64	49.06	5.64	−4.63**	1.63
Age at Year 2 follow-up	57.64	5.30	63.21	5.86	62.23	6.15	−5.10**	0.98
Age at Year 3 follow-up	69.24	5.01	74.33	5.98	73.46	6.10	−4.69**	0.87
Months between intake and Year 1	14.31	2.22	13.25	2.84	14.31	2.44	0.53	−1.06
Months between intake and Year 2	27.05	1.91	25.36	1.82	26.85	3.11	0.97	−1.49
Months between intake and Year 3	37.90	2.98	37.13	2.36	38.46	2.30	0.15	−1.33

* $p < 0.05$.
** $p < 0.01$.

school, or in the settings of local non-profit entities (Regional Centers) that contracted with the state to manage services to persons with developmental disabilities. As reported in Howard et al. (2005), Year 1 testing occurred an average of 14.3 months after intake. Thereafter, parents of all children were contacted annually to determine if they were interested in having their children participate in follow-up assessments. Table 1 shows the mean ages of the groups at each assessment and the intervals between assessments. On average, Year 2 testing occurred 23–34 months after intake (M = 27.0 months), and Year 3 assessments occurred 31–43 months after intake (M = 37.9 months).

The examiners selected standardized tests of cognitive skills, language skills, and adaptive behavior that were suited to each child's age and level of functioning. Howard et al. (2005) described the instruments used at intake and at Year 1. After Year 1, adaptive behavior was assessed using the Vineland Adaptive Behavior Scales (VABS). Nonverbal IQ was assessed using the Merrill-Palmer Scales of Development (although the Leiter International Performance Scale was used for one child in the IBT group in Year 3). Full-scale IQ was typically assessed after Year 1 using the developmentally appropriate Wechsler instrument, either the Wechsler Preschool and Primary Scale of Intelligence (WPPSI-III or WPPSI-Revised) or the Wechsler Intelligence Scale for Children (WISC-III or WISC-IV). However, one child in the IBT group was administered the Stanford-Binet Intelligence Scale (the 4th edition in Year 2 and the 5th edition in Year 3), and in Year 3 two children in the IBT group were administered the Differential Ability Scales, one IBT child was administered the Slosson Intelligence Test-Revised, and one IBT child was administered the Woodcock-Johnson Tests of Cognitive Abilities III. Receptive and expressive language skills were assessed using a variety of instruments. The most common was the Reynell Developmental Language Scales. Others included the Receptive One-Word Picture Vocabulary Test, the Expressive One-Word Picture Vocabulary Test, the Peabody Picture Vocabulary Test (3rd edition), the Expressive Vocabulary Test, and the Sequenced Inventory of Communication Development-Revised.

Measures for which developmental equivalents were available were converted to developmental quotients (DQs) for analysis using the formula DQ = 100 × developmental equivalent (months)/chronological age (months). When all children are the same age, there is no statistical difference between analyzing standard scores (SSs), developmental equivalents, and DQs. Unlike the other two measures, however, DQs allow valid comparisons to be made among children who have different chronological ages at the same assessment time, and automatically compensate for different intervals between assessment times (cf. Delmolino, 2006; Lord & Schopler, 1989).

2.3.2. Statistical analyses

As in our original study, statistical analyses focused on comparing the mean scores of children in the IBT group with those of children in the AP and GP groups; comparing the mean scores of children in the AP group with those of children in the GP group was of secondary interest. Accordingly, in this study we used the same multiple regression approach we employed in Howard et al. (2005). One term in the regression equation was a contrast that compared mean scores of the children in the IBT group with mean scores of the children in the AP and GP groups, while a second contrast term (orthogonal to the first) compared the mean scores of children in the AP group with the mean scores of children in the GP group. Both contrasts were tested simultaneously, together with two covariates (chronological age at diagnosis and parents' mean years of education) to control for group differences in the covariates.

Separate multiple regression analyses were performed for each of the four assessment times (intake, Year 1, Year 2, and Year 3). Repeated measures analyses examining all four assessment times at once were precluded because not all children were assessed at every follow-up. Restricting the analyses to children with complete assessment records would have eliminated more than half of the children from some analyses. Trends over the 3-year course of treatment were examined by using paired *t*-tests to compare each child's score at one assessment with his or her score at the following assessment.

For every dependent measure, we also determined whether each child achieved a favorable outcome. This was defined as a DQ or SS within the normal range of functioning (i.e., 85 or higher), or a DQ or SS that was at least 15 points (1 standard deviation) higher at the final assessment (Year 2 or Year 3) than at intake. This definition is logically similar to the reliable change index proposed by Jacobson and Truax (1991) for evaluating the effects of treatments. Chi-square tests were used to determine whether the percentage of children with a favorable outcome differed by treatment group, with a separate analysis conducted for each dependent measure.

3. Results

3.1. Ages and assessment times

The assessment chronology for all three groups is summarized in Table 1. Cells in the first five rows include descriptive statistics on chronological ages at diagnosis and at each assessment time. Data in the bottom three rows describe elapsed time between intake and later assessments. Data in the two rightmost columns represent comparisons of group means; asterisks indicate statistically significant differences. These data indicate that, at diagnosis and every subsequent assessment, the average child in the IBT group was younger than the average child in either comparison group; those differences were statistically significant. There was also a statistically significant difference between the mean ages of the AP and GP children at diagnosis, but not at any of the later assessments.

3.2. Analyses of standard scores and developmental quotients

Table 2 presents descriptive statistics and analyses of assessments of cognitive and adaptive skills for each group. Adaptive behavior scores (communication, social, and self-help skills) are expressed as developmental quotients (DQs), while cognitive skills scores and the composite adaptive behavior measure are expressed as standard scores (SSs). For each of the five measures, cells in the first four rows under each group's column list descriptive statistics from each assessment time; results of statistical comparisons of group means at each assessment time are shown in the two rightmost columns. All comparisons controlled for the child's age at diagnosis and the parents' years of education. Asterisks indicate statistical significance. As shown in the two rightmost columns, all Year 1 and Year 2 mean SSs and DQs were significantly higher for the IBT group than for the two comparison groups combined. There were no other statistically significant between-group differences; the mean scores for the IBT group and the two comparison groups combined did not differ significantly at intake, and the mean scores of the AP and GP groups did not differ significantly from each other at intake or at any of the other assessment times on any measure.

The cells in the bottom three rows for each of the five measures in Table 2 summarize changes in mean scores between successive assessments (intake to Year 1, Year 1 to Year 2, and Year 2 to Year 3). Asterisks denote statistically significant improvements (positive values) or declines (negative values) from one year to the next. The IBT group had statistically significant improvements on all measures from intake to Year 1. The AP group had a statistically significant improvement on the cognitive skills SS from intake to Year 1, and statistically significant declines in the self-help DQ and adaptive behavior composite SS from Year 1 to Year 2. The GP group had a statistically significant improvement on the social skills DQ from Year 1 to Year 2. No other changes were statistically significant.

The cells in the penultimate column in the bottom three rows for each measure in Table 2 represent comparisons of the mean change scores of the IBT group and the AP and GP groups combined. Asterisks indicate statistically significant differences in change scores between intake and Year 1 on all measures in favor of the IBT group. The cells in the rightmost

Pet. Reh. App.61

Table 2
Cognitive and adaptive skills scores at intake and Years 1–3, changes between successive assessments, and differences between groups.

Measure	Assessment	IBT group			AP group			GP group			IBT mean minus AP/GP mean	AP mean minus GP mean
		n	*M*	SD	*n*	*M*	SD	*n*	*M*	SD		
Cognitive (SS)	Intake	28	60.57	17.48	16	53.69	13.50	15	61.00	15.27	3.35	−7.31
	Year 1	26	89.88	20.87	16	62.13	19.63	16	69.13	15.13	24.26**	−7.00
	Year 2	22	86.59	20.47	13	59.38	13.38	13	66.08	17.35	23.86**	−6.69
	Year 3	21	89.43	23.99	14	64.43	24.84	13	71.77	19.45	21.47	−7.34
	Intake vs Year 1	25	27.44**	14.18	16	8.44*	15.04	15	8.27	17.20	19.09**	0.17
	Year 1 vs Year 2	20	−2.05	6.89	13	1.77	6.69	13	−1.69	6.79	−2.09	3.46
	Year 2 vs Year 3	17	2.94	9.06	11	1.09	8.62	11	6.55	13.45	−0.88	−5.45
Communication (DQ)	Intake	29	49.23	14.77	16	42.95	15.01	15	47.22	14.39	4.22	−4.27
	Year 1	26	76.47	22.48	16	46.52	21.63	16	52.38	22.34	27.02**	−5.86
	Year 2	22	73.46	28.93	13	45.94	20.50	13	51.26	22.56	24.86*	−5.33
	Year 3	19	72.84	29.46	14	49.40	29.36	13	57.68	22.74	19.45	−8.29
	Intake vs Year 1	26	26.23**	21.27	16	3.57	17.96	15	4.82	21.34	22.05**	−1.25
	Year 1 vs Year 2	20	−0.02	21.08	13	2.22	18.46	13	4.33	15.46	−3.29	-2.11
	Year 2 vs Year 3	16	2.57	21.79	11	4.68	14.68	11	5.68	10.67	−2.61	−1.00
Self-help (DQ)	Intake	29	60.65	16.59	16	56.63	15.54	15	60.30	17.13	2.24	−3.67
	Year 1	26	70.19	16.39	16	62.17	16.49	16	56.39	10.29	10.91*	5.78
	Year 2	22	69.35	16.99	13	52.13	16.46	13	60.68	16.39	12.95**	−8.55
	Year 3	19	67.13	15.78	14	56.11	19.25	13	61.92	13.67	8.22	−5.80
	Intake vs Year 1	26	9.86*	21.24	16	5.54	20.52	15	−3.41	17.93	8.65*	8.95
	Year 1 vs Year 2	20	0.68	17.85	13	−8.53*	14.07	13	5.46	15.26	2.22	−13.99*
	Year 2 vs Year 3	16	−2.66	8.33	11	3.31	18.15	11	1.71	17.01	−5.17	1.60
Social (DQ)	Intake	28	54.43	18.00	16	58.05	23.15	15	56.92	15.50	−3.07	1.13
	Year 1	26	69.47	19.05	16	58.44	28.89	16	49.55	20.21	15.47*	8.89
	Year 2	22	80.66	26.60	13	48.68	25.52	13	61.16	30.09	25.75**	−12.48
	Year 3	19	79.33	27.48	14	52.60	25.56	13	67.85	22.40	19.39	−15.25
	Intake vs Year 1	25	15.21**	22.99	16	0.39	23.86	15	−7.47	23.09	18.62*	7.86
	Year 1 vs Year 2	20	11.85**	17.14	13	−7.93	17.77	13	15.68*	20.45	7.98	−23.61**
	Year 2 vs Year 3[a]	16	3.37	14.81	11	1.57	17.41	11	7.27	19.86	−1.05	−5.70
Composite (SS)	Intake	26	72.00	7.73	16	69.81	10.48	13	69.69	8.54	2.24	0.12
	Year 1	26	81.15	10.95	16	69.25	12.91	16	67.63	10.45	12.72**	1.63
	Year 2	22	79.14	12.06	13	61.08	10.14	13	69.77	12.74	13.71**	−8.69
	Year 3	20	76.00	15.94	14	58.07	15.80	13	65.85	12.23	14.19	−7.77
	Intake vs Year 1	24	8.88**	10.94	16	−0.56	12.04	13	−1.31	11.24	9.77*	0.75
	Year 1 vs Year 2	20	−0.60	10.75	13	−6.54*	8.41	13	4.54	8.29	0.40	−11.08**
	Year 2 vs Year 3	16	−4.31	8.25	11	−4.09	11.23	11	−3.45	7.41	−0.54	−0.64

[a] Age at diagnosis is a significant covariate ($p < 0.05$).
* $p < 0.05$.
** $p < 0.01$.

column in the bottom three rows for each measure represent comparisons of the mean change scores between the AP and GP groups. Statistically significant differences were seen in the changes in self-help DQ, social DQ, and adaptive behavior composite SS from Year 1 to Year 2; in each case, the mean score declined for the AP group but increased for the GP group.

Table 3, which is formatted similarly to Table 2, presents nonverbal IQ, language, and motor skills DQs by group. Those skills were not assessed consistently after Year 1; some children were reassessed only after 2 years of treatment, while others were reassessed only after 3 years of treatment. Accordingly, Table 3 shows statistics for only three points in time: at intake; one year after intake; and either three years after intake or – if no assessment was made at Year 3 – two years after intake. For the IBT group, mean DQs on all measures except motor skills (which were in the normal range at intake) increased significantly from intake to Year 1. Expressive language skills also improved significantly from Year 1 to Year 2/3, while the motor skills mean DQ showed a decline over that same interval. For the AP group, the only statistically significant change was a decline in the mean motor skills DQ from Year 1 to Year 2/3. Over that same interval the GP group's mean expressive language DQ increased significantly, but no other statistically significant changes were found for that group. The penultimate right-hand column of Table 3 again shows several statistically significant differences in change scores favoring the IBT group over the two combined eclectic groups. As indicated in the final column, the only statistically significant difference between the AP and GP groups was on the motor skills DQ assessment at Year 2/3, with children in the AP group scoring lower, on average, than children in the GP group.

The group mean data described in Table 2 are represented graphically in Figs. 2–6, which also show individual scores and the numbers of children in each group whose scores were in the normal range (85 and above) on each measure at each assessment. Data from assessments conducted at intake and Year 1 mirror those reported in our previous paper (Howard et al., 2005): Mean scores on all measures for all three groups were comparable at intake, and the IBT group had significantly higher mean scores one year later than did the AP and GP groups. As discussed above, change scores indicated that, on average, children in the IBT group improved more than children in the other two groups after one year of treatment.

Pet. Reh. App.62

Table 3
Nonverbal IQ, language, and motor skills scores at intake, Year 1, and Year 2 or 3, changes between assessments, and differences between groups.

Measure	Assessment	IBT treatment group			AP treatment group			GP treatment group			IBT mean minus AP/GP mean	AP mean minus GP mean
		n	*M*	SD	*n*	*M*	SD	*n*	*M*	SD		
Non-verbal (DQ)	Intake	20	80.44	12.06	16	67.00	17.13	11	76.65	13.37	9.51	−9.65
	Year 1[b]	24	101.04	18.27	16	73.60	24.79	15	81.08	18.72	23.82^{**}	−7.47
	Year 2/3	24	98.05	17.92	15	69.33	22.08	14	82.20	21.74	22.50^{**}	−12.87
	Intake vs Year 1	20	20.31^{**}	14.97	16	6.61	18.56	11	2.42	13.42	15.41^{**}	4.19
	Year 1 vs Year 2/3	21	−2.20	8.65	15	−2.10	13.63	13	2.78	11.49	−2.36	−4.87
Receptive (DQ)	Intake[b]	29	48.79	20.87	16	45.44	15.23	15	47.29	13.59	2.45	−1.85
	Year 1	26	71.23	21.97	15	51.39	22.44	14	51.95	19.46	19.57^{*}	−0.55
	Year 2/3	25	74.46	25.08	15	49.53	24.61	13	60.31	18.75	19.93^{*}	−10.78
	Intake vs Year 1	26	22.53^{**}	18.31	15	5.27	13.12	13	2.77	11.96	18.42^{**}	2.50
	Year 1 vs Year 2/3	24	2.18	12.39	14	−0.27	20.81	12	4.39	10.49	0.31	−4.66
Expressive (DQ)	Intake	29	49.73	16.34	16	43.90	5.80	15	50.20	12.16	2.78	−6.30
	Year 1	26	69.24	23.20	15	47.31	24.42	14	48.08	14.35	21.56^{*}	−0.77
	Year 2/3	26	83.25	29.88	15	47.98	27.25	13	62.07	23.83	28.73^{*}	−14.10
	Intake vs Year 1	26	20.46^{**}	22.36	15	3.42	22.68	13	−2.84	12.29	19.95^{*}	6.26
	Year 1 vs Year 2/3	24	10.40^{*}	17.11	14	1.51	16.47	12	12.34^{*}	17.75	3.90	−10.82
Motor (DQ)	Intake	28	94.65	17.50	16	89.55	13.09	14	86.96	13.34	6.31	2.59
	Year 1	26	97.30	14.74	16	85.08	12.24	16	85.62	13.62	11.95^{*}	−0.54
	Year 2/3	25	90.17	12.64	12	74.00	13.24	14	86.31	15.83	9.54	-12.30^{*}
	Intake vs Year 1	25	0.63	18.23	16	−4.46	12.82	14	0.83	18.18	2.63	−5.29
	Year 1 vs Year 2/3	23	-8.44^{*}	18.05	12	-9.82^{*}	13.54	14	1.97	20.57	−4.97	−11.78

[b] Mean parental years of education is a significant covariate ($p < 0.05$).
* $p < 0.05$.
** $p < 0.01$.

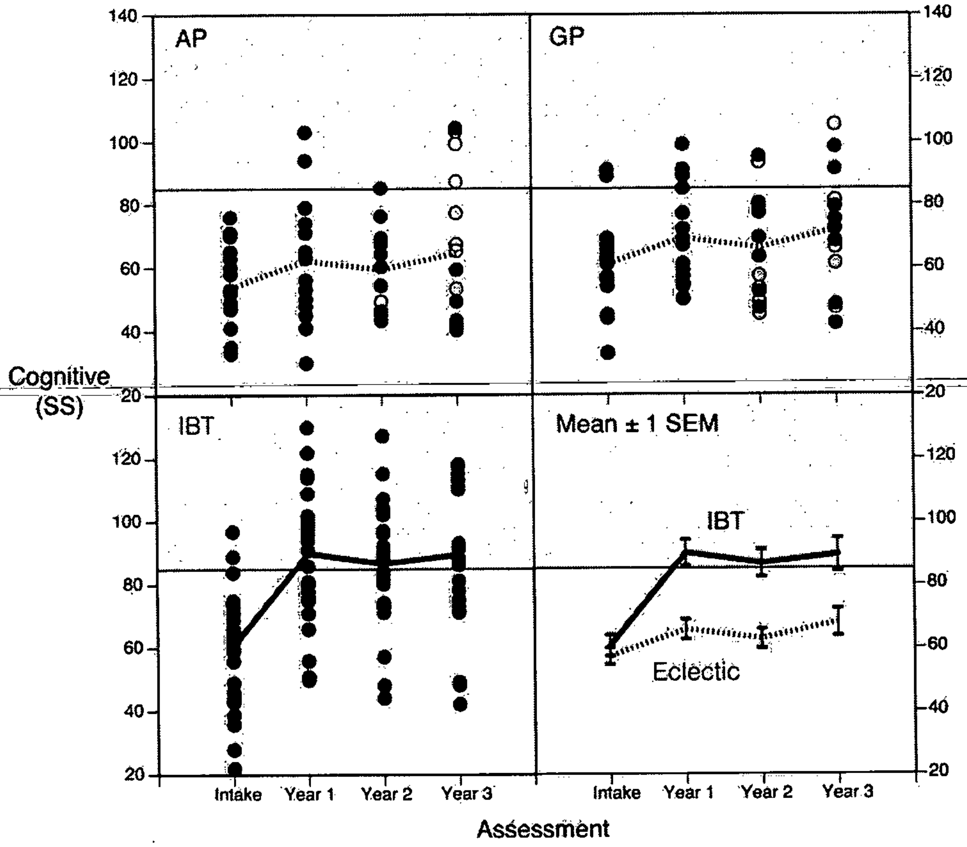

Fig. 2. Cognitive SSs at intake and 1–3 years later. Each dot represents the score for an individual child at that assessment time. Black dots indicate children who received their original treatment at the time of testing; white dots indicate children in the AP group who received GP treatment in the year preceding assessment, or children in the GP group who received AP treatment prior to assessment. Gray dots indicate children whose treatment prior to assessment was not recorded. Scores in the gray region of each panel are in the normal range (85 or higher). The lines in each panel connect the group mean scores at each assessment. The vertical bars in the lower right panel extend ±1 standard error around each group mean.

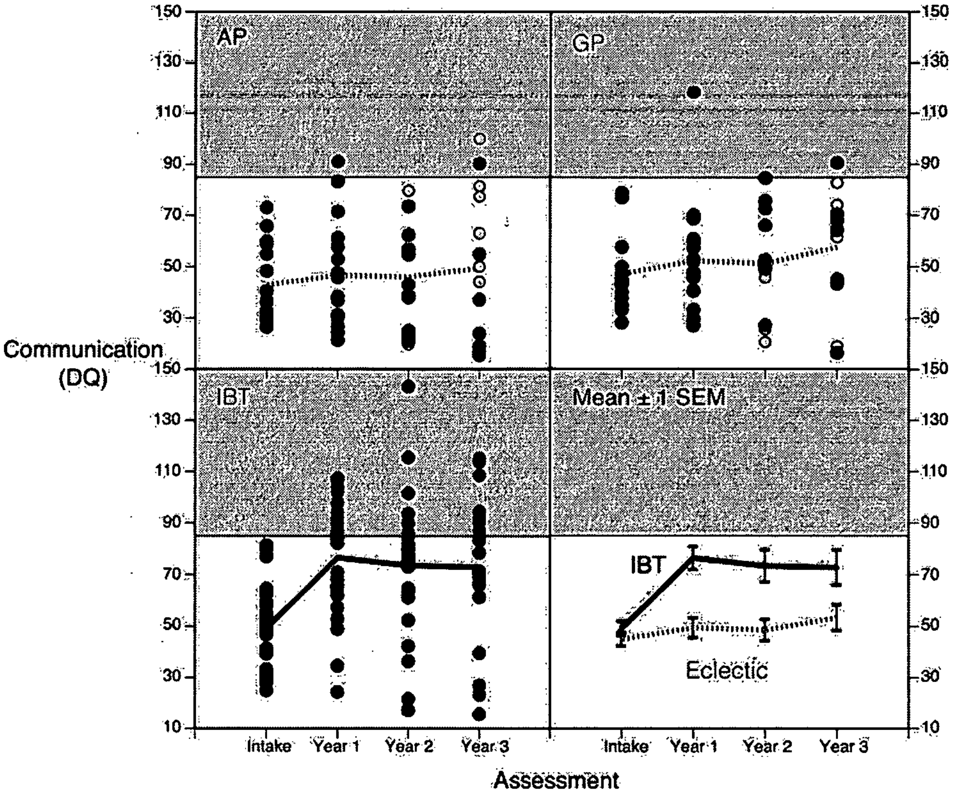

Fig. 3. Communication DQs at intake and 1–3 years later. See Fig. 2 caption for details.

