Predicting Outcomes of Interventions to Increase Social Competence in Children and Adolescents

A Dissertation Presented

by

Kodi Benjamin Arfer

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Experimental Psychology

Stony Brook University

May 2016

Stony Brook University

The Graduate School

Kodi Benjamin Arfer

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation.

Christian C. Luhmann – Dissertation Advisor
Associate Professor, Psychology

Richard J. Gerrig – Chairperson of Defense
Professor, Psychology

Matthew D. Lerner
Assistant Professor, Psychology

Frederick Shic
Assistant Professor, Computer Science, Yale University

This dissertation is accepted by the Graduate School

Charles Taber
Dean of the Graduate School

Abstract of the Dissertation

Predicting Outcomes of Interventions to Increase Social Competence in Children and Adolescents

by

Kodi Benjamin Arfer

Doctor of Philosophy

in

Experimental Psychology

Stony Brook University

2016

Autism is a common condition with often debilitating symptoms, and both the methods and results of behavioral treatments vary widely. It remains largely unknown which patients will respond best to which treatments, which is especially problematic because treatments are time-consuming and expensive. The present study is the first to use predictive data analysis to examine how the outcomes of behavioral interventions targeted at social competence can be statistically predicted from pretreatment measures. The study used five previously collected datasets, including patients aged 5 to 18, and including treatments such as skillstreaming, a Second Step program, and socio-dramatic affective-relational intervention. However, results indicated that by and large, pretreatment measures (other than the same instrument used for the outcome variable) were not predictive of outcomes. Follow-up analyses simulating the effects of treatment on broad populations weakly indicated that socio-dramatic affective-relational intervention would increase externalizing behavior overall, but also slightly increase self-control. Other kinds of pretreatment measures may be necessary to accurately predict treatment outcomes.

Table of Contents

Introduction

This study examined how treatment outcomes for autism can be predicted on the basis of pretreatment variables. Below, there is first described the characteristic features of autism, its significance for society, and what is known about its treatment, particularly how to decide between the many extant behavioral treatments. The section after that discusses predictive data analysis as an approach to addressing the treatment problem, how it differs from more familiar practices of data analysis, and how it can be expanded to an alternative approach to science in general. The introduction is concluded with a summary of this study's methods.

Autism and its treatment

Terms such as "autism", "autism spectrum disorder", "Asperger syndrome", and "pervasive developmental disorder" describe a variety of atypical patterns of behavior characterized primarily by atypical social interaction and secondarily by stereotyped interests, beginning in early childhood and remaining for life (Lai, Lombardo, & Baron-Cohen, 2014). This paper will refer to these conditions collectively as "autism" for short, although many writers distinguish autism per se from other conditions such as Asperger syndrome. About 1 in 160 people worldwide meet criteria for an autism diagnosis (Elsabbagh et al., 2012), comparable to the lifetime prevalence of schizophrenia of about 1 in 140 (McGrath, Saha, Chant, & Welham, 2008). The rate among 8-year-old children in the US is about 1 in 68 (Autism and Developmental Disabilities Monitoring Network, 2014). Diagnoses of autism have increased an order of magnitude since the 1980s (Cavagnaro, 2009), tracking a broadening of diagnostic criteria (Fisch, 2012). The social burden of autism is indicated by such findings as a 2.8 times increased death rate (Woolfenden, Sarkozy, Ridley, Coory, & Williams, 2012) and an estimated cost of $268 billion in the United States in 2015 (Leigh & Du, 2015). Autism and its treatment are regarded as a pressing public health issue.

Behavioral interventions of various kinds are widely used to treat autism, especially in children. (This paper will use the word "children" to refer to adolescents as well as prepubescents.) On the basis of a systematic research review, a panel of experts (Maglione, Gans, Das, Timbie, & Kasari, 2012) voted that there was low to moderate evidence in favor of the effectiveness of therapies from applied behavior analysis to social-skills training. The panel judged there was not enough evidence to compare the efficacies of the different therapies. Similarly, a review of reviews by Seida et al. (2009) concluded that "some form of treatment is favorable over no treatment. However, there is little evidence for the relative effectiveness of these treatment options." (p. 95) Wong et al. (2015) identified 27 different interventions as meeting criteria for evidence-based practice. Importantly, beneficial outcomes of early intervention seem to be maintained for years afterwards (e.g., McEachin, Smith, & Lovaas, 1993; Sallows & Graupner, 2005; Estes et al., 2015).