After the first year of treatment, the sharply accelerated trajectory for the IBT group relative to the two other groups did not continue for any measure except the social skills DQ, which increased again from Year 1 to Year 2 before leveling off (Fig. 5). The mean cognitive skills SS for the IBT group remained stable from Year 1 to Year 3 (Fig. 2), while the mean communication skills DQ, self-help skills DQ, and adaptive skills composite SS declined slightly (Figs. 3, 4 and 6, respectively). The mean scores for the GP and AP groups either increased slightly or declined from Year 1 to Year 3 on all measures except social skills DQs, which increased for the GP group (Fig. 5).

In general, the gaps that emerged between the means of the IBT group and the other two groups after one year of treatment remained fairly constant or expanded in favor of the IBT group in Years 2 and 3 (see the lower right-hand panels of Figs. 2–6). Although the mean scores for the children in the IBT group were higher than those of the children in the eclectic treatment group three years after intake, those differences were not statistically significant (see Tables 2 and 3). With one possible exception, that was not because children in the IBT group regressed or because those in the AP and GP groups improved substantially; rather, it was because some children lacked 3-year followup assessments, reducing the Year 3 sample sizes and precluding the detection of statistically significant differences among the group mean scores. The exception was the mean motor skills DQ for the IBT group, which declined slightly from intake to Year 2/3 but remained in the normal range. The AP group's mean motor skills DQs also declined over the course of treatment; that decline was statistically significant and resulted in a Year 3 mean that was below normal (see Table 3).

Given the large improvements in the IBT group after one year of treatment, it may seem surprising that continued treatment did not produce further large gains on most measures; rather, most mean scores remained stable or declined slightly in Years 2 and 3. That finding should be interpreted with caution, however, and in relation to the results for the other two groups. For example, the mean cognitive skills SS for the IBT group was in the normal range after one year of treatment, so further large increases were unlikely. The mean adaptive skills composite SS for the IBT group fell slightly over the course of treatment, but the means for the two other groups fell even more. One plausible explanation for the apparent declines in the mean VABS composite scores is that the programming for these young children emphasized skills other than those assessed by the VABS.

3.3. *Analyses of outcomes by type of treatment*

Additional analyses were conducted to ascertain the proportions of children in each group who achieved clinically important outcomes by the end of treatment, and the likelihood that each type of treatment would produce such outcomes.

Pet. Reh. App.64

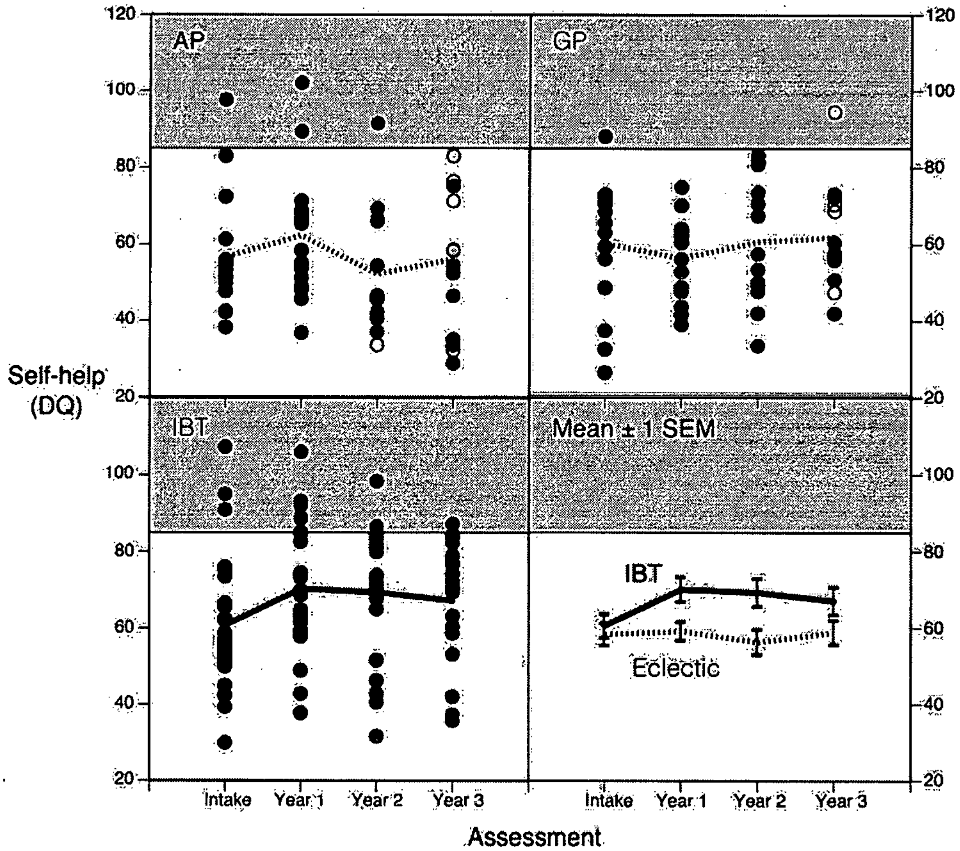

Fig. 4. Self-help DQs at intake and 1–3 years later. See Fig. 2 caption for details.

Table 4 shows the percentage of children in each group who had final (Year 2 or 3) scores in the normal range (i.e., ≥85; third column), final scores that were at least one standard deviation (≥15 points) higher than their intake scores (fifth column), and either of those favorable outcomes (penultimate column). Columns immediately to the right of each of those show odds ratios and probability ratios. To illustrate the odds ratio statistic, consider the data in the fourth column for the cognitive SS. For the IBT group, 61% had a final score ≥85 on that measure; the odds of achieving that favorable outcome were 0.607/(1 − 0.607) = 1.545. For the children in the AP and GP groups combined, 25% had final scores ≥85; the odds of this outcome were 0.250/(1 − 0.250) = 0.333. The ratio of those two odds is 1.545/0.333 = 4.64. This odds ratio of 4.64 is greater than the "neutral" value of 1, indicating that a favorable outcome on the cognitive SS was attained more often by children in the IBT group than by children in the two other groups combined. A likelihood ratio test, which is similar to a chi-square test, confirmed this difference as statistically significant. An odds ratio of 4.64, however, does not signify that children in the IBT group were 4.64 times more likely to have a favorable outcome than children in the AP and GP groups. Such an estimate is better provided by the probability ratio, which is shown in parentheses below each odds ratio in Table 4. The probability ratio for the cognitive SS example is 0.607 (the probability of a final score ≥85 for children in the IBT group) divided by 0.250 (the probability of a final score ≥85 for children in the AP and GP groups combined) = 2.43, indicating that children in the IBT group were 2.43 times more likely to achieve final cognitive SSs in the normal range than were children in the other two groups combined. Probability ratios are more readily interpreted than odds ratios, but statistical tests for group differences utilize odds ratios.

Table 4 shows that the overwhelming majority of the odds ratios and probability ratios favored IBT, indicating that clinically important outcomes as defined here were far more likely to be attained by children who received IBT than by children who received either of the other two treatments. The only exception was that final motor DQ scores were unlikely to be at least one standard deviation above the intake scores. As noted previously, that was likely due to a ceiling effect, in that the mean motor DQ for the IBT group was in the normal range at intake and stayed there over the course of treatment. Double asterisks in Table 4 show that the advantage for IBT children was more likely to be statistically significant when a favorable outcome was defined as a final score ≥85 than when it was defined as an increase of at least 15 points over intake.

Statistically significant differences between the AP and GP groups emerged only for an increase of 15 points or more over intake for social, motor, and adaptive skills composite scores. For those three measures, the odds of a favorable outcome were higher for the GP group than for the AP group. For the cognitive, receptive, and self-help measures, children in the AP group

Pet. Reh. App.65

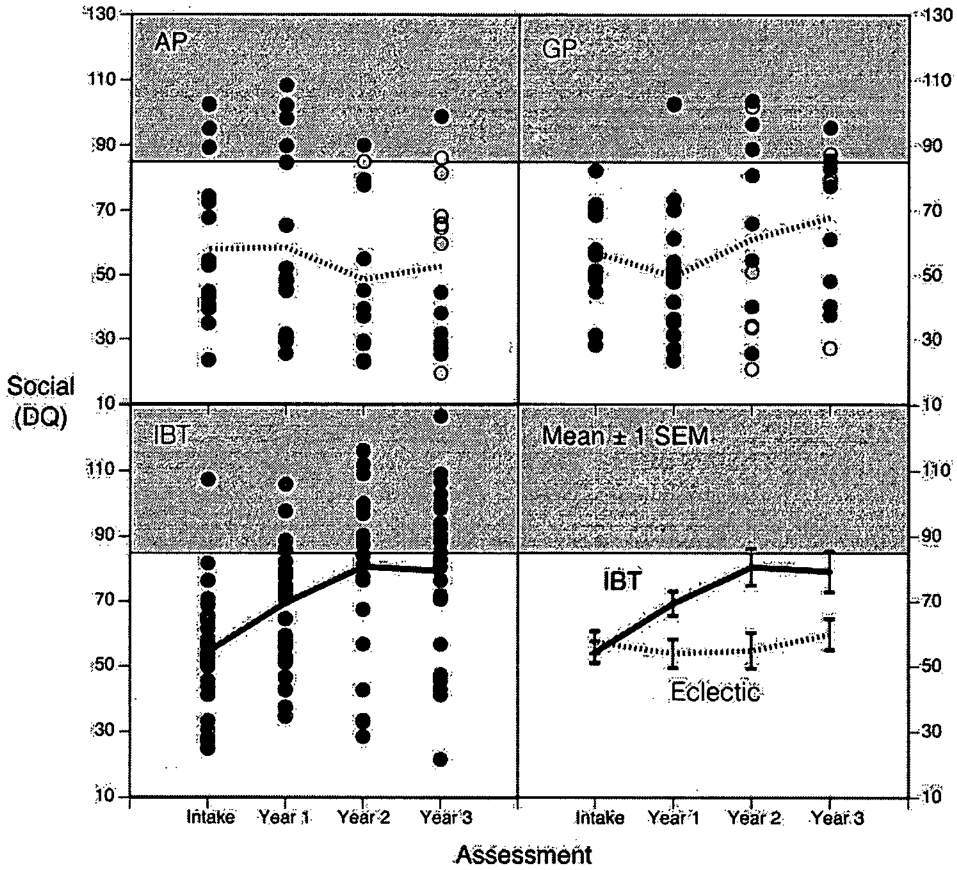

Fig. 5. Social DQs at intake and 1–3 years later. See Fig. 2 caption for details.

were more likely to have favorable outcomes than children in the GP group, though none of those differences was statistically significant. Collectively, these analyses suggest that neither of the comparison treatments was likely to result in favorable outcomes, and that combining the AP and GP groups did not mask any important group differences in outcomes.

Fig. 7 is a graphic representation of the percentages of children in the IBT group and the combined AP and GP groups who had scores in the normal range at each assessment. At intake, those percentages were comparably small for both groups on all measures except the motor skills DQ, on which fairly large proportions of both groups (57% IBT, 47% AP/GP combined) had scores in the normal range. By the end of treatment, a larger percentage of children in the IBT group than in the AP/GP group had scores in the normal range on all measures except the self-help DQ.

Individuals with final scores that were in the normal range (≥85) or at least one standard deviation above intake scores can be readily identified in Fig. 8. In this figure, each child's score on each measure is plotted as a function of his or her score at intake (on the x-axis) and the change from intake to the final assessment (on the y-axis; the final assessment was made at Year 2 if the child was not assessed at Year 3). Final scores in the normal range appear in the dark gray region of each panel, and scores representing increases of at least one standard deviation over intake are in the light gray regions. Both regions are populated by more children in the IBT group (closed circles) than by children in the other two groups (open symbols). That is, more of the children who received IBT had final outcomes that constituted clinically important changes over baseline than did children who received either of the other two treatments.

An important question is whether children in this study who attained normal levels of functioning at any point maintained those levels over the course of treatment. That question is difficult to answer, because only a portion of the children in each group had scores in the normal range at any assessment time, and not all children were assessed at both Year 2 and Year 3. Nevertheless, the question is sufficiently important to merit an attempt to answer it. For this analysis, children were classified into four categories of outcomes: (a) scored <85 one year and remained <85 the next year; (b) scored <85 one year but scored ≥85 the next year (i.e., transitioned to a normal range of functioning); (c) scored ≥85 one year but scored <85 the next year (i.e., regressed); and (d) scored ≥85 one year and remained ≥85 the following year. Those categories were then combined across measures to calculate the probability of each of the four outcomes for each year-to-year assessment transition. Separate calculations were made for children in the IBT group and for children in the combined AP and GP groups.

Results of these analyses are illustrated by the Sankey diagram shown in Fig. 9. In this figure, arrows are not just proportional in width to the quantities the represent; they are also horizontal if they represent children who maintained

Pet. Reh. App.66

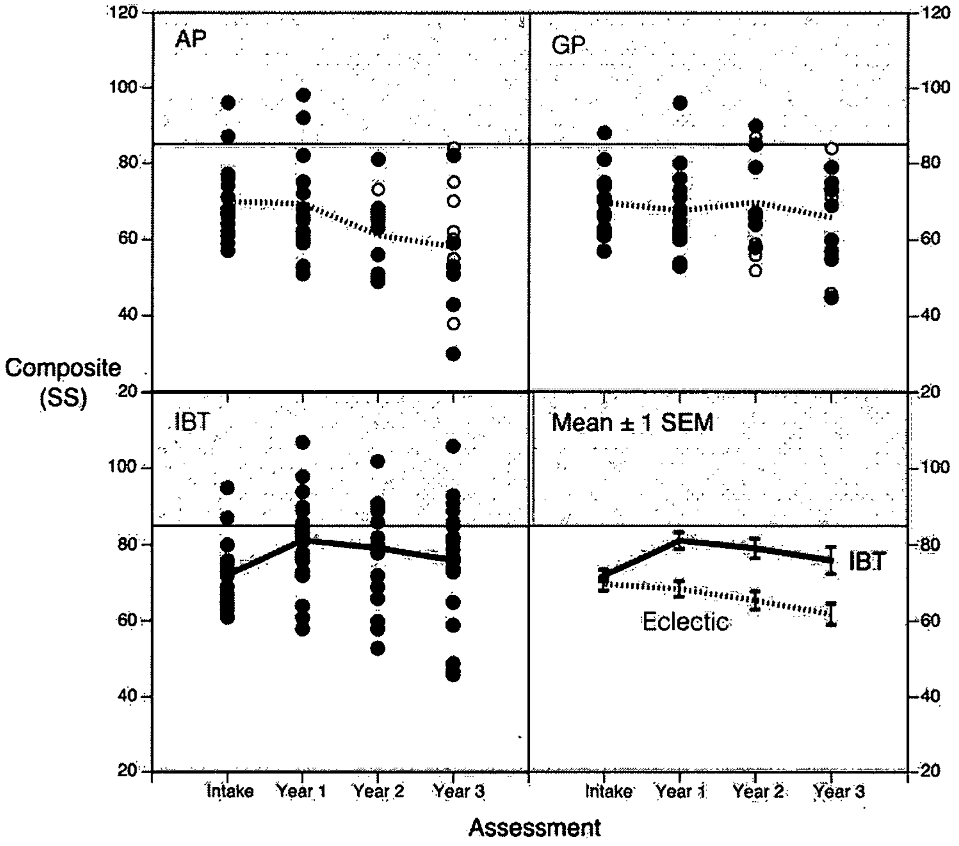

Fig. 6. Composite adaptive skills SSs at intake and 1–3 years later. See Fig. 2 caption for details.

assessed levels of functioning, they slant upward for children who improved, and they slant downward for children who regressed from one year to the next. The figure should be interpreted cautiously, because it represents data that were collapsed across measures and is based upon other suboptimal manipulations. Nevertheless, several intriguing trends are suggested. One is that most children who moved from below-normal to normal-range functioning did so after one year of treatment. For both groups, the probability of moving into the normal range was higher from intake to Year 1 than from Year 1 to Year 2, or from Year 2 to Year 3 (indicated by the upward-slanting arrows in Fig. 9). Stated differently, the prospect of achieving scores in the normal range diminished with each additional year of treatment, but the likelihood of scoring in the normal range was substantially and consistently higher for children in the IBT group than for children in the AP/GP groups combined at all three years post-intake (as shown by the percentages in the upward-slanting arrows). For children in the AP/GP group, if a score ≥85 was not attained after one year of treatment, the prospects for attaining a normal score were extremely dim.

A second general trend, confirming analyses presented in preceding tables and figures, is that children in the IBT group were far more likely to score in the normal range at all three post-intake assessments than were children in the two comparison groups. Further, percentages shown in the upward slanting arrows indicate that children in the IBT group were more than three times as likely as children in the AP and GP groups to have scores that moved them from the below-normal to the normal range at Years 1–3. That advantage was not limited to Year 1 scores; it remained relatively consistent throughout all three years of the study.

A final trend, illustrated by the downward slanting arrows in Fig. 9, is that regressions from normal to below-normal range scores were much more common for children in the AP/GP group than for children in the IBT group. In fact, children in the AP/GP group were 3.45 times as likely to regress as to advance during the first year of treatment, 4.45 times more likely to regress than advance during the second year of treatment, and 4.91 times more likely to regress than advance in the third year of treatment. The opposite pattern was seen for children in the IBT group, where advancements were 2.48 times as likely as regressions during the first year of treatment. Advancements and regressions occurred about equally often between Year 1 and Year 2 for the IBT group (the ratio was 1.08 in favor of advancements), but in the third year of treatment advancements were 1.75 times as frequent as regressions. Collectively, these findings suggest that children who received IBT were much more likely to attain and maintain normal levels of functioning than were children who received either of the other treatments.

Table 4
Percent of children with favorable outcomes, and odds ratios and probability ratios for each measure.

Measure	Group	Final score ≥ 85	Odds ratio (probability ratio)	Final score ≥ 15 points above intake	Odds ratio (probability ratio)	Either desirable outcome	Odds ratio (probability ratio)
Cognitive (SS)	IBT	61% (n = 28)	4.64^{**} (2.43)	81% (n = 27)	8.00^{**} (2.30)	82% (n = 28)	8.78^{**} (2.39)
	AP/GP combined	25% (n = 32)		35% (n = 31)		34% (n = 32)	
	AP	25% (n = 16)	1.00 (1.00)	38% (n = 16)	1.20 (1.13)	38% (n = 16)	1.32 (1.20)
	GP	25% (n = 16)		33% (n = 15)		31% (n = 16)	
Non-verbal (DQ)	IBT	85% (n = 27)	8.40^{**} (2.10)	60% (n = 20)	3.00 (1.80)	85% (n = 27)	6.52^{**} (1.82)
	AP/GP combined	41% (n = 32)		33% (n = 27)		47% (n = 32)	
	AP	31% (n = 16)	0.45 (0.63)	31% (n = 16)	0.80 (0.86)	44% (n = 16)	0.78 (0.88)
	GP	50% (n = 16)		36% (n = 11)		50% (n = 16)	
Receptive (DQ)	IBT	26% (n = 27)	5.08^{*} (4.02)	78% (n = 27)	8.17^{**} (2.59)	85% (n = 27)	10.45^{**} (2.40)
	AP/GP combined	6% (n = 31)		30% (n = 30)		35% (n = 31)	
	AP	6% (n = 16)	0.93 (0.94)	31% (n = 16)	1.14 (1.09)	38% (n = 16)	1.20 (1.13)
	GP	7% (n = 15)		29% (n = 14)		33% (n = 15)	
Expressive (DQ)	IBT	46% (n = 28)	5.85^{**} (3.60)	82% (n = 28)	9.20^{**} (2.46)	82% (n = 28)	9.66^{**} (2.55)
	AP/GP combined	13% (n = 31)		33% (n = 30)		32% (n = 31)	
	AP	13% (n = 16)	0.93 (0.94)	31% (n = 16)	0.82 (0.88)	31% (n = 16)	0.91 (0.94)
	GP	13% (n = 15)		36% (n = 14)		33% (n = 15)	
Commun-ication (DQ)	IBT	36% (n = 28)	3.89^{*} (2.86)	68% (n = 28)	2.25 (1.40)	75% (n = 28)	3.00^{*} (1.50)
	AP/GP combined	13% (n = 32)		48% (n = 31)		50% (n = 32)	
	AP	13% (n = 16)	1.00 (1.00)	38% (n = 16)	0.40 (0.63)	38% (n = 16)	0.36 (0.60)
	GP	13% (n = 16)		60% (n = 15)		63% (n = 16)	
Self-help (DQ)	IBT	11% (n = 28)	3.72 (3.43)	39% (n = 28)	1.36 (1.22)	43% (n = 28)	1.65 (1.37)
	AP/GP combined	3% (n = 32)		32% (n = 31)		31% (n = 32)	
	AP	0% (n = 16)	0.00 (0.00)	38% (n = 16)	1.65 (1.41)	38% (n = 16)	1.80 (1.50)
	GP	6% (n = 16)		27% (n = 15)		25% (n = 16)	
Social (DQ)	IBT	54% (n = 28)	4.12^{*} (2.45)	67% (n = 27)	2.77 (1.59)	71% (n = 28)	3.65^{*} (1.76)
	AP/GP combined	22% (n = 32)		42% (n = 31)		41% (n = 32)	
	AP	13% (n = 16)	0.31 (0.40)	25% (n = 16)	0.22^{*} (0.42)	25% (n = 16)	0.26 (0.44)
	GP	31% (n = 16)		60% (n = 15)		56% (n = 16)	
Motor (DQ)	IBT	57% (n = 28)	1.51 (1.22)	19% (n = 27)	0.91 (0.93)	57% (n = 28)	1.51 (1.22)
	AP/GP combined	47% (n = 32)		20% (n = 30)		47% (n = 32)	
	AP	31% (n = 16)	0.27 (0.50)	0% (n = 16)	0.00^{**} (0.00)	31% (n = 16)	0.27 (0.50)
	GP	63% (n = 16)		43% (n = 14)		63% (n = 16)	
Composite (SS)	IBT	36% (n = 28)	8.33^{**} (5.71)	16% (n = 25)	1.65 (1.55)	36% (n = 28)	5.37^{*} (3.81)
	AP/GP combined	6% (n = 32)		10% (n = 29)		9% (n = 32)	
	AP	0% (n = 16)	0.00 (0.00)	0% (n = 16)	0.00^{*} (0.00)	0% (n = 16)	0.00^{*} (0.00)
	GP	13% (n = 16)		23% (n = 13)		19% (n = 16)	

* Odds ratio differs significantly from 1 ($p < 0.05$).
** Odds ratio differs significantly from 1 ($p < 0.01$).

Pet. Reh. App.67

4. Discussion

4.1. Differential treatment outcomes

Our 2005 study evaluated outcomes for 61 children with autism who received just over one year of either IBT or one of two eclectic interventions. Although the three groups were similar at intake, children who received IBT had significantly higher mean scores after one year of treatment than those who received eclectic interventions. The present study extended

Pet. Reh. App.68

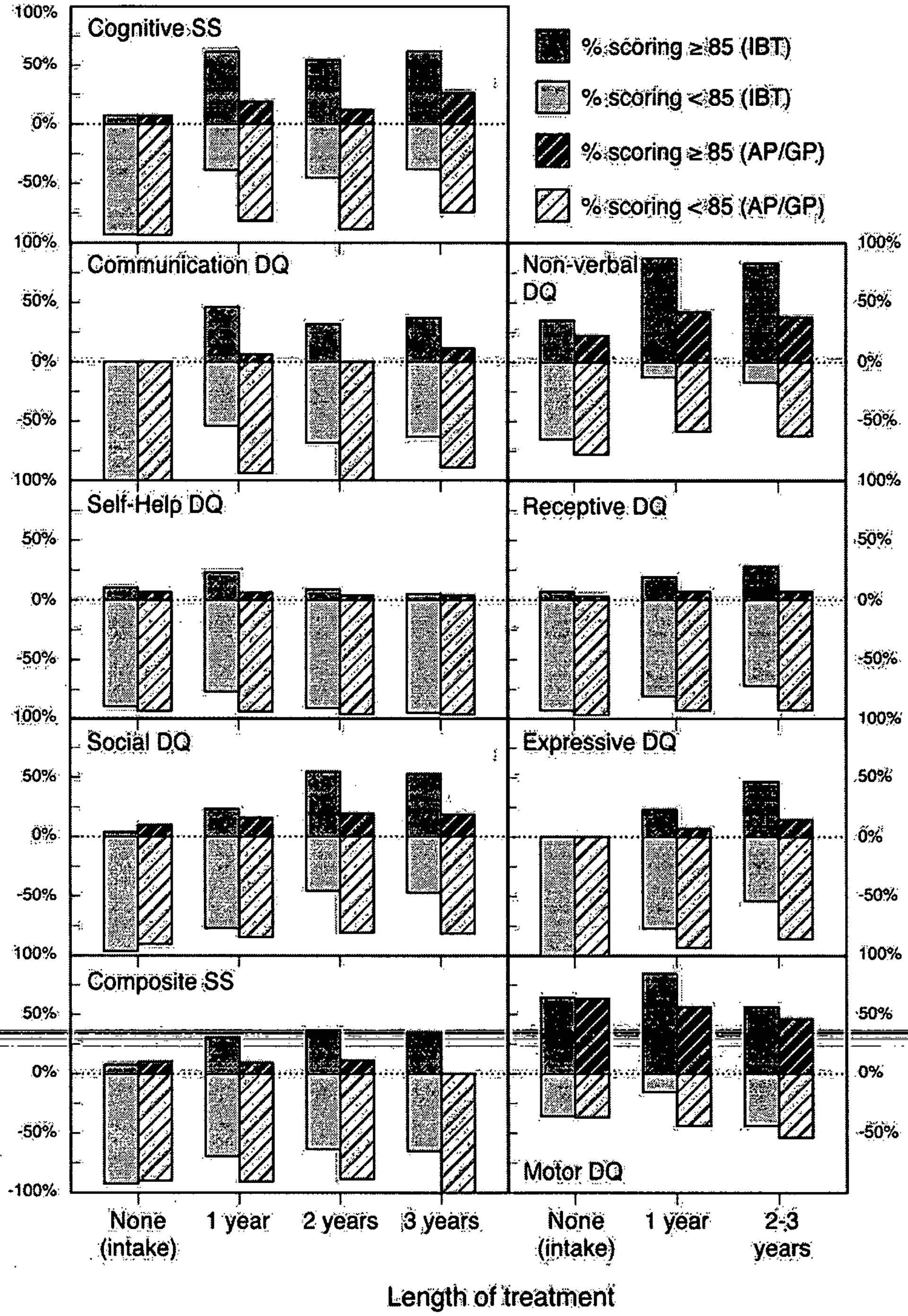

Fig. 7. Percent of children in each treatment group with a score in the normal range (SS or DQ ≥85) at intake and 1–3 years after intake.

those findings by showing that the largest gains generally occurred in the first year of treatment and in IBT children only, and that the advantage experienced by IBT children after one year of treatment was maintained throughout the second and third years of treatment. Indeed, three years after treatment began, mean scores on standardized assessments of cognitive, language, adaptive, and motor skills were higher for children in the IBT group than they were for children in the eclectic intervention groups.

At their final assessment, 61% of the children who received IBT tested within the average range of cognitive functioning, compared with only 25% of the children who received eclectic treatment. That is, children in the IBT group were more than twice as likely to attain a cognitive skills score in the normal range as children in the two eclectic intervention groups. Final

Pet. Reh. App.69

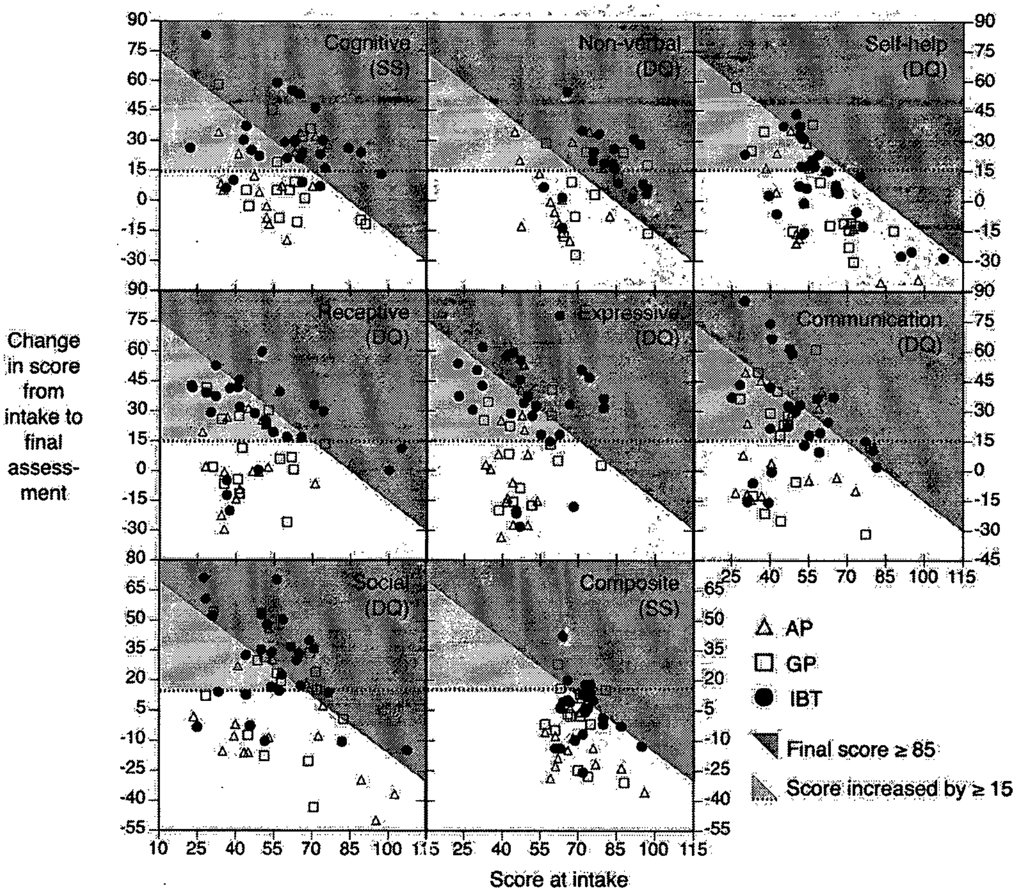

Fig. 8. Scores for individual children on each measure, plotted as a function of the value at intake along the x-axis, and the change from intake to Year 3 (or Year 2 if the child was not assessed at Year 3) along the y-axis. Scores of children in the IBT group are shown as solid circles, scores of children in the AP group are represented by open triangles, and scores of children in the GP group are shown as open squares. Final scores in the normal range (≥85) appear in the dark gray region of each panel, and final scores <85 but at least 15 points higher than at intake (i.e., above the dotted line in each panel) appear in the light gray region of each panel.

assessment scores on other measures showed similar patterns. Compared to children who received eclectic interventions, children who received IBT were twice as likely to score in the normal range on the final assessment of nonverbal skills, approximately three times as likely to score in the normal range on the final assessments of communication and adaptive skills, approximately four times as likely to score within the normal range on the final assessments of receptive and expressive communication skills, and almost six times more likely to have a final adaptive behavior skills composite score within the normal range.

As they were at Year 1, average outcomes at Years 2 and 3 were worse for children in the AP and GP groups than for children in the IBT group, while average outcomes for the two eclectic intervention groups did not differ significantly from each other. The mean score for the GP group was higher than the mean score for the AP group on some measures in some years, but there were no statistically reliable differences between outcomes produced by the two eclectic treatments. Additionally, both eclectic treatments performed substantially worse than IBT in producing standardized test scores in the normal range of functioning, and neither eclectic treatment was more likely than the other to produce a favorable outcome. The results for the AP intervention might be surprising to some readers because that intervention was intensive and designed specifically for children with autism. Despite these features, no child from the AP group scored in the normal range on the final assessment of adaptive functioning. In contrast, more than one-third of the children in the IBT group achieved a normal-range score on the final assessment of adaptive skills. These findings are especially important given the critical contribution of adaptive skills to independent functioning throughout the lifespan.

Although scores in the normal range are certainly desirable outcomes, so are other clinically significant improvements. Changes in test scores that do not reach the normal range may nonetheless reflect the acquisition of many skills that enhance independent functioning, which in turn produces economic savings due to reduced need for specialized services (Jacobson, Mulick, & Green, 1998; Motiwala, Gupta, Lilly, Ungar, & Coyte, 2006). About one-third of the children in this study who received AP or GP interventions had final scores on tests of cognitive or adaptive skills that were at least 15 points higher than

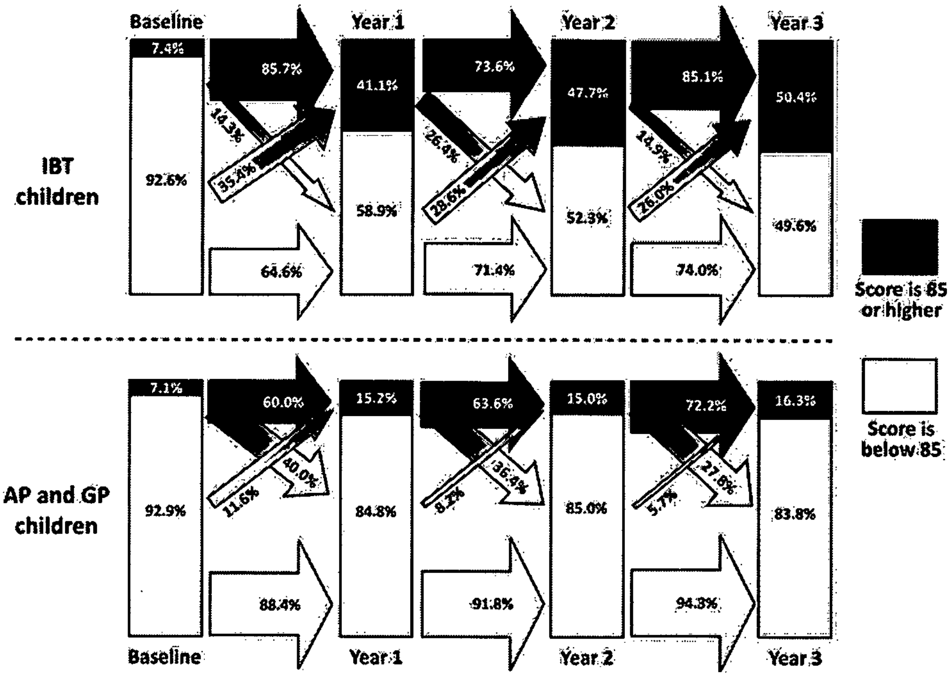

Fig. 9. Percentages of children who scored in the normal range (gray region) or below 85 (white region) at each assessment time, and who transitioned from one of those ranges to another on successive assessments. Horizontal arrows indicate maintenance of scores in the normal range (gray arrows) or in the below-normal range (white arrows). Upward-slanting arrows indicate changes from below-normal to normal-range scores, and downward-slanting arrows represent changes from normal to below-normal range scores.

their intake scores, suggesting that those interventions may produce some benefit for some children with autism. Children in the IBT group, however, were more than twice as likely as children in the other two groups to show changes of that magnitude over the course of treatment. Differences on most other measures were somewhat smaller but equally clear and in the same direction. Motor skills scores were an exception, as they were somewhat more likely to increase by at least 15 points among children in the AP and GP groups than among children in the IBT group. However, that difference was not statistically significant.

The multiple regression approach we used for most of our statistical analyses accommodated individual differences (e.g., in parental education and age at diagnosis), but of course those analyses focused on group data. Group comparisons are appropriate for determining which of two or more treatments is generally most effective; however, we urge caution in relying exclusively on group statistics to prognosticate about individual children. It is clear from the individual data presented here (Figs. 2–6) that not all children within each treatment group responded similarly to that treatment. Research correlating child characteristics with differential outcomes might help identify categories of children who are more or less likely to respond well to a given treatment on average, but more precise information about the effects of treatments on individuals with varying characteristics could be gleaned from studies using single-case research designs, perhaps in combination with elements of between-groups designs (Green, 2008; Guyatt et al., 2008; Larson, 1990; Morgan & Morgan, 2001; Powers et al., 2006). Research methods that focus on changes in individual behavior with treatment could also enable analyses of the differential effectiveness of elements of multicomponent treatments like IBT (e.g., Heyvaert, Maes, Van den Noortgate, Kuppens, & Onghena, 2012) as well as treatment targets that function as behavioral cusps to bring the individual's behavior into contact with new contingencies of reinforcement, thereby producing even more widespread behavior change (Rosales-Ruiz & Baer, 1997).

4.2. Changes over the course of treatment

In this study, the changes that occurred during the first year of treatment were generally maintained throughout the second and third years for children in all three groups. Group mean scores in Years 2 and 3 tended to remain within ±5 points of the corresponding group means at the end of Year 1, with the large differences in favor of IBT after one year largely persisting throughout Years 2 and 3. Other studies comparing IBT with eclectic treatment over similar time periods have produced similar findings (Cohen et al., 2006; Eikeseth et al., 2007). One difference is that the IBT advantage was larger after a mean of 31.4 months of treatment than after one year of treatment in the study by Eikeseth et al. (2007). That may be related to the fact that the children studied by Eikeseth et al. were older when they started treatment than the children in our study and the study by Cohen et al.

Pet. Reh. App.71

(2006), but it might also reflect differences in other child characteristics or the treatment packages (e.g., variations in targets, priorities, procedures, etc.).

Measures of some skill domains in our study deviated from the trends just described. For instance, the mean motor and self-help scores for the IBT group were higher than those for either eclectic intervention group at the end of Year 1, but the differences between the final group means on those measures were not statistically significant. That was at least partly due to reduced sample sizes at Year 3. It should also be reiterated that motor skills were not delayed substantially for any of the groups at intake, and motor and self-help skills were not among the highest priority treatment targets for many of the children who received IBT.

The fact that most of the largest improvements in the IBT group occurred after one year of treatment might lead some to conclude that there is little benefit in extending treatment beyond the first year. Such a conclusion might be warranted if there were compelling evidence to support predictions that improvements would persist if treatment were to end after one year. Our study cannot speak to that hypothesis, because none of the children in the IBT group received just one year of treatment. Nor are we aware of other studies that have tested that hypothesis directly. One group of researchers did, however, evaluate the performances of 23 young children with autism two years after they had completed a 2-year course of IBT (Kovshoff, Hastings, & Remington, 2011). They found that a subgroup of 9 children who had statistically significant increases on tests of cognitive and adaptive skills during treatment maintained those gains after two years with no treatment, but the scores of the other 14 children decreased significantly. Analyses showed that the first subgroup had higher baseline scores and received more intensive treatment than did the second subgroup. Although limited, those findings corroborate our clinical observations that terminating IBT prematurely can be detrimental to many children with autism.

Ending IBT after one year might also be justified if it were reasonably certain that extending treatment would be unlikely to produce further clinically significant gains. Again, we have found no compelling evidence to support that prediction. On the contrary, some children in our IBT group made marked improvements in Years 2 and 3 (e.g., see the upward-pointing arrows in Fig. 9). Other researchers have also documented meaningful improvements occurring in the second, third, and fourth year of IBT (e.g., Cohen et al., 2006; Eikeseth et al., 2007; Sallows & Graupner, 2005). We speculate that given the pervasive and substantial skill deficits exhibited by many young children with autism, one and even two years of IBT is not likely to produce gains that will persist over long periods of time without specialized intervention. The first 1–2 years of IBT are typically focused on building many basic, foundational skills. Further intensive treatment seems essential for solidifying those repertoires and for building the more complex social, language, and academic skills required to function successfully in regular school and community settings.

4.3. Limitations

Participants in this study were not randomly assigned to groups; instead, treatment assignments primarily reflected parental preferences and education team decisions. In Howard et al. (2005), however, we demonstrated empirically that the three groups were functionally equivalent at intake. The only statistically significant group differences were in parental education (parents of children in the IBT group averaged one year more of education than parents of children in the AP and GP groups) and age at diagnosis (children in the IBT group were diagnosed an average of 5 months earlier than children in the GP group, who in turn were diagnosed an average of 4 months earlier than children in the AP group). Both variables were controlled for statistically in subsequent data analyses, though control was rarely necessary because individual scores almost never covaried with parental education or age at diagnosis.

Another limitation is that some children switched between the AP and GP treatments during Years 2 and 3. We have no information about the reasons for those shifts, but it would be unusual for an education team to recommend moving a child out of an effective program and for the child's family to approve such a change. Therefore, we speculate that the changes may speak to the lack of efficacy of either eclectic approach. The data showed that neither eclectic treatment reliably produced meaningful benefits, and when children switched from one eclectic treatment to the other, there was rarely any improvement with the new treatment. These findings imply that the two eclectic treatments were essentially indistinguishable in their efficacy, and that our analyses and conclusions were not compromised by the fact that some children switched from one eclectic treatment to the other.

The impact of mortality on our findings should be considered. Virtually all children were assessed in all domains at intake and Year 1, but participation rates were lower in subsequent years. The reduced sample sizes forced us to combine data from Years 2 and 3 to analyze outcomes for the nonverbal intelligence, receptive language, expressive language, and motor skills measures. That precluded mapping developmental trajectories for those domains as precisely as we did for other domains. It is important to note, however, that mortality does not seem to have biased the overall findings. In fact, imputation analyses suggest that the group differences we observed were not artifacts of mortality; if anything, the advantage of IBT over the eclectic treatments would likely have been greater if more comprehensive assessment data were available for Years 2 and 3.

The primary limitation of our study may be that there were no measures of the integrity with which any of the treatments was delivered, as we reported in Howard et al. (2005). Additionally, each treatment comprised a number of components, and it was not feasible to parse out the contributions of individual components to the outcomes. Nonetheless, our findings converge with those of other studies in which IBT and a comparison eclectic treatment program had similar elements, intensity, and duration (e.g., Cohen et al., 2006; Eikeseth et al., 2007). They add to the growing body of evidence that IBT

Pet. Reh. App.72

produces significantly larger increases on standardized measures of cognitive and adaptive functioning than other treatments. Although those measures do not capture all repertoires that may be influenced by intervention, they are considered more objective than indices like classroom placement, and correlate positively with other measures of overall and long-term functioning. Thus, there is general consensus among autism researchers that protocols for evaluating treatment effects must include certain standardized instruments (e.g., Eldevik et al., 2009, 2010; Fein et al., 2013; Martin, Bibby, Mudford, & Eikseth, 2003; Mundy, 1993; Wolery & Garfinkle, 2002), Collectively, this study and others that used such protocols clearly indicate that IBT is an effective, evidence-based treatment for young children diagnosed with autism.

Acknowledgements

This study was supported in part by Valley Mountain Regional Center and California State University, Stanislaus. We also wish to express our appreciation to Jessica Bailey for her assistance.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ridd.2014.08.021.

References

Bondy, A. S., & Frost, L A. (1994). *The picture exchange communication system: Training manual.* Cherry Hill, NJ: Pyramid.

Cohen, H., Amerine-Dickens, M., & Smith, T. (2006). Early intensive behavioral treatment: Replication of the UCLA model in a community setting. *Developmental and Behavioral Pediatrics, 27,* S145-S155.

Delmolino, L M. (2006). Brief report: Use of DQ for estimating cognitive ability in young children with autism. *Journal of Autism and Developmental Disorders, 36,* 959-963.

Eikeseth, S. (2009). Outcome of comprehensive psycho-educational interventions for young children with autism. *Research in Developmental Disabilities, 30,* 158-178.

Eikeseth, S., Smith, T., Jahr, E., & Eldevik, S. (2002). Intensive behavioral treatment at school for 4- to 7-year-old children with autism: A 1-year comparison controlled study. *Behavior Modification, 2002,* 49-68.

Eikeseth, S., Smith, T., Jahr, E., & Eldevik, S. (2007). Outcome for children with autism who began intensive behavioral treatment between ages 4 and 7: A comparison controlled study. *Behavior Modification, 31,* 264-278.

Eldevik, S., Eikeseth, S., Jahr, E., & Smith, T. (2006). Effects of low-intensity behavioral treatment for children with autism and mental retardation. *Journal of Autism and Developmental Disorders, 36,* 211-224.

Eldevik, S., Hastings, R. P., Hughes, J. C., Jahr, E., Eikeseth, S., & Cross, S. (2009). Meta-analysis of early intensive behavioral intervention for children with autism. *Journal of Clinical Child & Adolescent Psychology, 38,* 439-450.

Eldevik, S., Hastings, R. P., Hughes, J. C., Jahr, E., Eikeseth, S., & Cross, S. (2010). Using participant data to extend the evidence base for intensive behavioral intervention for children with autism. *American Journal on Intellectual and Developmental Disabilities, 115,* 381-405.

Eldevik, S., Hastings, R. P., Jahr, E., & Hughes, J. C. (2012). Outcomes of behavioral intervention for children with autism in mainstream pre-school settings. *Journal of Autism and Developmental Disorders, 42,* 210-220.

Fein, D., Barton, M., Eigsti, I., Kelley, E., Naigles, L., Schultz, R. T., & Tyson, K. (2013). Optimal outcome in individuals with a history of autism. *Journal of Child Psychology and Psychiatry, 54,* 195-205.

Fischer, J. L., Howard, J. S., Sparkman, C. R., & Moore, A. G. (2009). Establishing generalized syntactical responding in young children with autism. *Research in Autism Spectrum Disorders, 4,* 76-88.

Green, G. (2008). Single-case research methods for evaluating treatments for ASD. In S. C. Luce, D. S. Mandell, C. Mazefsky, & W. Seibert (Eds.), *Autism in Pennsylvania: A symposium issue of the Speaker's Journal of Pennsylvania Policy* (pp. 119-132). Harrisburg, PA: Legislative Office for Research Liaison, Pennsylvania House of Representatives.

Green, G. (2011). Early intensive behavior analytic intervention for autism spectrum disorders. In E. Mayville & J. Mulick (Eds.), *Behavioral foundations of effective autism treatment* (pp. 183-199). Cornwall-on-Hudson, NY: Sloan Publishing.

Green, G., Brennan, L. C., & Fein, D. (2002). Intensive behavioral treatment for a toddler at high risk for autism. *Behavior Modification, 26,* 69-102.

Guyatt, G., Rennie, D., Meade, M., & Cook, D. J. (2008). *Users' guides to the medical literature: A manual for evidence-based clinical practice* (2nd ed.). New York: McGraw-Hill Professional.

Heyvaert, M., Maes, B., Van den Noortgate, W., Kuppens, S., & Onghena, P. (2012). A multilevel meta-analysis of single-case and small-n research on interventions for reducing challenging behavior in persons with intellectual disabilities. *Research in Developmental Disabilities, 33,* 766-780.

Howard, J. S., Sparkman, C. R., Cohen, H. G., Green, G., & Stanislaw, H. (2005). A comparison of intensive behavior analytic and eclectic treatments for young children with autism. *Research in Developmental Disabilities, 26,* 359-383.

Jacobson, J. W., Mulick, J. A., & Green, G. (1998). Cost-benefit estimates for early intensive behavioral intervention for young children with autism - General model and single state case. *Behavioral Interventions, 13,* 201-226.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59,* 12-19.

Kovshoff, H., Hastings, R., & Remington, B. (2011). Two-year outcomes for children with autism after the cessation of early intensive behavioral intervention. *Behavior Modification, 35,* 427-450.

Larson, E. B. (1990). N-of-1 clinical trials: A technique for improving medical therapeutics. *Western Journal of Medicine, 152,* 52-56.

Lord, C., & Schopler, E. (1989). The role of age at assessment, developmental level, and test in the stability of intelligence scores in young autistic children. *Journal of Autism and Developmental Disorders, 19,* 483-499.

Lovaas, O. I. (1987). Behavioral treatment and normal educational and intellectual functioning in young autistic children. *Journal of Consulting and Clinical Psychology, 55,* 3-9.

Martin, N., Bibby, P., Mudford, O. C., & Eikseth, S. (2003). Toward the use of a standardized assessment for young children with autism: Current assessment practices in the UK. *Autism, 7,* 321-330.

Morgan, D. L., & Morgan, R. K. (2001). Single-participant research design: Bringing science to managed care. *American Psychologist, 56,* 119-127.

Motiwala, S. S., Gupta, S., Lilly, M. B., Ungar, W. J., & Coyte, P. C. (2006). The cost-effectiveness of expanding intensive behavioural intervention to all autistic children in Ontario. *Healthcare Policy, 1*(2), 135-151.

Mundy, P. (1993). Normal versus high-functioning status in children with autism. *American Journal on Mental Retardation, 97,* 381-384.

National Autism Center (2009). *National standards report.* Randolph, MA: National Autism Center.

Pet. Reh. App.73

Powers, S. C., Piazza-Waggoner, C., Jones, J. S., Ferguson, K. S., Daines, C., & Acton, J. D. (2006). Examining clinical trial results with single-subject analysis: An example involving behavioral and nutrition treatment for young children with cystic fibrosis. *Journal of Pediatric Psychology, 31*, 574–581.

Reichow, B., & Wolery, M. (2009). Comprehensive synthesis of early intensive behavioral interventions for young children with autism based on the UCLA Young Autism Project Model. *Journal of Autism and Developmental Disorders, 39*, 23–41.

Remington, B., Hastings, R. P., Kovshoff, H., Espinosa, F., Jahr, E., Brown, T., & Ward, N. (2007). Early intensive behavioral intervention: Outcomes for children with autism and their parents after two years. *American Journal on Mental Retardation, 112*, 418–438.

Rogers, S. J., & Vismara, L. A. (2008). Evidence-based comprehensive treatments for early autism. *Journal of Clinical Child and Adolescent Psychology, 37*, 8–38.

Rosales-Ruiz, J., & Baer, D. M. (1997). Behavioral cusps: A developmental and pragmatic construct for behavioral analysis. *Journal of Applied Behavior Analysis, 30*, 533–544.

Sallows, G. O., & Graupner, T. D. (2005). Intensive behavioral treatment for children with autism: Four-year outcome and predictors. *American Journal on Mental Retardation, 110*, 417–438.

Schmidt, M. (2008). The Sankey diagram in energy and material flow management Part I: History. *Journal of Industrial Ecology, 12*, 82–94.

Smith, T., Groen, A. D., & Wynne, J. W. (2000). Randomized trial of intensive early intervention for children with pervasive developmental disorder. *American Journal on Mental Retardation, 105*. 269-285.

Wolery, M., & Garfinkle, A. N. (2002). Measures in intervention research with young children who have autism. *Journal of Autism and Developmental Disorders, 32*, 463–478.

Zachor, D. A., Ben-Itzchak, E., Rabinovich, A., & Lahat, E. (2007). Change in autism core symptoms with intervention. *Research in Autism Spectrum Disorders, 1*, 304–307.

Human Mutation

OFFICIAL JOURNAL

HGVS
HUMAN GENOME
VARIATION SOCIETY
www.hgvs.org

Performance of Mutation Pathogenicity Prediction Methods on Missense Variants

Janita Thusberg,[1,2] Ayodeji Olatubosun,[1] and Mauno Vihinen[1,3]*

[1]*Institute of Biomedical Technology, FI-33014 University of Tampere, Finland;* [2]*Buck Institute for Age Research, Novato, California;* [3]*Research Center, Tampere University Hospital, Tampere, Finland*

Communicated by Christophe Béroud

Received 20 May 2010; accepted revised manuscript 7 December 2010.
Published online 22 February 2011 in Wiley Online Library (www.wiley.com/humanmutation). DOI 10.1002/humu.21445

ABSTRACT: Single nucleotide polymorphisms (SNPs) are the most common form of genetic variation in humans. The number of SNPs identified in the human genome is growing rapidly, but attaining experimental knowledge about the possible disease association of variants is laborious and time-consuming. Several computational methods have been developed for the classification of SNPs according to their predicted pathogenicity. In this study, we have evaluated the performance of nine widely used pathogenicity prediction methods available on the Internet. The evaluated methods were MutPred, nsSNPAnalyzer, Panther, PhD-SNP, PolyPhen, PolyPhen2, SIFT, SNAP, and SNPs&GO. The methods were tested with a set of over 40,000 pathogenic and neutral variants. We also assessed whether the type of original or substituting amino acid residue, the structural class of the protein, or the structural environment of the amino acid substitution, had an effect on the prediction performance. The performances of the programs ranged from poor (MCC 0.19) to reasonably good (MCC 0.65), and the results from the programs correlated poorly. The overall best performing methods in this study were SNPs&GO and MutPred, with accuracies reaching 0.82 and 0.81, respectively.
Hum Mutat 32:358–368, 2011. © 2011 Wiley-Liss, Inc.

KEY WORDS: method evaluation; bioinformatics; pathogenicity prediction; SNPs

Introduction

Most human genetic variation is represented by single nucleotide polymorphisms (SNPs), and many of them are believed to cause phenotypic differences between individuals. Owing to the application of high-throughput sequencing methods, the number of identified variants in the human genome is growing rapidly, but identifying those variations responsible for specific phenotypes is a laborious task. The ability to discriminate between pathogenic and benign variants computationally could significantly aid targeting disease-causing mutations by helping in the selection and prioritization of likely candidates from a pool of data. A subset of SNPs occur at protein coding regions in the genome, and from a medical point of view particularly interesting ones are the nonsynonymous SNPs (nsSNPs) that lead to an amino acid substitution at the protein level (referred here to as missense variants). nsSNPs may affect gene function through their effect on the structure and/or function of the encoded protein.

Prediction of the possible disease-association of missense variants is a difficult problem because an amino acid substitution can affect the biological function of a gene product in a number of ways [Thusberg and Vihinen, 2009]. An amino acid substitution may disrupt sites that are critical in protein function, such as catalytic residues or ligand-binding pockets. A missense mutation may as well lead to alterations in the structure, folding, or stability of the protein product, thereby altering or preventing the function of the protein. On the other hand, amino acid substitutions do not necessarily affect protein function. Effects of missense mutations are often the most difficult to predict while the consequences of most deletions, insertions, and nonsense mutations are rather self-evident.

Many methods have been developed for the computational prediction of the phenotypic effect of nsSNPs. Some of them are for the study of very specific mechanisms, whereas others are developed to predict whether a variation is harmful or benign. All of the variation tolerance methods evaluated in this study follow a similar procedure in which a missense variant is first labeled with properties related to the damage it may cause to the protein structure or function. The resulting feature vector is then utilised to decide whether the variant is pathogenic or not. The methods differ in the properties of the variant they take into account in the prediction, as well as in the nature and possible training of the classification method used for decision making. The nine widely used methods evaluated in this study are based on evolutionary information (Panther [Thomas et al., 2003], PhD-SNP SVM-Profile [Capriotti et al., 2006], and SIFT [Ng and Henikoff, 2001]), or a combination of protein structural and/or functional parameters and multiple sequence alignment derived information (MutPred [Li et al., 2009], nsSNPAnalyzer [Bao et al., 2005], PolyPhen [Ramensky et al., 2002], PolyPhen2 [Adzhubei et al., 2010], SNAP [Bromberg and Rost, 2007], and SNPs&GO [Calabrese et al., 2009]). The machine-learning methods utilize neural networks (NN) (SNAP), random forests (RF) (MutPred, nsSNPAnalyzer), or support vector machines (SVMs) (PhD-SNP, SNPs&GO) for classification, whereas the other methods classify variants according to empirically derived rules (PolyPhen), Bayesian methods (PolyPhen2), or mathematical operations (SIFT, Panther) (Table 1).

*Correspondence to: Mauno Vihinen, Institute of Biomedical Technology, FI-33014 University of Tampere, Finland. E-mail: mauno.vihinen@uta.fi

Contract grant sponsors: The Tampere Graduate School in Biomedicine and Biotechnology; The Sigrid Jusélius Foundation; The Academy of Finland; The Medical Research Fund of Tampere University Hospital.

Pet. Reh. App.75

Table 1. Summary of the Evaluated Methods

Method	Based on	Training set	Conservation analysis	Structural attributes	Annotations	Website
MutPred	RF	HGMD, Swiss-Prot	SIFT, Pfam, PSI-BLAST	Predicted attributes	–	http://mutpred.mutdb.org/
nsSNPAnalyzer	RF	Swiss-Prot	SIFT	Homologue mapping	–	http://snpanalyzer.uthsc.edu/
Panther	Alignment scores	–	Panther library, HMMs	–	–	http://www.pantherdb.org/tools/csnpScoreForm.jsp
PhD-SNP	SVM	Swiss-Prot	Sequence environment, sequence profiles	–	–	http://gpcr2.biocomp.unibo.it/cgi/predictors/PhD-SNP/PhD-SNP.cgi
PolyPhen	Empirical rules	–	PSIC profiles	Homologue mapping/predictions	Swiss-Prot	http://genetics.bwh.harvard.edu/pph/
PolyPhen2	Bayesian classification	Swiss-Prot, neutral pseudo-mutations	PSIC profiles	Homologue mapping/predictions	Pfam domain	http://genetics.bwh.harvard.edu/pph2/
SIFT	Alignment scores	–	MSAs	–	–	http://sift.jcvi.org/
SNAP	NN	PMD, neutral pseudo-mutations	PSIC profiles, Pfam, PSI-BLAST	Predictions	–	http://rostlab.org/services/snap/
SNPs&GO	SVM	Swiss-Prot	Sequence environment, sequence profiles, Panther	–	GO	http://snps-and-go.biocomp.unibo.it/snps-and-go/

GO, Gene Ontology; HGMD, Human Gene Mutation Database; HMM, Hidden Markov model; NN, neural network; MSA, multiple sequence alignment; PMD, Protein Mutant Database; PSIC, position-specific independent counts; RF, random forest; SVM, support vector machine.

As mutation data and information about the genotypes of individuals accumulate, understanding the molecular level effects of variations and elucidating their possible disease association is an important research challenge [Karchin, 2009; Mooney, 2005; Ng and Henikoff, 2006; Steward et al., 2003; Thusberg and Vihinen, 2009]. Numerous locus-specific databases (LSDBs) have been established for the collection, analysis, and distribution of disease-related variation information in certain genes. Data for several genes is available, for example, in the protein knowledgebase SwissProt [Yip et al., 2004] and PhenCode [Giardine et al., 2007], which is a database that connects human variant data with phenotypic information from LSDBs with genomic data from the ENCODE project and other resources in the UCSC Genome Browser [Raney et al., 2011]. SNP information is available in dbSNP [Sherry et al., 2001], a genetic variation database. Several tools for the prediction of the phenotypic consequences of missense variants are available, but without knowledge about the quality of predictions, choosing the best method and evaluating the reliability of its outcome is impossible. We therefore performed the first comprehensive systematic evaluation of nine bioinformatics tools predicting the phenotypic effects of missense variants.

Materials and Methods

Datasets

We built a positive dataset (referred to as pathogenic dataset) of 19,335 missense mutations from the PhenCode database [Giardine et al., 2007] (downloaded in June 2009), registries in IDbases [Piirilä et al., 2006] and from 18 individual LSDBs, and a negative (neutral) dataset of 21,170 human nonsynonymous coding SNPs with an allele frequency >0.01 and chromosome sample count >49 from the dbSNP database [Sherry et al., 2001] build 131. The SNP data was filtered so that none of the dbSNP entries included in our dataset contained OMIM links to minimize the number of disease-associated SNPs in the neutral dataset. Entries annotated as "putative" or "predicted" were also left out. In addition, the neutral dataset was searched against the pathogenic dataset in order to remove possible duplicates and further minimise the probability of having false negative cases in the set. The PhenCode data was filtered so that only SNPs annotated as disease causing in the SwissProt database were taken into our pathogenic dataset. Swiss-Prot provides high-quality hand-curated information about the possible disease-relation of nsSNPs, derived from literature [Yip et al., 2008]. The complementing LSDB data was retrieved manually from each database. The pathogenic and neutral datasets contained 1,190 and 9,011 proteins, respectively, of which 445 and 1,205 were found to have three-dimensional structure coordinates in the Protein Data Bank (PDB) [Berman et al., 2000]. The datasets are available for download at our Website (http://bioinf.uta.fi).

Both datasets were run by all of the nine methods studied here. The number of results from nsSNPAnalyzer is much smaller than the original number of cases in the input data, because the program only accepts mutations in those sequences for which a homologous protein is found in the ASTRAL database [Chandonia et al., 2004]. A large number of proteins in our dataset did not match with any entry in the database, thus limiting the number of cases that could be analysed by nsSNPAnalyzer.

Two kinds of subdatasets were constructed from the original pathogenic and neutral datasets. First, a structural subdataset was compiled from the part of both datasets for which structural data was available in the PDB, to study the effect of available structure data on prediction performance. Second, for probing the effect of using Swiss-Prot-derived data as part of the pathogenic testing set, we constructed a subdataset containing only pathogenic variants not present in Swiss-Prot. The corresponding neutral dataset was compiled by randomly selecting an equal number of variants from the original neutral test set.

To test whether the differences in method performance with these subdatasets was caused by smaller testing set size, we constructed 100 sample datasets each containing 1,000 pathogenic and 1,000 neutral variants randomly picked from the original datasets, and compared the average MCCs obtained with the MCCs from the subdatasets.

The Pathogenic-or-not Pipeline (PON-P) [Thusberg and Vihinen, 2009] was used for the submission of sequences and variants into the analysis programs nsSNPAnalyzer, Panther, PhD-SNP, PolyPhen, PolyPhen2, SIFT, and SNAP. PON-P is a service that simultaneously submits the input data provided by the user to selected prediction methods. MutPred and SNPs&GO were run locally at the corresponding laboratories by the developers of the methods.

Prediction Methods

The effects of mutations and SNPs were predicted by the programs MutPred [Li et al., 2009], nsSNPAnalyzer [Bao et al.,

Pet. Reh. App.76

2005], Panther [Thomas et al., 2003], PhD-SNP [Capriotti et al., 2006], PolyPhen [Ramensky et al., 2002], PolyPhen2 [Adzhubei et al., 2010], SIFT [Ng and Henikoff, 2001], SNAP [Bromberg and Rost, 2007], and SNPs&GO [Calabrese et al., 2009]. Key properties of the methods are listed in Table 1. The default parameters of all programs were applied, and only the protein sequence and missense variant were given as input information for each program, as in a normal user situation of unknown variant analysis.

MutPred

MutPred is a Random Forest-based classification method that utilizes several attributes related to protein structure, function, and evolution. MutPred utilizes the SIFT method [Ng and Henikoff, 2003] for defining the evolutionary attributes, along with PSI-BLAST, transition frequencies [Bromberg and Rost, 2007], and Pfam profiles [Finn et al., 2010]. In MutPred, structural descriptors include prediction of secondary structure and solvent accessibility by the method PHD [Rost, 1996], transmembrane helix prediction by TMHMM [Krogh et al., 2001], coiled-coil structure prediction by MARCOIL [Delorenzi and Speed, 2002], stability prediction by I-Mutant 2.0 [Capriotti et al., 2005], B-factor prediction [Radivojac et al., 2004], and disorder prediction by DisProt [Peng et al., 2006]. Function-related attributes include predictions of DNA-binding residues [Ahmad et al., 2004], catalytic residues, calmodulin-binding targets [Radivojac et al., 2006], and posttranslational modification sites [Daily et al., 2005; Iakoucheva et al., 2004; Radivojac et al., 2010]. The MutPred method estimates effects of an amino acid substitution on the set of defined properties of a protein and based on those estimates, predicts whether an amino acid substitution is likely to have phenotypic effects.

nsSNPAnalyzer

nsSNPAnalyzer is a machine-learning method that integrates multiple sequence alignment (MSA) and protein structure analysis to classify missense variants. The input protein sequence is searched against the ASTRAL database [Chandonia et al., 2004] for homologous protein structures, and extracts features of the environment of the substitution from the obtained structure, namely, the solvent accessibility, environmental polarity, and secondary structure. The SIFT method [Ng and Henikoff, 2003] is used for calculating the normalised probability of the substitution in the MSA, and the similarity and dissimilarity between the mutated, that is, original, and mutant residue is also taken into account. The program then uses a Random Forest classifier trained by a dataset prepared from the Swiss-Prot database [Yip et al., 2004] to classify the variant to be disease-associated or functionally neutral.

Panther

The Panther Evolutionary Analysis of Coding SNPs (referred simply to as Panther in this article) calculates substitution position-specific evolutionary conservation (subPSEC) scores based on alignments of evolutionarily related proteins to predict the pathogenicity. The alignments are obtained from the PANTHER library of protein families based on Hidden Markov Models (HMMs). The subPSEC score describes the amino acid probabilities, in particular, positions among evolutionarily related sequences, and the values range from 0 (neutral) to about −10 (most likely to be deleterious). The cutoff for classifying a missense variant to be pathogenic can be defined by the user, but the authors of the method advice to use a cutoff of −3 for classification [Thomas et al., 2003].

PhD-SNP

PhD-SNP is a prediction method based on single-sequence and sequence profile based support vector machines trained on Swiss-Prot variants [Yip et al., 2004]. The single-sequence SVM (SVM-Sequence) classifies the missense variant to be pathogenic or neutral based on the nature of the substitution and properties of the neighboring sequence environment. The profile-based SVM (SVM-Profile) utilizes sequence profile information taken from MSAs, and classifies the variant according to the ratio between the frequencies of the wild-type and substituted residue. A decision tree algorithm chooses which one of the two SVMs described above is to be used at each case based on the occurrence of wild-type and mutant amino acids at the given position.

PolyPhen

PolyPhen (Polymorphism Phenotyping) uses a rule-based cutoff system to classify variants. It initially characterises the input missense variant by various sequence, structure, and phylogeny based descriptors. The sequence-based characterisation includes SWALL database [Johnson and Todd, 2000] annotations for sequence features, a transmembrane predictor TMHMM [Krogh et al., 2001] and PHAT [Ng et al., 2000] transmembrane-specific matrix score for substitutions at predicted transmembrane regions, the Coils2 program [Lupas et al., 1991] for prediction of coiled coil regions, and the SignalP [Nielsen et al., 1997] program to predict signal peptide regions. Phylogenetic information is derived by constructing a profile matrix from aligned sequences by the PSIC (Position-Specific Independent Counts) software [Sunyaev et al., 1999]. The structural descriptors are obtained by mapping the missense variant onto the corresponding or similar protein and then using the DSSP program [Kabsch and Sander, 1983] for secondary structure information, solvent-accessible surface, and φ–ψ dihedral angles. In addition, PolyPhen calculates the normalized accessible surface area and changes in accessible surface propensity resulting from the amino acid substitution, change in residue side chain volume, region of the Ramachandran map, normalized B factor, and loss of a hydrogen bond according to the Hbplus program [McDonald and Thornton, 1994]. The SWALL database annotations are utilized in the structure analysis such that the program checks whether the substitution site is in spatial contact with critical residues annotated to be involved in forming binding sites or active sites. Additionally, the contacts of the substituted residue with ligands or subunits of the protein molecule are checked. After characterising the variant, PolyPhen applies empirically derived rules based on the characteristics to predict whether a missense variant is damaging or benign.

PolyPhen2

PolyPhen2 utilizes a combination of sequence- and structure-based attributes for the description of an amino acid substitution, and the effect of mutation is predicted by a naive Bayesian classifier. The sequence-based features include PSIC scores and MSA properties, and position of mutation in relation to domain boundaries as defined by Pfam [Finn et al., 2010]. The structure-derived features

Pet. Reh. App.77

are solvent accessibility, changes in solvent accessibility for buried residues, and crystallographic B-factor.

SIFT

SIFT (Sorting Intolerant From Tolerant) makes inferences from sequence similarity using mathematical operations. SIFT constructs an MSA and considers the position of the missense variant and the type of the amino acid change. Based on the amino acids appearing at each position in the MSA, SIFT calculates the probability that a missense variant is tolerated conditional on the most frequent amino acid being tolerated.

SNAP

SNAP (Screening for Nonacceptable Polymorphisms) is a neural network-based tool for the prediction of the effect of a missense variant. The method utilises evolutionary information from PSI-BLAST [Altschul et al., 1997] frequency profiles and PSIC [Sunyaev et al., 1999], transition frequencies for mutations, biophysical characteristics of the substitution, secondary structural information, and relative solvent accessibility values predicted by PROFsec/PROFacc [Rost, 1996; Rost and Sander, 1994], chain flexibility predicted by PROFbval [Schlessinger et al., 2006], protein family evolutionary information, and information about domain boundaries from Pfam [Finn et al., 2010], and Swiss-Prot annotations [Bairoch and Apweiler, 2000] to classify a missense variant. The training sets for the NN were constructed from Protein Mutant Database (PMD) [Kawabata et al., 1999] data complemented by a set of neutral pseudomutations generated by the authors of the method as described in Bromberg and Rost [2007].

SNPs&GO

SNPs&GO is an SVM classifier based on mutation type and sequence environment information, sequence profiles taken from MSAs, predictions from the program Panther [Thomas et al., 2003], and a function-based log-odds score describing information about protein function defined by Gene Ontology (GO) terms [Ashburner et al., 2000].

From the output of the programs, we only took the binary prediction (pathogenic/neutral) into consideration without taking into account any confidence values provided by some of the programs. Panther provides a numerical output rather than a binary classification (subPSEC score), which we converted to a binary prediction using a cutoff point of −3 as recommended in [Thomas et al., 2003]. PolyPhen and PolyPhen2 classify the effects of a missense variant into three categories: "Probably pathogenic," "Possibly pathogenic," and "Benign." We converted these into binary classifications in two ways, first by considering only the "Probably pathogenic" class as pathogenic and the "Possibly pathogenic" and "Benign" classes as neutral, and second, by considering both the "Probably pathogenic" and "Possibly pathogenic" classes as pathogenic, and the "Benign" class as neutral. These two ways of classifying the variants are referred to as PolyPhen(2)a and PolyPhen(2)b in this study, respectively.

Determination of Secondary Structural Elements and Accessible Surface Areas

The 3D structure coordinates of proteins were obtained from the PDB. Secondary structural information and accessible surface area (ASA) values for each mutation site were assigned by the program STRIDE [Frishman and Argos, 1995]. We classified residues with ASAs $\leq 10\%$ as buried and with ASAs $\geq 25\%$ as exposed, similarly as in a previous study [Khan and Vihinen, 2010].

Determination of Structural Classes of Proteins

The CATH database version 3.3 [Orengo et al., 1997] was used to group studied proteins according to their secondary and tertiary structure types.

Statistical Analyses

The quality of the predictions is described by six parameters: accuracy, precision, sensitivity, specificity, negative predictive value (NPV) and Matthews correlation coefficient (MCC). In the following equations, *tp*, *tn*, *fp*, and *fn* refer to the number of true positives, true negatives, false positives and false negatives, respectively.

$$\text{Accuracy} = \frac{tp+tn}{tp+tn+fp+fn}$$

$$\text{Precision} = \frac{tp}{tp+fp}$$

$$\text{Specificity} = \frac{tn}{fp+tn}$$

$$\text{Sensitivity} = \frac{tp}{tp+fn}$$

$$\text{NPV} = \frac{tn}{tn+fn}$$

$$\text{MCC} = \frac{tp \times tn - fn \times fp}{\sqrt{(tp+fn)(tp+fp)(tn+fn)(tn+fp)}}$$

The MCC [Matthews, 1975] is a very important evaluation statistic as it is unaffected by the differing proportion of neutral and pathogenic datasets predicted by the different programs. Because of its insensitivity to differing test set sizes, it gives a more balanced assessment of performance than the other performance measures [Baldi et al., 2000].

To be able to correlate the quality parameters for different programs with different sizes of test sets containing different amounts of pathogenic and neutral cases, the numbers of neutral cases were normalized to be equal to the number of pathogenic cases for each program.

Substitution statistics for both the pathogenic and neutral datasets were analyzed by comparing the frequencies of the substitutions with the expected values that were calculated using the distribution of all amino acids in the datasets. For the original residues, the expected values were calculated with regard to their codon diversity thereby taking into account all possible amino acid substitutions. The chi-square test was used to determine the significance of the results and chi-square was calculated as:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

where f_o is the observed frequency and f_e is the expected frequency for an amino acid. The p-values were estimated in a one-tailed fashion.

Correlations between the program outputs were calculated by counting all of the common cases and those predicted correctly, and using Spearman's rank correlation coefficient.

Results

Test Set Features

The distributions of mutated and mutant amino acids in both pathogenic and neutral datasets are biased (Table 2), and only a few residues occur as expected on the grounds of codon diversity. In the pathogenic dataset (mutation data), A, C, G, M, R, W, and Y are overrepresented among the original (mutated) amino acid residues, whereas E, F, I, K, L, N, Q, S, T, and V are significantly underrepresented. These results are in line with previous observations for distributions of disease-causing mutations in protein secondary structural elements [Khan and Vihinen, 2007], except for the overrepresentation of A and Y, and underrepresentation of L, N, S, and V in our data. In the neutral dataset, the distributions of many amino acids differ from the distributions in the pathogenic set. Most importantly, cysteines are highly underrepresented among the substituted positions, as opposed to their frequent mutation in the pathogenic dataset. This might be due to the important role of cysteines in folding of many proteins as they are capable of forming disulphide bonds, and therefore the substitution of cysteines in proteins transported through endoplasmic reticulum by any other residue can rarely be neutral in terms of protein structure and function. Other differences between the datasets are the underrepresentation of mutated glycine, tryptophan, and tyrosine residues in the neutral set as opposed to their frequent mutation in the pathogenic set, and the overrepresentation of isoleucine, asparagine, threonine, and valine residues in the neutral variation data, contrasting their underrepresentation in the mutation data.

The distributions of mutant or substituting amino acids are also very biased in both pathogenic and neutral datasets, and the amino acid residues I, P, R, T, V, and Y have opposite distributions in the mutation and neutral sets. Interestingly, proline residues are highly overrepresented among the substituting residues in the mutation dataset, and underrepresented in the negative set.

Table 2. Amino Acid Distributions in the Pathogenic (Mutations) and Neutral (SNPs) Datasets

	Wild-type residues/pathogenic variants					Wild-type residues/neutral variants			
	Observed	Expected	χ^2	*P*-value		Observed	Expected	χ^2	*P*-value
A	1224	252.5	**3737.28*****	0.000	A	1852	1449.4	**111.82*****	3.91E-26
C	943	468.1	**481.79*****	8.71E-107	C	424	473.9	*5.24**	0.022
D	950	988.7	1.52	0.218	D	991	1017.8	0.70	0.401
E	994	1449.8	*143.32****	5.02E-33	E	1273	1530.4	*43.31****	4.68E-11
F	537	766.1	*68.53****	1.25E-16	F	458	766.0	*123.83****	9.16E-29
G	2087	1355.0	**395.42*****	5.46E-88	G	1182	1374.1	*26.85****	2.20E-07
H	554	528.8	1.20	0.273	H	530	555.0	1.12	0.289
I	642	911.4	*79.64****	4.49E-19	I	996	924.5	**5.53***	0.019
K	497	1173.9	*390.28****	7.20E-87	K	774	1223.0	*164.85****	9.87E-38
L	1497	2068.4	*157.84****	3.35E-36	L	1270	2113.0	*336.34****	4.00E-75
M	520	435.5	**16.39*****	5.16E-05	M	642	442.2	**90.32*****	2.03E-21
N	605	754.4	29.59***	5.35E-08	N	894	777.0	**17.61*****	2.71E-05
P	1192	1252.8	2.95	0.086	P	1277	1323.3	1.62	0.203
Q	454	970.0	*274.52****	1.17E-61	Q	875	1028.1	*22.79****	1.81E-06
R	2797	1136.4	**2426.45*****	0.000	R	2376	1168.5	**1247.88*****	2.40E-273
S	1135	1681.4	*177.55****	1.66E-40	S	1648	1793.0	*11.72***	0.001
T	802	1087.9	*75.12****	4.42E-18	T	1482	1145.7	**98.72*****	2.91E-23
V	919	1246.3	*85.93****	1.86E-20	V	1682	1263.7	**138.46*****	5.78E-32
W	376	254.4	**58.17*****	2.41E-14	W	167	251.8	*28.54****	9.17E-08
Y	610	553.1	**5.85***	0.016	Y	377	549.8	*54.31****	1.71E-13
All	19335	19335			All	21170	21170		
	Mutant residues/pathogenic variants					Mutant residues/neutral variants			
A	622	1267.9	*329.01****	1.58E-73	A	1061	1388.20	*77.12****	1.61E-18
C	1233	563.5	**795.45*****	5.26E-175	C	722	617.0	**17.88*****	2.36E-05
D	900	633.9	**111.67*****	4.22E-26	D	666	694.1	1.14	0.286
E	719	563.5	**42.91*****	5.72E-11	E	825	617.0	**70.14*****	5.53E-17
F	623	633.9	0.19	0.664	F	855	694.1	**37.30*****	1.01E-09
G	922	1232.7	*78.29****	8.90E-19	G	1376	1349.6	0.52	0.473
H	918	633.9	**127.29*****	1.61E-29	H	967	694.1	**107.30*****	3.83E-25
I	619	950.9	*115.85****	5.14E-27	I	1139	1041.1	**9.20****	0.002
K	834	563.5	**129.85*****	4.41E-30	K	1171	617.0	**497.49*****	3.34E-110
L	1225	1796.1	*181.62****	2.15E-41	L	1390	1966.6	*169.06****	1.19E-38
M	534	317.0	**148.61*****	3.50E-34	M	828	347.0	**666.52*****	5.72E-147
N	662	633.9	1.24	0.265	N	845	694.1	**32.81*****	1.02E-08
P	1609	1267.9	**91.78*****	9.67E-22	P	1176	1388.2	*32.44****	1.23E-08
Q	808	563.5	**106.09*****	7.05E-25	Q	1056	617.0	**312.40*****	6.56E-70
R	2084	1831.4	**34.85*****	3.56E-09	R	1431	2005.2	*164.41****	1.23E-37
S	1502	1796.1	*48.17****	3.91E-12	S	1691	1966.6	*38.63****	5.13E-10
T	1012	1267.9	*51.64****	6.68E-13	T	1517	1388.2	**11.95****	0.001
V	1195	1267.9	*4.19**	0.041	V	1589	1388.2	**29.05*****	7.07E-08
W	638	246.5	**621.62*****	3.32E-137	W	471	269.9	**149.78*****	1.93E-34
Y	676	493.1	**67.88*****	1.74E-16	Y	394	539.9	*39.41****	3.44E-10
All	19335	19335			All	21170	21170		

The chi-square values in italics identify residues that are underrepresented and the values in bold identify overrepresented residues in comparison to random distributions derived theoretical codon usage frequencies. Significance levels are $^{*}P<0.05$; $^{**}P<0.01$; $^{***}P<0.001$.

Proline is a known secondary structure breaker [Chou and Fasman; 1974] and therefore mutations to P are often pathogenic.

Performance of Prediction Methods

To evaluate the performance of the programs predicting the pathogenicity of missense variants, we used six measures: accuracy, precision (or positive predictive value, PPV), specificity, sensitivity, NPV, and MCC. The values for these measures are presented in Table 3 for all the missense variants. SNPs&GO performed best in terms of accuracy (0.82), precision (0.90), specificity (0.92), and MCC (0.65), but sensitivity was higher in six other methods, and MutPred, Panther, PolyPhen2b, and SNAP performed better in terms of NPV. nsSNPAnalyzer performed worst in terms of MCC (0.19), accuracy (0.60), NPV (0.60), and precision (0.59). The two versions of PolyPhen have very similar overall performance; however, PolyPhen2 is recommended because the quality measures are more balanced.. The version classifying "Probably pathonegenic," PolyPhen2a, as harmful is somewhat better than the other option.

In Table 3, the results are presented for the subset of cases for which structural information could be assigned. The performance of all methods was generally worse except for sensitivity, which is better for all methods. SNPs&GO performed best also in the structural subcategory considering accuracy, precision, specificity, and MCC, and MutPred was the best method in terms of sensitivity and NPV.

To test whether the poor performance was due to the smaller dataset size we sampled the full dataset results for those cases for which structural data was not available. We then compared the average MCC values of the samples to those obtained for the full dataset. The 100 sample datasets each contained randomly picked 1,000 neutral and 1,000 pathogenic variations. The average MCCs of the sample datasets were comparable to the MCCs of the full dataset in the case of Panther (average sample MCC 0.53), PhD-SNP (0.43), PolyPhen2b (0.39), and SNAP (0.47). For the other methods the MCC values were rather close when comparing the full dataset to the subdataset. We conclude that the large differences in the MCCs of the programs between the full dataset and the set for which structures were available (Table 3) were not due to the differences in the sizes of these datasets but were caused by some other factors, that is, differences in the performance of the methods when predicting on different types of data.

We also performed the analyses for a dataset that consisted only of LSDB-derived mutations not found in SwissProt (Table 3). This was done as some methods have been trained with Swiss-Prot disease-causing mutations. Because all methods (except SNPs&GO),

Table 3. Performance of Prediction Methods

	MutPred	nsSNPAnalyzer	Panther	PhD-SNP	PolyPhen1a	PolyPhen 1b	PolyPhen 2a	PolyPhen 2b	SIFT	SNAP	SNPs&GO
Performance of prediction methods (full data)											
tp[a]	13829	4360	9689	11900	10093	14285	13807	16206	10464	16000	13736
fn[a]	2507	2778	2859	6896	9185	4993	5102	2703	4856	2146	5487
tn[a]	15891	1319	8676	16788	17669	13671	13863	10199	12188	8190	17028
fp[a]	4557	943	2797	4377	3199	7197	6010	9674	7433	6387	1382
cases +[a]	16336	7138	12548	18796	19278	19278	18909	18909	15320	18146	19223
cases −[a]	20448	2262	11473	21165	20868	20868	19873	19873	19621	14577	18410
Accuracy[b]	0.81	0.60	0.76	0.71	0.69	0.70	0.71	0.69	0.65	0.72	0.82
Precision[b]	0.79	0.59	0.76	0.75	0.77	0.68	0.71	0.64	0.64	0.67	0.90
Specificity[b]	0.78	0.58	0.76	0.79	0.85	0.66	0.70	0.51	0.62	0.56	0.92
Sensitivity[b]	0.85	0.61	0.77	0.63	0.52	0.74	0.73	0.86	0.68	0.88	0.71
NPV[b]	0.84	0.60	0.77	0.68	0.64	0.72	0.72	0.78	0.66	0.83	0.76
MCC[b]	0.63	0.19	0.53	0.43	0.39	0.40	0.43	0.39	0.30	0.47	0.65
Performance of prediction methods (3D structure)											
tp[a]	5625	2857	3934	5041	4563	5980	5814	6726	4303	6751	5887
fn[a]	517	1603	1009	2411	3074	1657	1842	930	1329	714	1746
tn[a]	1101	569	735	1090	1361	1070	1163	843	904	700	1378
fp[a]	697	527	441	754	462	753	672	992	901	777	318
cases +[a]	6142	4460	4943	7452	7637	7637	7656	7656	5632	7465	7633
cases −[a]	1798	1096	1176	1844	1823	1823	1835	1835	1805	1477	1696
Accuracy[b]	0.76	0.58	0.71	0.63	0.67	0.68	0.70	0.67	0.63	0.69	0.79
Precision[b]	0.70	0.57	0.68	0.62	0.70	0.65	0.67	0.62	0.60	0.63	0.80
Specificity[b]	0.61	0.52	0.63	0.59	0.75	0.59	0.63	0.46	0.50	0.47	0.81
Sensitivity[b]	0.92	0.64	0.80	0.68	0.60	0.78	0.76	0.88	0.76	0.90	0.77
NPV[b]	0.88	0.59	0.75	0.65	0.65	0.73	0.72	0.79	0.68	0.83	0.78
MCC[b]	0.55	0.16	0.43	0.27	0.35	0.38	0.40	0.37	0.27	0.42	0.58
Performance of prediction methods (pathogenic dataset only from LSDBs, not in SwissProt)											
tp	2240	1175	1368	1436	1651	2410	2190	2764	2131	2615	2547
fn	899	862	1252	2158	1943	1184	1361	787	1145	917	952
tn	2655	212	1508	2842	3004	2333	2334	1705	2073	1382	2898
fp	804	165	501	752	534	1205	1028	1657	1268	1069	259
cases +[a]	3139	2037	2620	3594	3594	3594	3551	3551	3276	3532	3499
cases −[a]	3459	377	2009	3594	3538	3538	3362	3362	3341	2451	3157
Accuracy[b]	0.74	0.57	0.64	0.6	0.65	0.66	0.66	0.64	0.64	0.65	0.82
Precision[b]	0.75	0.57	0.68	0.66	0.75	0.66	0.67	0.61	0.63	0.63	0.90
Specificity[b]	0.77	0.56	0.75	0.79	0.85	0.66	0.69	0.51	0.62	0.56	0.92
Sensitivity[b]	0.71	0.58	0.52	0.4	0.46	0.67	0.62	0.78	0.65	0.74	0.73
NPV[b]	0.73	0.57	0.61	0.57	0.61	0.67	0.64	0.70	0.64	0.68	0.77
MCC[b]	0.48	0.14	0.28	0.21	0.33	0.33	0.31	0.30	0.27	0.31	0.66

[a] Total number of cases used by the given program (not normalized).
[b] Accuracy, precision, specificity, sensitivity, NPV, and MCC are calculated from normalised numbers.

Pet. Reh. App.80

and not only the ones trained on Swiss-Prot data, performed worse in this subcategory, we claim our results are not biased, even though we acknowledge that a perfectly fair comparison between methods trained on different datasets cannot be made.

To study the effect of residue types, the mutated and mutant amino acids were assigned into six groups according to their physicochemical properties: hydrophobic (C, F, I, L, M, V, W, and Y), positively charged (H, K, and R), negatively charged (D and E), conformational (G and P), polar (N, Q, and S), and A and T [Shen and Vihinen, 2004]. There were small differences in accuracy and precision of the methods for different types of wild-type or mutant amino acids, but their sensitivity and MCC were dependent on the physicochemical properties of the wild-type and mutant amino acids (Fig. 1). The methods were more sensitive to mutations at conformational, hydrophobic, and positively charged amino acids than mutations at polar residues or A and T (Fig. 1). MCC differed as well depending on the nature of the original residue position, and substitutions at hydrophobic positions were predicted best by most methods. Panther predicted mutations at hydrophobic and positively charged residues with equal performance, and MutPred and SNPs&GO performed better predicting conformational residues. Mutations affecting negatively charged residues had the lowest MCCs by most methods, except for PolyPhen1b, which predicted other classes better than the conformational class, and MutPred, nsSNPAnalyzer, and SNPs&GO, which had the lowest MCC when predicting the effects of mutations altering A and T residues (Fig. 1). The sensitivity and MCC of the methods also varied in predicting the effects of different types of mutant residues. All the methods performed best when the substituting residue was charged, and in the case of nsSNPAnalyzer, polar residues were predicted better than negatively charged residues, and SNAP predicted polar residues better than positively charged residues.

Differences in prediction sensitivity could also be seen at the level of individual amino acids. Predictions for substitutions at C, W, and Y were clearly more sensitive than at other residues by all methods (Fig. 2A). A similar trend was also seen when looking at mutant amino acids: mutations to the aforementioned residues were predicted with better sensitivity (Fig. 2A). The sensitivity of PolyPhen2b and SNAP varied less at individual residues than that of the other programs.

The results for the substitutions in the secondary structural elements are shown in Figure 2B. All of the programs predicted

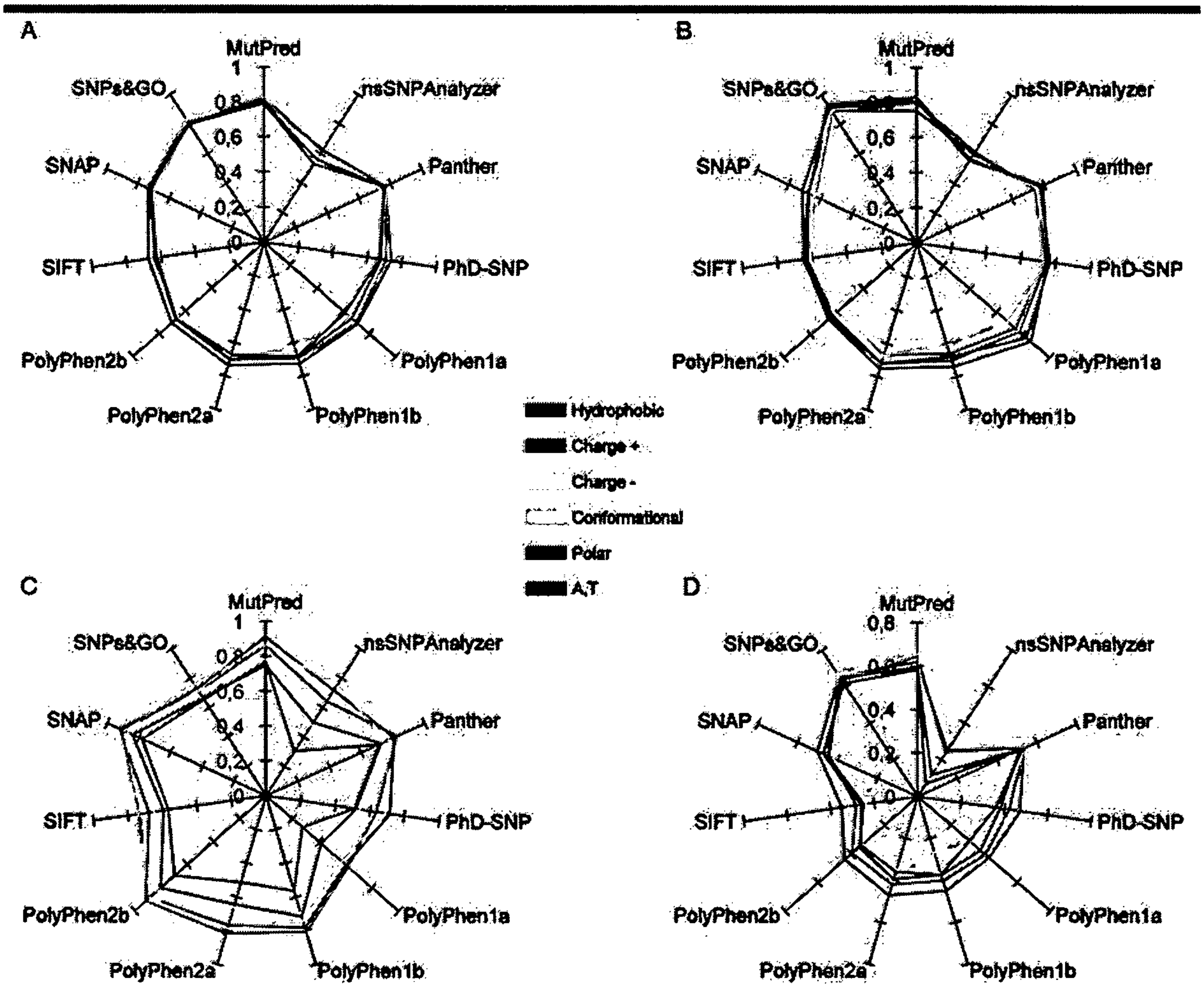

Figure 1. The values of the quality parameters, accuracy, precision, sensitivity, and Matthews correlation coefficient (MCC) for different classes of substituted amino acids. **A:** accuracy, **B:** precision, **C:** sensitivity, and **D:** MCC. Abbreviations: Charge+, positively charged. Charge −, negatively charged. [Color figures can be viewed in the online issue, which is available at wileyonlinelibrary.com]

Pet. Reh. App.81

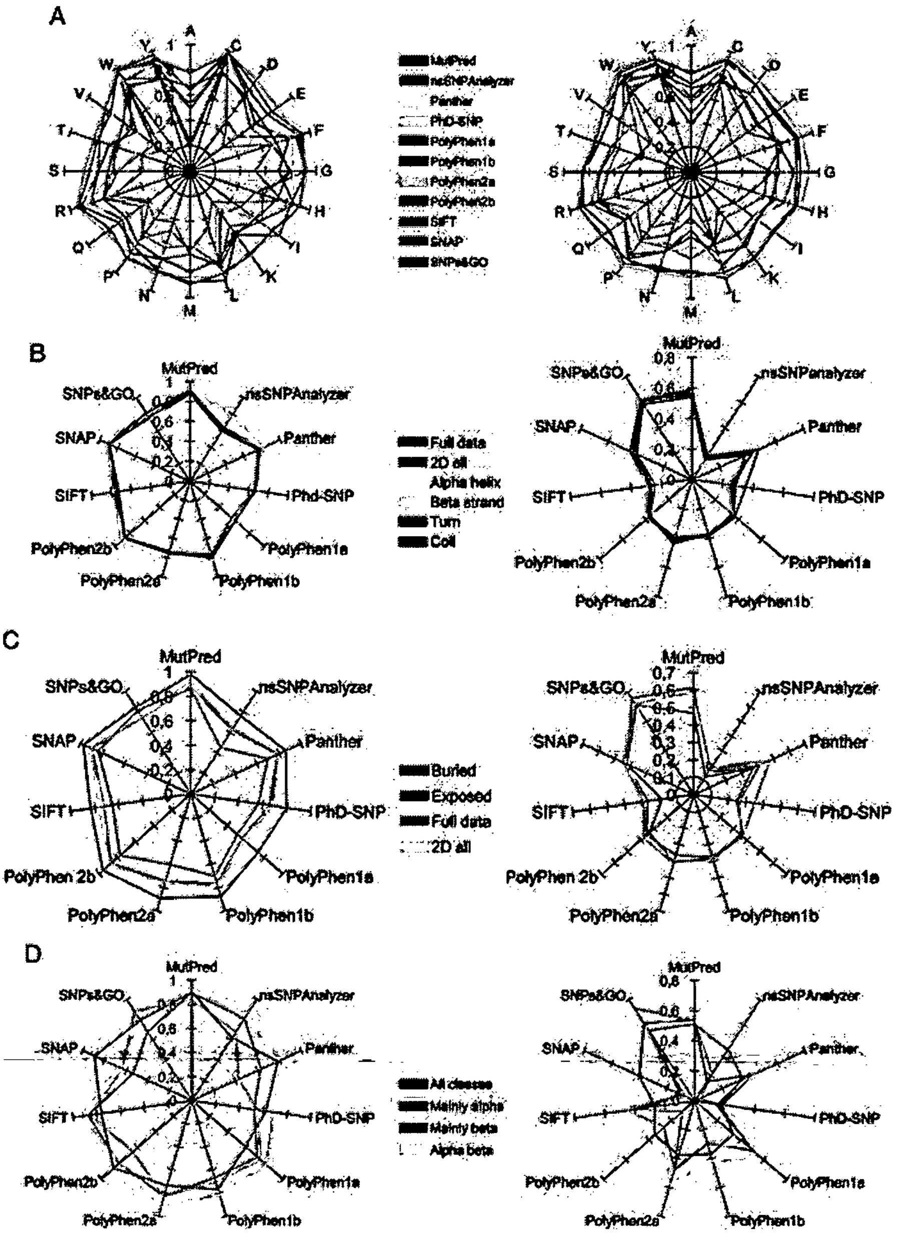

Figure 2. The values of sensitivity and Matthews correlation coefficient (MCC) for different types of amino acid substitutions. **A**: Sensitivity in different amino acid residues. Left: mutated (original) amino acids, right: substituting (mutant) amino acids. **B**: Sensitivity (left) and MCC (right) for amino acid substitutions at different secondary structural elements. **C**: Sensitivity (left) and MCC (right) for amino acid substitutions according to the accessible surface area (ASA) of the position (buried ASA ≤10%, exposed ASA ≥25%). **D**: Sensitivity (left) and MCC (right) for amino acid substitutions at different protein structural classes. [Color figures can be viewed in the online issue, which is available at wileyonlinelibrary.com]

the effects of substitutions at different secondary structures with almost equal accuracy and precision. Sensitivity and MCC values showed more variation with secondary structure. In terms of MCC, MutPred, nsSNPAnalyzer, PolyPhen1b, and PolyPhen2b predicted amino acid substitutions at strands best, whereas Panther, PolyPhen1a, SNAP, and SNPs&GO performed best at turns. PhD-SNP and SIFT predicted substitutions positioned at α-helices best, and PolyPhen2a at coils. The differences in MCC were not striking. Except for Panther, PhD-SNP, and SNPs&GO, all methods were most sensitive when predicting the effects of amino acid substitutions at strands. Solvent-accessible surface areas of the positions did not markedly affect prediction accuracy or precision, but all the methods were more sensitive when predicting the effects of substitutions at buried positions (Fig. 2C). MCC for most methods was better at exposed than buried positions, except for PolyPhen1a and PolyPhen2a, which performed better at buried positions. MCCs for PolyPhen1b and SNAP did not differ with solvent accessibility of the position. These results are not in line

Pet. Reh. App.82

Table 4. Pairwise Prediction Correlations

	MutPred	nsSNPAnalyzer	Panther	PhD-SNP	PolyPhen 1a	PolyPhen 1b	PolyPhen 2a	PolyPhen 2b	SIFT	SNAP	SNPs&GO
MutPred		8721	22645	36300	36522	36522	35198	35198	32705	29674	34066
nsSNPAnalyzer	4620		7237	9225	9380	9380	9353	9353	8270	8609	9145
Panther	15296	3589		23671	23869	23869	23406	23406	21540	20713	22555
PhD-SNP	23955	4389	14838		39659	39659	38254	38254	34532	32203	37095
PolyPhen1a	22125	4386	13961	22756		40146	38485	38485	34683	32533	37324
PolyPhen1b	22208	4965	14701	22170	23764		38485	38485	34683	32533	37324
PolyPhen2a	22234	4777	14728	21871	22383	23156		38782	33686	31790	36317
PolyPhen2b	20911	5012	14288	20042	19656	22412	24006		33686	31790	36317
SIFT	18807	4302	12623	18879	18207	18985	18645	17833		28726	32434
SNAP	18877	4750	13307	18004	17024	19811	19321	19945	16393		30987
SNPs&GO	23220	4672	14285	23333	22544	22206	22042	20569	18135	18833	
MutPred		53.0	67.5	66.0	60.6	60.8	63.2	59.4	57.5	63.6	68.2
nsSNPAnalyzer	0.36		49.6	47.6	46.8	52.9	51.1	53.6	52.0	55.2	51.1
Panther	0.54	0.37		62.7	58.5	61.6	62.9	61.0	58.6	64.2	63.3
PhD-SNP	0.57	0.35	0.51		57.4	55.9	57.2	52.4	54.7	55.9	62.9
PolyPhen1a	0.43	0.44	0.46	0.45		59.2	58.2	51.1	52.5	52.3	60.4
PolyPhen1b	0.43	0.47	0.50	0.43	0.66		58.2	58.2	54.7	60.9	59.5
PolyPhen2a	0.49	0.44	0.51	0.45	0.56	0.58		61.9	55.3	55.3	60.7
PolyPhen2b	0.44	0.42	0.49	0.40	0.46	0.57	0.72		52.9	62.7	56.6
SIFT	0.41	0.53	0.48	0.45	0.45	0.52	0.50	0.51		57.0	55.9
SNAP	0.46	0.41	0.51	0.44	0.44	0.54	0.52	0.53	0.53		60.8
SNPs&GO	0.50	0.25	0.39	0.44	0.39	0.38	0.38	0.34	0.34	0.39	

Upper table: the number of cases shared by two programs (upper right triangle). The number of cases predicted correctly (lower left triangle). Lower table: The number of cases predicted correctly, reported as a percentage (upper right triangle). Pairwise correlation (lower left triangle).

with a previous study [Mort et al., 2010], where a sequence conservation based method yielded results of lower accuracy when predicting the effects of solvent-exposed residues.

CATH classifies proteins as mainly α-helical or β-stranded, mixed α- and β-structures (α–β), or as having few secondary structures. Interestingly, none of the proteins included in this analysis was assigned into the few secondary structures class. The predictions differed with respect to sensitivity and MCC depending on which protein class a mutation appeared (Fig. 2D). Most programs were more sensitive to amino acid substitutions in the α–β class of proteins, but SNPs&GO predicted substitutions best in the mainly β-class. nsSNPAnalyzer predicted those mutations occurring in α–β and α-helical proteins or domains with equal sensitivity. MCCs varied significantly with the structural class of proteins, especially in the predictions by nsSNPAnalyzer, PolyPhen1b, PolyPhen2a, and 2b, and SNPs&GO. The results were generally better for the α–β class of proteins, but nsSNPAnalyzer predicted substitutions at α-helical proteins best and SNPs&GO performed best with proteins in the mainly β-class.

To further evaluate the performance of the programs we compared them in a pairwise fashion (Table 4). The numbers of cases that were shared by the programs varied because the number of cases that could be predicted by each program varied as described in the Materials and Methods section. The largest percentage of correctly predicted cases by two programs was 68.2% (for the combination of MutPred and SNPs&GO). On average, the fraction of correctly predicted cases between any two programs was 57.7%. The correlations between two programs were highest for MutPred and PhD-SNP (0.57), and for PolyPhen 1 and 2 (0.57 for the less stringent b versions, and 0.56 for the a versions) (without taking into account the higher correlation between PolyPhen1a or 2a and PolyPhen1b or 2b that are different forms of the same program). Correlation was lowest for nsSNPAnalyzer and SNPs&GO (0.25).

Discussion

In this study we evaluated how reliably the pathogenicity of missense mutants can be predicted, and whether selected features of the variant or the structural context affect prediction performance. The processing of the vast and increasing amount of genetic variation data requires the development of automatic annotation tools to determine the potential pathological character of a given variant. Prioritizing the most interesting and likely pathogenic cases for experimental analysis is another important application of the tested prediction methods.

To our knowledge, no comprehensive evaluation of the performance of missense variant pathogenicity predictors has been made outside the performance studies of individual methods in the context of their development. We selected test sets that have not been used in the training of the methods as such, but a subset of the pathogenic dataset is comprised of mutations from Swiss-Prot, and some methods (MutPred, nsSNPAnalyzer, PhD-SNP, PolyPhen2, and SNPs&GO) have used Swiss-Prot mutations in the training of the method. Testing of the performance of a method with the same cases it was trained on would lead into biased results, so that those methods trained on SwissProt mutations would have an advantage over the other methods. However, because the pathogenic dataset includes a large number of LSDB variations not found in SwissProt, we claim the test set was not similar to the training sets to the extent that it would advantage those methods trained on SwissProt data. Further, we tested the methods with cases coming only from LSDBs. With this dataset the performance decreased with all methods, whether trained on Swiss-Prot data or not, except for SNPs&GO. This indicates that the good performance of SNP&GO was not a result of that it has previously been exposed to the test dataset during its training phase. Furthermore, the poor performance of PhD-SNP indicates the method did not benefit from the possible identical cases in the data used for training and testing. However, it is impossible to construct a large testing dataset that would not share any cases with the original training sets of any of the methods, especially when the specific contents of the training sets are rarely published.

The neutral dataset was generated from dbSNP entries that had a frequency higher than 1% when there was data at least for 25 individuals (50 chromosomes). This way the number of false negatives could be minimized in the test set.

Pet. Reh. App.83

There are still other pathogenicity predictors that we did not evaluate. SNPs3D [Yue et al., 2006] was not included in this study because it does not allow submission of user-defined amino acid substitutions. Similarly, LS-SNP [Karchin et al., 2005] is an annotated database of SNPs, not a prediction method for any user-provided variant, although often referred to as a prediction method for nsSNP pathogenicity. The Auto-Mute predictor of disease potential of human nsSNPs [Barenboim et al., 2008] was left out from the analysis because the program did not allow batch submission. PMut [Ferrer-Costa et al., 2005] could not be tested because the server did not return predictions.

Overall, we found SNPs&GO and MutPred to be clearly the most reliable predictors for our dataset of genetic variants. The accuracies of all the methods were in the range of 0.60–0.82, and precision ranged from 0.59 to 0.90. More variation among the methods was seen when considering the sensitivities and MCC values that ranged from 0.52 to 0.88 and 0.19 to 0.65, respectively. The local structural context of a mutated residue did not dramatically affect predictor performance in most cases but most methods showed variance in their prediction power at the level of protein tertiary structure classification and at different mutated positions.

Studies have shown that combining information obtained from the multiple sequence alignment and three-dimensional protein structure can increase prediction performance [Bromberg and Rost, 2007; Saunders and Baker, 2002]. According to our results, this is not always the case. Panther operates solely on sequence-based evolutionary information, and it is one of the best performing methods, outperforming all the methods incorporating structural information in the prediction, except for MutPred, which uses sequence-derived structural predictions as features in combination with evolutionary information. Furthermore, although nsSNPAnalyzer uses the SIFT method for the evolutionary analysis and also includes structure-derived features, its overall performance is below that for SIFT, except for an increase in specificity in the structure subset of data. However, the two best performing predictors include both protein structural or functional and MSA-derived information in the prediction.

It is very difficult to determine whether the notable differences in the performance of these methods are caused by differences in the features utilized by the methods or the training datasets. For example, SNPs&GO uses GO annotations as a feature, and GO is biased toward genes involved in diseases. The PDB is biased as well, containing structures of mostly well-studied proteins, which include products of disease-related genes. Therefore, one would expect SNPs&GO would perform better in predicting the effects of missense variants in proteins that have structures in the PDB as they are likely to have GO annotation as well—and in fact, it performs worse. One factor that very probably affects prediction reliability is the quality of multiple sequence alignment. Because all of the methods studied here use MSA as input to the prediction, the quality of the provided MSA should be very carefully assessed. For many of the methods, we did not find documentation how the MSA is constructed when the user provides just the query sequence as input. For example, an automatic BLAST search often performed by the programs may lead into construction of an MSA that contains multiple versions of the same sequence or paralog sequences, affecting the resulting conservation analysis. The MSA should contain a selection of closely and distantly related sequences in order to effectively yield a conservation signal.

In conclusion, those methods that performed best had high accuracy (reaching 0.82, SNPs&GO), precision (0.90, SNPs&GO), specificity (0.92, SNPs&GO), sensitivity (0.88, SNAP), and NPV (0.84, MutPred). Matthews correlation coefficient reached the value of 0.65 at best (SNPs&GO). There is no single method that could be rated as best by all parameters, so the user should consider what aspects would be most valuable considering the nature of the data analysed. Furthermore, some methods require 3D structure coordinates, limiting the number of cases that can be analyzed (nsSNPAnalyzer), and some methods are at least currently too slow for high-throughput analyses (SNAP). Although some of the existing methods perform reasonably well, development of new, more reliable methods is certainly needed. Complementary methods could be combined in a metaserver to yield more reliable predictions.

Acknowledgments

The authors thank Pier Luigi Martelli and Rita Casadio from the University of Bologna, Biao Li from Indiana University, and Sean Mooney and Vidhya Krishnan from the Buck Institute for Age Research, for cooperation in running of data.

References

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. Nat Methods 7:248–449.

Ahmad S, Gromiha MM, Sarai A. 2004. Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. Bioinformatics 20:477–486.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25–29.

Bairoch A, Apweiler R. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 28:45–48.

Baldi P, Brunak S, Chauvin Y, Andersen CAF, Nielsen H. 2000. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16:412–424.

Bao L, Zhou M, Cui Y. 2005. nsSNPAnalyzer: identifying disease-associated non-synonymous single nucleotide polymorphisms. Nucleic Acids Res 33:W480–W482.

Barenboim M, Masso M, Vaisman, II, Jamison DC. 2008. Statistical geometry based prediction of nonsynonymous SNP functional effects using random forest and neuro-fuzzy classifiers. Proteins 71:1930–1939.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The Protein Data Bank. Nucleic Acids Res 28:235–242.

Bromberg Y, Rost B. 2007. SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic Acids Res 35:3823–3835.

Calabrese R, Capriotti E, Fariselli P, Martelli PL, Casadio R. 2009. Functional annotations improve the predictive score of human disease-related mutations in proteins. Hum Mutat 30:1237–1244.

Capriotti E, Calabrese R, Casadio R. 2006. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. Bioinformatics 22:2729–2734.

Capriotti E, Fariselli P, Casadio R. 2005. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res 33: W306–W310.

Chandonia JM, Hon G, Walker NS, Lo Conte L, Koehl P, Levitt M, Brenner SE. 2004. The ASTRAL Compendium in 2004. Nucleic Acids Res 32:D189–D192.

Chou PY, Fasman GD. 1974. Prediction of protein conformation. Biochemistry 13: 222–245.

Daily KM, Radivojac P., Dunker AK. 2005. Intrinsic disorder and protein modifications: building an SVM predictor for methylation. IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB: 475–481.

Delorenzi M, Speed T. 2002. An HMM model for coiled-coil domains and a comparison with PSSM-based predictions. Bioinformatics 18:617–625.

Ferrer-Costa C, Gelpí JL, Zamakola L, Parraga I, de la Cruz X, Orozco M. 2005. PMUT: a web-based tool for the annotation of pathological mutations on proteins. Bioinformatics 21:3176–3178.

Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer EL, Eddy SR, Bateman A. 2010. The Pfam protein families database. Nucleic Acids Res 3:D211–D222.

Pet. Reh. App.84

Frishman D, Argos P. 1995. Knowledge-based protein secondary structure assignment. Proteins 23:566–579.

Giardine B, Riemer C, Hefferon T, Thomas D, Hsu F, Zielenski J, Sang Y, Elnitski L, Cutting G, Trumbower H, and others. 2007. PhenCode: connecting ENCODE data with mutations and phenotype. Hum Mutat 28:554–562.

Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, Obradovic Z, Dunker AK. 2004. The importance of intrinsic disorder for protein phosphorylation. Nucleic Acids Res 32:1037–1049.

Johnson GC, Todd JA. 2000. Strategies in complex disease mapping. Curr Opin Genet Dev 10:330–334.

Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637.

Karchin R. 2009. Next generation tools for the annotation of human SNPs. Brief Bioinform 10:35–52.

Karchin R, Diekhans M, Kelly L, Thomas DJ, Pieper U, Eswar N, Haussler D, Sali A. 2005. LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. Bioinformatics 21:2814–2820.

Kawabata T, Ota M, Nishikawa K. 1999. The protein mutant database. Nucleic Acids Res 27:355–357.

Khan S, Vihinen M. 2007. Spectrum of disease-causing mutations in protein secondary structures. BMC Struct Biol 7:56.

Khan S, Vihinen M. 2010. Performance of protein stability predictors. Hum Mutat 31:675–684.

Krogh A, Larsson B, von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305:567–580.

Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, Mooney SD, Radivojac P. 2009. Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics 25:2744–2750.

Lupas A, Van Dyke M, Stock J. 1991. Predicting coiled coils from protein sequences. Science 252:1162–1164.

Matthews BW. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 405:442–451.

McDonald IK, Thornton JM. 1994. Satisfying hydrogen bonding potential in proteins. J Mol Biol 238:777–793.

Mooney S. 2005. Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis. Brief Bioinform 6:44–56.

Mort M, Evani US, Krishnan VG, Kamati KK, Baenziger PH, Bagchi A, Peters B, Sathyesh R, Li B, Sun Y, Xue B, Shah NH, Kann MG, Cooper DN, Radivojac P, Mooney SD. 2010. In silico functional profiling of human disease-associated and polymorphic amino acid substitutions. Hum Mutat 31:335–346.

Ng PC, Henikoff S. 2001. Predicting deleterious amino acid substitutions. Genome Res 11:863–874.

Ng PC, Henikoff S. 2003. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res 31:3812–3814.

Ng PC, Henikoff S. 2006. Predicting the effects of amino acid substitutions on protein function. Annu Rev Genomics Hum Genet 7:61–80.

Ng PC, Henikoff JG, Henikoff S. 2000. PHAT: a transmembrane-specific substitution matrix. Predicted hydrophobic and transmembrane. Bioinformatics 16:760–766.

Nielsen H, Engelbrecht J, Brunak S, von Heijne G. 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng 10:1–6.

Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. 1997. CATH—a hierarchic classification of protein domain structures. Structure 5: 1093–1108.

Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z. 2006. Length-dependent prediction of protein intrinsic disorder. BMC Bioinformatics 7:208.

Piirilä H, Väliaho J, Vihinen M. 2006. Immunodeficiency mutation databases (IDbases). Hum Mutat 27:1200–1208.

Radivojac P, Obradovic Z, Smith DK, Zhu G, Vucetic S, Brown CJ, Lawson JD, Dunker AK. 2004. Protein flexibility and intrinsic disorder. Protein Sci 13: 71–80.

Radivojac P, Vacic V, Haynes C, Cocklin RR, Mohan A, Heyen JW, Goebl MG, Iakoucheva LM. 2010. Identification, analysis, and prediction of protein ubiquitination sites. Proteins 78:365–380.

Radivojac P, Vucetic S, O'Connor TR, Uversky VN, Obradovic Z, Dunker AK. 2006. Calmodulin signaling: analysis and prediction of a disorder-dependent molecular recognition. Proteins 63:398–410.

Ramensky V, Bork P, Sunyaev S. 2002. Human non-synonymous SNPs: server and survey. Nucleic Acids Res 30:3894–3900.

Raney BJ, Cline MS, Rosenbloom KR, Dreszer TR, Learned K, Barber GP, Meyer LR, Sloan CA, Malladi VS, Roskin KM, Suh BB, Hinrichs AS, Clawson H, Zweig AS, Kirkup V, Fujita PA, Rhead B, Smith KE, Pohl A, Kuhn RM, Karolchik D, Haussler D, Kent WJ. 2011. 0ENCODE whole-genome data in the UCSC Genome Browser (2011 update). Nucleic Acids Res 39(Database issue):871–875.

Rost B. 1996. PHD: predicting one-dimensional protein structure by profile-based neural networks. Methods Enzymol 266:525–539.

Rost B, Sander C. 1994. Conservation and prediction of solvent accessibility in protein families. Proteins 20:216–226.

Saunders CT, Baker D. 2002. Evaluation of structural and evolutionary contributions to deleterious mutation prediction. J Mol Biol 322:891–901.

Schlessinger A, Yachdav G, Rost B. 2006. PROFbval: predict flexible and rigid residues in proteins. Bioinformatics 22:891–893.

Shen B, Vihinen M. 2004. Conservation and covariance in PH domain sequences: physicochemical profile and information theoretical analysis of XLA-causing mutations in the Btk PH domain. Protein Eng Des Sel 17:267–276.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29:308–311.

Steward RE, MacArthur MW, Laskowski RA, Thornton JM. 2003. Molecular basis of inherited diseases: a structural perspective. Trends Genet 19:505–513.

Sunyaev SR, Eisenhaber F, Rodchenkov IV, Eisenhaber B, Tumanyan VG, Kuznetsov EN. 1999. PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. Protein Eng 12: 387–394.

Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A. 2003. PANTHER: a library of protein families and subfamilies indexed by function. Genome Res 13:2129–2141.

Thusberg J, Vihinen M. 2009. Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods. Hum Mutat 30: 703–714.

Yip YL, Famiglietti M, Gos A, Duek PD, David FP, Gateau A, Bairoch A. 2008. Annotating single amino acid polymorphisms in the UniProt/Swiss-Prot knowledgebase. Hum Mutat 29:361–366.

Yip YL, Scheib H, Diemand AV, Gattiker A, Famiglietti LM, Gasteiger E, Bairoch A. 2004. The Swiss-Prot variant page and the ModSNP database: a resource for sequence and structure information on human protein variants. Hum Mutat 23: 464–470.

Yue P, Melamud E, Moult J. 2006. SNPs3D: candidate gene and SNP selection for association studies. BMC Bioinformatics 7:166.

Pet. Reh. App.85

SCIENTIFIC REPORTS
natureresearch

OPEN

Identification of pathogenic missense mutations using protein stability predictors

Lukas Gerasimavicius, Xin Liu & Joseph A. Marsh✉

Attempts at using protein structures to identify disease-causing mutations have been dominated by the idea that most pathogenic mutations are disruptive at a structural level. Therefore, computational stability predictors, which assess whether a mutation is likely to be stabilising or destabilising to protein structure, have been commonly used when evaluating new candidate disease variants, despite not having been developed specifically for this purpose. We therefore tested 13 different stability predictors for their ability to discriminate between pathogenic and putatively benign missense variants. We find that one method, FoldX, significantly outperforms all other predictors in the identification of disease variants. Moreover, we demonstrate that employing predicted absolute energy change scores improves performance of nearly all predictors in distinguishing pathogenic from benign variants. Importantly, however, we observe that the utility of computational stability predictors is highly heterogeneous across different proteins, and that they are all inferior to the best performing variant effect predictors for identifying pathogenic mutations. We suggest that this is largely due to alternate molecular mechanisms other than protein destabilisation underlying many pathogenic mutations. Thus, better ways of incorporating protein structural information and molecular mechanisms into computational variant effect predictors will be required for improved disease variant prioritisation.

Advances in next generation sequencing technologies have revolutionised research of genetic variation, increasing our ability to explore the basis of human disorders and enabling huge databases covering both pathogenic and putatively benign variants[1,2]. Novel sequencing methodologies allow the rapid identification of variation in the clinic and are helping facilitate a paradigm shift towards precision medicine[3,4]. Despite this, however, it remains challenging to distinguish the small fraction of variants with medically relevant effects from the huge background of mostly benign human genetic variation.

A particularly important research focus is single nucleotide variants that lead to amino acid substitutions at the protein level, i.e. missense mutations, which are associated with more than half of all known inherited diseases[5,6]. A large number of computational methods have been developed for the identification of potentially pathogenic missense mutations, i.e. variant effect predictors. Although different approaches vary in their implementation, a few types of information are most commonly used, including evolutionary conservation, changes in physiochemical properties of amino acids, biological function, known disease association and protein structure[7]. While these predictors are clearly useful for variant prioritisation, and show a statistically significant ability to distinguish known pathogenic from benign variants, they still make many incorrect predictions[8-10], and the extent to which we can rely on them for aiding diagnosis remains limited[11].

An alternative approach to understanding the effects of missense mutations is with computational stability predictors. These are programs that have been developed to assess folding or protein interaction energy changes upon mutation (change in Gibbs free energy – ΔΔG in short). This can be achieved by approximating structural energy through linear physics-based pairwise energy scoring functions, their empirical and knowledge-based derivatives, or a mixture of such energy terms. Statistical and machine learning methods are employed to parametrise the scoring models. These predictors have largely been evaluated against their ability to predict experimentally determined ΔΔG values. Great effort has been previously made to assess stability predictor performance in producing accurate or well-correlated energy change estimates upon mutation, as well as assessing their shortfalls, such as biases arising from destabilising variant overrepresentation in training sets and lack of

MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh EH4 2XU, UK. ✉email: joseph.marsh@igmm.ed.ac.uk

self-consistency predicting forward–backward substitutions[12–18]. Several predictors have since been shown to alleviate such issues through their specific design or have been improved in this regard[14,19,20]. Moreover, the practical utility of stability predictors has been demonstrated through their extensive usage in the fields of protein engineering and design[21–23].

Although computational stability predictors have not been specifically designed to identify pathogenic mutations, they are very commonly used when assessing candidate disease mutations. For example, publications reporting novel variants will often include the output of stability predictors as evidence in support of pathogenicity[24–27]. This relies essentially upon the assumption that the molecular mechanism underlying many or most pathogenic mutations is directly related to the structural destabilisation of protein folding or interactions[28–31]. However, despite their widespread application to human variants, there has been little to no systematic assessment of computational stability predictors for their ability to predict disease mutations. A number of studies have assessed the real-world utility for individual protein targets and families using certain stability predictors[32–36]. However, numerous computational stability predictors have now been developed and, overall, we still do not have a good idea of which methods perform best for the identification of disease mutations, and how they compare relative to other computational variant effect predictors.

In this work, we explore the applicability and performance of 13 methodologically diverse structure-based protein stability predictors for distinguishing between pathogenic and putatively benign missense mutations. We find that FoldX significantly outperforms all other stability predictors for the identification of disease mutations, and also demonstrate the practical value of using predicted absolute ΔΔG values to account for potentially overstabilising mutations. However, this work also highlights the limitations of stability predictors for predicting disease, as they still miss many pathogenic mutations and perform worse than many variant effect predictors, thus emphasising the importance of considering alternate molecular disease mechanisms beyond protein destabilisation.

Results

We tested 13 different computational stability predictors on the basis of accessibility, automation or batching potential, computation speed, as well as recognition—and included FoldX[37], INPS3D[38], Rosetta[37], PoPMusic[39], I-Mutant[40], SDM[41], SDM2[42], mCSM[43], DUET[44], CUPSAT[45], MAESTRO[46], ENCoM[47] and DynaMut[48] (Table 1). We ran each predictor against 13,508 missense mutations from 96 different high-resolution (< 2 Å) crystal structures of disease-associated monomeric proteins. Our disease mutation dataset was comprised of 3,338 missense variants from ClinVar[2] annotated as pathogenic or likely pathogenic, and we only included proteins with at least 10 known pathogenic missense mutations occurring at residues present in the structure. We compared these to 10,170 missense variants observed in the human population, taken from gnomAD v2.1[1], which we refer to as "putatively benign". We acknowledge that it is likely that some of these gnomAD variants could be pathogenic under certain circumstances (e.g. if observed in a homozygous state, if they cause late-onset disease, or there is incomplete penetrance), or they may be damaging but lead to a subclinical phenotype. However, the large majority of gnomAD variants will be non-pathogenic, and we believe that our approach of represents a good test of the practical utilisation of variant effect predictors, where the main challenge is in distinguishing severe pathogenic mutations from others observed in the human population. While filtering by allele frequency would give us variants that are more likely to be truly benign, it would also dramatically reduce the size of the dataset (e.g. only ~ 1% of missense variants in gnomAD have an allele frequency > 0.1%). Thus, we have not filtered the gnomAD variants (other than to exclude known pathogenic variants present in the ClinVar set).

To investigate the utility of the computational stability predictors for the identification of pathogenic missense mutations, we used receiver operating characteristic (ROC) plots to assess the ability of ΔΔG values to distinguish between pathogenic and putatively benign mutations (Fig. 1A). This was quantifed by the area under the curve (AUC), which is equal to the probability of a randomly chosen disease mutation being assigned a higher-ranking score than a random benign one. Of the 13 tested structure-based ΔΔG predictors, FoldX performs the best as a predictor of human missense mutation pathogenicity, with an AUC value of 0.661. This is followed by INPS3D at 0.640, Rosetta at 0.617 and PoPMusic at 0.614. Evaluating the performance through bootstrapping, we found that the difference between FoldX and other predictors is significant, with a p value of 2×10^{-4} compared to INSP3D, 1×10^{-7} for Rosetta and 8×10^{-9} for PoPMusiC. The remaining predictors show a wide range of lower performance values.

Two predictors, ENCoM and DynaMut, stand out for their unusual pattern in the ROC plots, with a rotated sigmoidal shape where the false positive rate becomes greater than the true positive rate at higher levels. Close inspection of the underlying data shows that this is indicative of the predicted energy change distribution tails for the disease-associated class extending both directions away from the putatively benign missense mutation score density. This suggests that a considerable portion of pathogenic missense mutations are predicted by these methods to excessively stabilise the protein.

While the analysis (Fig. 1A) assumes that protein destabilisation should be indicative of mutation pathogenicity, it also possible for mutations that increase protein stability to cause disease[49,50]. Recent research has shown that absolute ΔΔG values, which treat stabilisation and destabilisation equivalently, may be better indicators of disease association[51,52]. Therefore, we repeated the analysis using absolute ΔΔG values (Fig. 1B). This improved the performance of most predictors, while not reducing the performance of any. The most drastic change was observed for ENCoM, which improved from worst to fifth best predictor, with an increase in AUC from 0.495 to 0.619. However, the top four predictors, FoldX, INPS3D, Rosetta and PoPMuSiC, improve only slightly and do not change in ranking.

Using the ROC point distance to the top-left corner[53], we establish the best disease classification ΔΔG value for each predictor when assessing general perturbation (Table 2). It is interesting to note that FoldX demonstrates

Predictor	Link	Description
DynaMut[48]	https://biosig.unimelb.edu.au/dynamut/	Consensus predictor which uses outputs from Bio3D, ENCoM and DUET to assess the impact of mutations on protein stability. Due to its nature, the predictor leverages multiple methodologies, such as normal mode analysis and statistical potentials
ENCoM[47]	No longer available as a stand-alone server, but available from DynaMut	A prediction method based on normal mode analysis that relates changes in vibrational entropy upon mutation to changes in protein stability. Uses coarse-grained protein representations that accounts for residue properties
DUET[44]	https://biosig.unimelb.edu.au/duet/stability	A machine-learnt consensus predictor that leverages output from SDM and mCSM, integrated using support vector machines
SDM[41]	No longer available as a stand-alone server (succeeded by the SDM2 webserver), but available from DynaMut	A knowledge-based energy potential, derived using evolutionary environment-specific residue substitution propensities
FoldX[76]	https://foldxsuite.crg.eu/	A full-atom force field consisting of physics-based interaction and entropic terms, parametrised on empirical training data. Allows to easily run predictions on multi-chain assemblies
Rosetta[37]	https://www.rosettacommons.org/home	Rosetta macromolecular modelling software suite, which includes algorithms for stability impact prediction. Driven by a scoring function that is a linear combination of statistical and empirical energy terms. Highly modular and customisable
INPS3D[38]	https://inpsmd.biocomp.unibo.it/inpsSuite/default/index3D	INPS3D builds upon its sequence and physicochemical conservation-based predecessor INPS, and employs structure-derived features such as solvent accessibility and local energy differences. The predictor is trained by employing support vector regression
mCSM[43]	https://biosig.unimelb.edu.au/mcsm/stability	A machine-learned approach that evaluates structural signature changes imparted by mutations. Derives graph representation of physicochemical and geometric residue environment features
SDM2[42]	https://marid.bioc.cam.ac.uk/sdm2/prediction	Updated version of SDM, a knowledge-based potential, which uses environment-specific residue substitution tables, information on residue conformation and interactions, as well as packing density and residue depth, to assess protein stability changes
CUPSAT[45]	https://cupsat.tu-bs.de/	Prediction method that uses a residue torsion angle potential and an environment-specific atom pair potential (an improvement upon amino-acid potentials) to assess stability changes
PoPMuSiC[39]	https://soft.dezyme.com/query/create/pop	A potential consisting of 13 statistical terms, volume difference between the wild-type and mutant residues, as well as the solvent accessibility of the original residue to differentiate core and surface substitutions
MAESTRO[46]	https://pbwww.che.sbg.ac.at/maestro/web	Combines 3 statistical scoring functions of solvent exposure and residue pair distances, as well as 6 protein properties, in a machine-learning framework to derive a consensus stability impact prediction
I-Mutant 3.0[40]	https://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant3.0/I-Mutant3.0.cgi	A machine-learning derived method that takes into account mutated residue spatial environment in terms of surrounding residue types and surface accessibility

Table 1. Protein stability predictors used in this study.

the best classification performance when utilising 1.58 kcal/mol as the stability change threshold, which is remarkably close to the value of 1.5 kcal/mol previously suggested and used in a number of other works when assessing missense mutation impact on stability[13,35,54]. Of course, these threshold values should be considered far from absolute rules, and there are many pathogenic and benign mutations above and below the thresholds for all predictors. For example, nearly 40% of pathogenic missense mutations have FoldX values lower than the threshold, whereas approximately 35% of putatively benign variants are above the threshold.

To account for the class imbalance between putatively benign and pathogenic variants (roughly 3-to-1) in our dataset, we also performed precision-recall curve analysis. While the AUC of PR curves, unlike ROC, does not have a straightforward statistical interpretation, we again based the predictor performance according to this metric. From Fig. S1, it is apparent that the top four best predictors, according to both raw and absolute ΔΔG values, remain the same as in the ROC analysis—FoldX, INPS3D, Rosetta and PoPMuSiC, respectively.

We also calculated ROC AUC values for each protein separately and compared the distributions across predictors (Fig. 2). FoldX again performs much better than other stability predictors for the identification of pathogenic mutations, with a mean ROC of 0.681, compared to INPS3D at 0.655, Rosetta at 0.627, PoPMuSiC at 0.621, and ENCoM at 0.630. Notably, the protein-specific performance was observed to be extremely heterogeneous across all predictors. While some predictors performed extremely well ($AUC > 0.9$) for certain proteins, each predictor has a considerable number of proteins for which they perform worse than random classification ($AUC < 0.5$).

Using the raw and absolute ΔΔG scores, we explored the similarities between different predictors by calculating Spearman correlations for all mutations between all pairs of predictors (Fig. S2). It is apparent that, outside of improved method versions and their predecessors, as well as consensus predictors and their input components, independent methods do not show correlations above 0.65. Furthermore, correlations on the absolute scale appear to slightly decrease in the majority of cases, with exceptions like ENCoM becoming more correlated with FoldX and INPS3D, while at the same time decoupling from DynaMut—a consensus predictor which uses it as input. Interestingly, FoldX and INSP3D, the best two methods, only correlate at 0.50 and 0.48 for raw and absolute ΔΔG values, respectively, which could indicate potential for deriving a more effective consensus methodology.

Pet. Reh. App:88

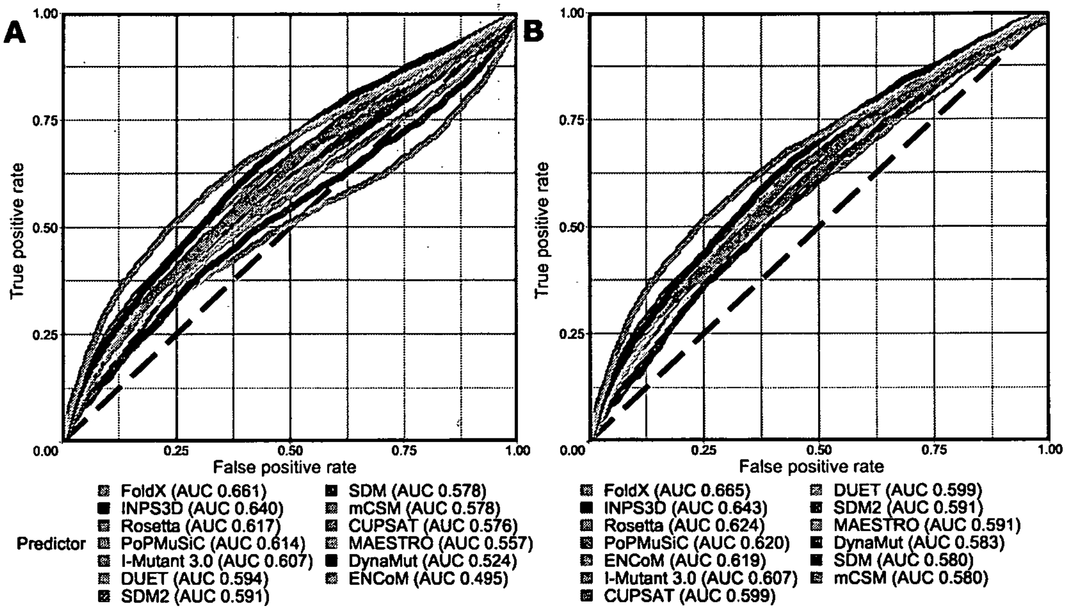

Figure 1. Using ΔΔG values from protein stability predictors to discriminate between pathogenic and putatively benign missense variants. Receiver operating characteristic (ROC) curves are plotted for each predictor, with the classification performance being presented next to its name in the form of area under the curve (AUC). (**A**) ROC curves for classification performance using native ΔΔG value scale for each predictor. (**B**) ROC curves for predictor classification performance when using absolute ΔΔG values. The figure was generated in R v3.6.3 (https://www.r-project.org) using ggplot2 v3.3.0 (https://ggplot2.tidyverse.org/), both freely available.

Predictor	Absolute ΔΔG threshold	False positive rate (95% confidence interval)	True positive rate (95% confidence interval)
FoldX	1.578	0.339–0.357	0.591–0.624
INPS3D	0.674	0.389–0.409	0.595–0.628
Rosetta	1.886	0.390–0.409	0.572–0.605
PoPMuSiC	0.795	0.417–0.437	0.584–0.618
CUPSAT	1.455	0.415–0.434	0.549–0.583
MAESTRO	0.321	0.418–0.437	0.544–0.578
SDM	1.025	0.350–0.370	0.477–0.511
SDM2	0.875	0.365–0.385	0.510–0.544
mCSM	0.889	0.433–0.453	0.542–0.575
DUET	0.803	0.400–0.421	0.548–0.582
I-Mutant 3.0	0.915	0.405–0.424	0.545–0.578
ENCoM	0.221	0.415–0.436	0.598–0.632
DynaMut	0.476	0.446–0.467	0.570–0.605

Table 2. Best stability predictor classification thresholds according to 'distance-to-corner' metric. The performance metrics and their 95% confidence intervals were derived from 2000 bootstraps of the data.

Finally, we compared the performance of protein stability predictors to a variety of different computational variant effect predictors (Fig. 3). Importantly, we excluded any predictors trained using supervised learning techniques, as well as meta-predictors that utilise the outputs of other predictors, thus including only predictors we labelled as unsupervised and empirical in our recent study[10]. This is due to the fact that predictors based upon supervised learning are likely to have been directly trained on some of the same mutations used in our evaluation dataset, making a fair comparison impossible[10,55]. A few predictors perform substantially better than FoldX, with the best performance seen for SIFT4G[56], a modified version of the SIFT algorithm[57]. Interestingly, FoldX and INPS3D are the only stability predictors to outperform the BLOSUM62 substitution matrix[58]. On the other hand, all stability predictors performed better than a number of simple evolutionary constraint metrics.

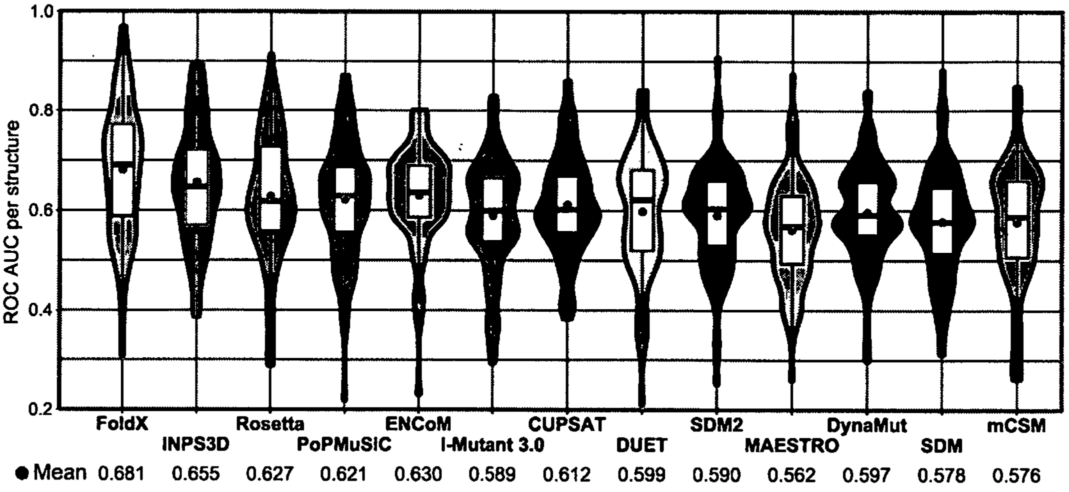

Figure 2. The heterogeneity of protein-specific missense variant classification performance. All the stability predictors exhibit very high degrees of heterogeneity in their protein-specific performance, as measured by the ROC AUC on a per-protein basis. Absolute ΔΔG values were used during protein-specific tool assessment. The mean performance of each predictor is indicated by a red dot and numerically showcased below the plot. Boxes inside the violins illustrate the interquartile range (IQR) of the protein-specific performance points, with the whiskers measuring 1.5 IQR. Boxplot outliers are designated by black dots. The figure was generated in R v3.6.3 (https://www.r-project.org) using ggplot2 v3.3.0 (https://ggplot2.tidyverse.org), both freely available.

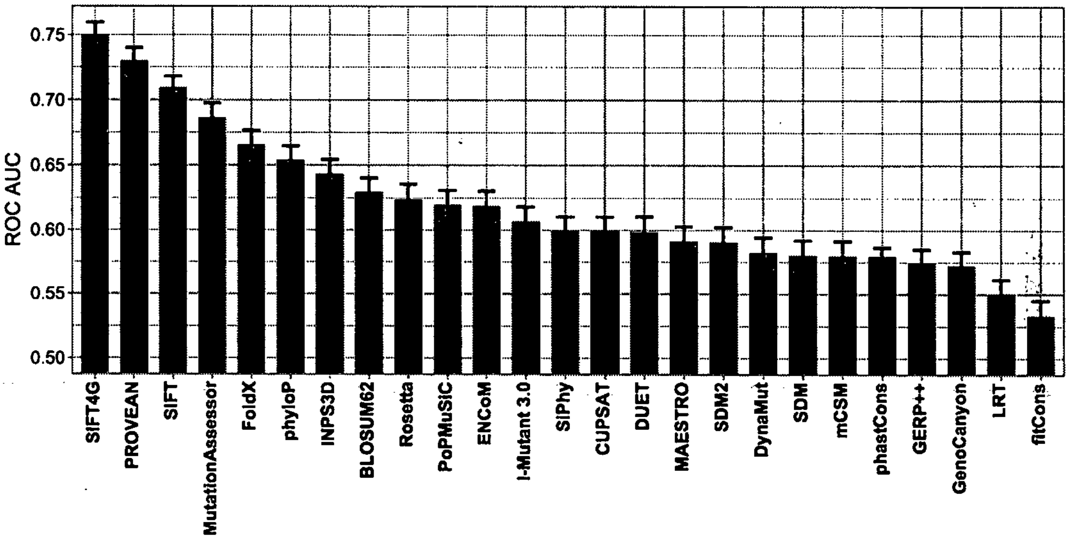

Figure 3. Performance comparison of protein stability and variant effect predictors for identifying pathogenic variants. Error bars indicate the 95% confidence interval of the ROC AUC as derived through bootstrapping. Stability predictors are shown in red, while other variant effect prediction methods are shown in green. Absolute ΔΔG values were used for stability-based methods. The figure was generated in R v3.6.3 (https://www.r-project.org) using ggplot2 v3.3.0 (https://ggplot2.tidyverse.org), both freely available.

Discussion

The first purpose of this study was to compare the abilities of different computational stability to distinguish between known pathogenic missense mutations and other putatively benign variants observed in the human population. In this regard, FoldX is the winner, clearly outperforming the other ΔΔG prediction tools. It also has the advantage of being computationally undemanding, fairly easy to run, and flexible in its utilisation. Compared to other methods that employ physics-based terms, FoldX introduces a few unique energy terms into its potential, notably the theoretically derived entropy costs for fixing backbone and side chain positions[59]. However, the main reason behind its success is likely the parametrisation of the scoring function, resulting from the well optimised design of the training and validation mutant sets, which aimed to cover all possible residue structural environments[60]. Interestingly, while the form of the FoldX function, consisting of mostly physics-based energy terms, has not seen much change over the years, newer knowledge-based methods, which leverage statistics derived from the abundant sequence and structure information, demonstrate poorer and highly varied performance. However, it is important to emphasise that the performance of FoldX does not necessarily mean that it is the best predictor of experimental ΔΔG values or true (de)stabilisation, as that is not what we are testing here. We also note the strong performance of INPS3D, which ranked a clear second in all tests. It has the advantage of being available as a webserver, thus making it simple for users to test small numbers of mutations without installing any software.

There are two factors likely to be contributing to the improvement in the identification of pathogenic mutations using absolute ΔΔG values. First, while most focus in the past has been on destabilising mutations, some pathogenic missense mutations are known to stabilise protein structure. As an example, the H101Q variant of chloride intracellular channel 2 (CLIC2) protein, which is thought to play a role in calcium ion signalling, leads to developmental disabilities, increased risk to epilepsy and heart failure[61]. The CLIC2 protein is soluble, but requires insertion into the membrane for its function, with a flexible loop connecting its domains being functionally implicated in a necessary conformational rearrangement. The histidine to glutamine substitution, which occurs in the flexible loop, was predicted to have an overall stabilising energetic effect due to conservation of weak hydrogen bonding, but also the removal of charge that the protonated histidine exerted on the structure[61]. The ΔΔG predictions were followed up by molecular dynamics simulations, which supported the previous conclusions by showing reduced flexibility and movement of the N-terminus, with functional assays also revealing reduced membrane integration of the CLIC2 protein in line with the rigidification hypothesis[62]. However, other interesting examples of negative effects of over-stabilisation exist in enzymes and protein complexes, manifesting through the activity-stability trade-off, rigidification of co-operative subunit movements, dysregulation of protein–protein interactions, and turnover[49,50,63].

In addition, it may be that some predictors are not as good at predicting the direction of the change in stability upon mutation. That is, they can predict structural perturbations that will be reflected in the magnitude of the ΔΔG value, but are less accurate in their prediction of whether this will be stabilising or destabilisng. For example, ENCoM and DynaMut predict nearly half of pathogenic missense mutations to be stabilising (41% and 44%, respectively), whereas FoldX predicts only 13%. While FoldX, Rosetta and PoPMuSiC are all driven by scoring functions consisting of a linear combination of physics- and statistics-based energy terms, ENCoM is based on normal mode analysis, and relates the assessed entropy changes around equilibrium upon mutation to the state of free energy. DynaMut, a consensus method, integrates the output from ENCoM and several other predictors (Table 1) into its score[48]. The creators of ENCoM found that their method is less biased at predicting stabilising mutations[64]. From our analysis, we are unable to confidently say anything about what proportion of pathogenic mutations are stabilising versus destabilising, or about which methods are better at predicting the direction of stability change, but this is clearly an issue that needs more attention in the future.

The second purpose of our study was to try to understand how useful protein stability predictors are for the identification of pathogenic missense mutations. Here, the answer is less clear. While all methods show some ability to discriminate between pathogenic and putatively benign variants, it is notable and perhaps surprising that all methods except FoldX and INPS3D performed worse than the simple BLOSUM62 substitution matrix, which suggests that these methods may be relatively limited utility for variant prioritisation. Even FoldX was unequivocally inferior to multiple variant effect predictors, suggesting that it should not be relied upon by itself for the identification of disease mutations.

One reason for the limited success of stability predictors in the identification of disease mutations is that predictions of ΔΔG values are still far from perfect. For example, a number of studies have compared ΔΔG predictors, showing heterogeneous correlations with experimental values on the order of $R = 0.5$ for many predictors[12,13,65]. However, a recent work has also revealed problems with the noise in experimental stability data used to benchmark the prediction methods, generally assessed through correlation values[66]. Taking noise and data distribution limitations into account, it is estimated that with currently available experimental data the best ΔΔG predictor output correlations should be in the range 0.7–0.8, while higher values would suggest overfitting[66]. As such, even assuming that 'true' ΔΔG values were perfectly correlated with mutation pathogenicity, we would still expect these computational predictors to misclassify many variants.

The existence of alternate molecular mechanisms underlying pathogenic missense mutations is also likely to be a major contributor to the underperformance of stability predictors compared to other variant effect predictors. At the simplest level, our analysis does not consider intermolecular interactions. Thus, given that pathogenic mutations are known to often occur at protein interfaces and disrupt interactions[67,68], the stability predictors would not be likely to identify these mutations in this study. We tried to minimise the effects of this by only considering crystal structures of monomeric proteins, but the existence of a monomeric crystal structure does not mean that a protein does not participate in interactions. Fortunately, FoldX can be easily applied to protein complex structures, so the effects of mutations on complex stability can be assessed.

Pathogenic mutations that act via other mechanisms may also be missed by stability predictors. For example, we have previously shown that dominant-negative mutations in ITPR1[69] and gain-of-function mutations in PAX6[70] tend to be mild at a protein structural level. This is consistent with the simple fact that highly destabilising mutations would not be compatible with dominant-negative or gain-of-function mechanisms. Similarly, hypomorphic mutations that cause only a partial loss of function are also likely to be less disruptive to protein structure than complete loss-of-function missense mutations[71].

These varying molecular mechanisms are all likely to be related to the large heterogeneity in predictions we observe for different proteins in Fig. 2. Similarly, the specific molecular and cellular contexts of different proteins could also limit the utility of ΔΔG values for predicting disease mutation. For example, even weak perturbations in haploinsufficient proteins could lead to a deleterious phenotype. At the same time, intrinsically stable proteins, proteins that are overabundant or functionally redundant could tolerate perturbing variants without such high

Pet. Reh. App.91

$\Delta\Delta G$ variants being associated with disease. Finally, in some cases, mildly destabilising mutations can unfold local regions, leading to proteasome mediated degradation of the whole protein[34,36,72].

There could be considerable room for improvement in $\Delta\Delta G$ predictors and their applicability to disease mutation identification. Recently emerged hybrid methods, such as VIPUR[73] and SNPMuSiC[74], show promise of moving in the right direction, as they assess protein stability changes upon mutation while attempting to increase the interpretability and accuracy by taking the molecular and cellular contexts into account. However, none of the mentioned hybrid methods employ FoldX, which, given our findings here, may be a good strategy. Rosetta is also promising due to its tremendous benefit demonstrated in protein design. It should be noted that the protocol used for Rosetta in our work utilised rigid backbone parameters, due to the computation costs and time constraints involved in allowing backbone flexibility. An accuracy-oriented Rosetta protocol, or the "cartesian_ddg" application in the Rosetta suite, which allows structure energy minimisation in Cartesian space, may lead to better performance[37,75].

The ambiguity of the relationship between protein stability and function is exacerbated by the biases of the various stability prediction methods, which arise in their training, like overrepresentation of destabilising variants, dependence on crystal resolution and residue replacement asymmetry. Having observed protein-specific performance heterogeneity, we suggest that in the future focus could be shifted to identifying functional and structural properties of proteins, which could be most amenable to structure and stability-based prediction of mutation effects. Additionally, a recent work has showcased the use of homology models in structural analysis of missense mutation effects associated with disease, demonstrating utility that rivals experimentally derived structures, and thus expanding the possible resource pool that could be taken advantage of for structure-based disease prediction methods[30]. Further, our disease-associated mutations set likely contains variants causing disease through other mechanisms, that do not manifest through strong perturbation of the structure, making accurate evaluation impossible. To allow better stability-based predictors, it is important to have robust annotation of putative variant mechanisms, which is currently lacking due to non-existent experimental characterisation. We hope our results encourage new hybrid approaches, which make full use of the best available tools and resources to increase our ability to accurately prioritise putative disease mutations for further study, and elucidate the relationship between disease and stability changes.

Methods

Pathogenic and likely pathogenic missense mutations were downloaded from the ClinVar[2] database on 2019-04-17, while putatively benign variants were taken from gnomAD v2.1[1]. Any ClinVar mutations were excluded from the gnomAD set. We searched for human protein-coding genes with at least 10 ClinVar mutations occurring at residues present in a single high-resolution (< 2 Å) crystal structure of a protein that is monomeric in its first biological assembly in the Protein Data Bank. We excluded non-monomeric structures due to the fact that several of the computational predictors can only take a single polypeptide chain into consideration.

FoldX 5.0[76] was run locally using default settings. Importantly, the 'RepairPDB' option was first used to repair all structures. Ten replicates were performed for each mutation to calculate the mean.

The Rosetta suite (2019.14.60699 release build) was tested on structures first pre-minimised using the minimize_with_cst application and the following flags: -in:file:fullatom; -ignore_unrecognized_res -fa_max_dis 9.0; -ddg::harmonic_ca_tether 0.5; -ddg::constraint_weight 1.0; -ddg::sc_min_only false. The ddg_monomer application was run according to a rigid backbone protocol with the following argument flags: -in:file:fullatom; -ddg:weight_file ref2015_soft; -ddg::iterations 50; -ddg::local_opt_only false; -ddg::min_cst false; -ddg::min true; -ddg::ramp_repulsive true ;-ignore_unrecognized_res.

Predictions by ENCoM, DUET and SDM were extracted from the DynaMut results page, as it runs them as parts of its own scoring protocol. mCSM values from DynaMut coincided perfectly with values from the separate mCSM web server, and thus the server values were used, as DynaMut calculations yielded less results due to failing on more proteins.

All other stability predictors were accessed through their online webservers with default settings by employing the Python RoboBrowser web scrapping library. Variant effect predictors were run in the same way as described in our recent benchmarking study[10].

Method performance was analysed in R using the PRROC[77] and pROC[78] packages, and AUC curve differences were statistically assessed through 10,000 bootstraps using the roc.test function of pROC. For DynaMut, I-Mutant 3.0, mCSM, SDM, SDM2 and DUET, the sign of the predicted stability score was inverted to match the convention of increased stability being denoted by a negative change in energy. For the precision-recall analysis, we used a subset of the mutation dataset, containing 9,498 ClinVar and gnomAD variants, which had no missing prediction values for any of the stability-based methods. This is because a few of the predictors were unable to give predictions for all mutations (e.g. they crashed on certain structures), and for the precision-recall analysis, it is crucial that all predictors are tested on exactly the same dataset. We also show that the relative performance of the top predictors remains the same in the ROC analysis using this smaller dataset (Table S1).

All mutations and corresponding structures and predictions are provided in Table S2.

Received: 11 June 2020; Accepted: 31 August 2020
Published online: 21 September 2020

References

1. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581
(2020).