This paper will be concerned chiefly with the social symptoms of autism, especially in

verbal, high-functioning people, without severe comorbidity or very low intelligence. Broadly, these symptoms can be described as abnormal or missing social engagement, particularly with other people of the same age. Specifically, autistic people tend not to show interest in others, reciprocate the social actions or emotions of others, react to implicit social cues, or interpret or produce nonverbal communicative behavior such as facial expressions and gestures (Attwood, 2000; Otero, Schatz, Merrill, & Bellini, 2015). These deficits seem to stem at least partly from lack of ability rather than mere lack of interest, since in laboratory tests, autistic people show reduced ability to read facial expressions (Baron-Cohen, Wheelwright, & Jolliffe, 1997) or understand non-literal uses of language such as irony (Kaland et al., 2002). As one might expect, a prominent class of treatment for social deficits is social-skills training, in which an instructor teaches everyday social skills such as greeting and making eye contact, often to adolescent or adult patients in small groups (Lerner, White, & McPartland, 2012). Most social-skills training programs include direct instruction, role-playing or other forms of practice, and feedback (Wong et al., 2015). The research base on the effectiveness of social-skills training is still developing, but is considered promising (Miller, Vernon, Wu, & Russo, 2014; Soorya et al., 2014).

A common strategy for behavioral intervention that complicates evaluation is individualized treatment. Clinicians often attempt to adapt autism treatment to the particular strengths and needs of patients. The panel of Maglione et al. (2012) endorsed this practice, and Seida et al. (2009) observed that so did many other research reviews. The need for individualized treatment is suggested by the finding that, as with nearly all treatments for mental disorder, the outcomes of a single treatment vary widely between patients (e.g., Smith, Groen, & Wynn, 2000; Dawson et al., 2010). And yet just as extant research provides little guidance to choosing among the many types of behavioral interventions, it provides little information as to how treatment should differ based on the patient's characteristics.

Many writers have offered suggestions on the basis of qualitative research or case studies as to how treatment could be individualized. For example, Choque Olsson, Rautio, Asztalos, Stoetzer, and Bölte (2016) mention that it could be useful to focus on group cohesion in social-skills training groups for adolescents, whereas snack-time and playacting seem more useful for younger children. Bottema-Beutel, Mullins, Harvey, Gustafson, and Carter (2016) suggest adjusting how much a peer group is told about a patient's diagnosis or symptoms according to the patient's own preferences. And Smith and Sharp (2013) suggest having patients join social groups such as clubs chosen based on patients' personal interests and their degree and kind of sensory hypersensitivity. But these ideas have not been directly tested. More generally, qualitative research gives little indication of which variables such individualization strategies would affect, how strong the effects would be, and how sure we can be that the effects would be in the desired direction.

The practice of individualizing treatment is often taken to entail adjusting the internal features of a single treatment. For example, Stahmer, Schreibman, and Cunningham (2011), pointing out that parenting autistic children can be stressful, discuss the possibility of deciding on the basis of a parent's level of stress whether a given treatment should be administered by the parent or by a clinician. A coarser approach to individualization, which is more closely related to the present study, is to decide which of several treatments (including no treatment) to provide to each person. For example, Stahmer et al. suggest that the Picture Exchange Communication System, a visually focused treatment, may be more effective for children with low ability to initiate joint attention than Responsive Education and Prelinguistic Milieu Teaching, a vocally focused treatment (Yoder & Stone, 2006).

The question of how to treat a given patient takes on additional importance when we consider the costs of treatment. There is a broad consensus that treatment should be intensive, with the panel of Maglione et al. (2012) suggesting that autistic children receive a minimum of 25 hours a week of comprehensive intervention. Multiplying this by the growing number of autism diagnoses implies that providing all recommended treatment would be extraordinarily expensive, in terms of time as well as money. Treatment can still be cost-effective considering that the symptoms of autism can cost society millions of dollars per person (Ganz, 2007; Järbrink & Knapp, 2001). What would be helpful is to be able to predict treatment outcomes, and thus assess cost-effectiveness, on an individual basis.

Predictive data analysis

Making treatment decisions on an individual basis is a good candidate for the use of predictive data analysis. With predictive methods, one can estimate variables that are difficult to measure directly, or that are not measurable at all because they involve future events. For example, we would like to predict the outcome of an autism treatment on a given child before actually administering the treatment, so we can decide which treatment is best, how intensive the treatment should be, when the treatment should begin and end, whether treatment should be administered at all (if the symptoms are mild or spontaneous recovery is likely), and so on. In short, we want to make good treatment decisions.

Prediction can be useful, but it involves many subtle difficulties. One of the first difficulties concerns ambiguity in the word "prediction". Typically, when clinical researchers speak of prediction, they say that a theory predicts an overall finding regarding the ordering of variables. For example, the extreme male brain theory of autism predicts that boys are more likely to show autistic traits than girls (Baron-Cohen, 2002). In statistics, however, "prediction" refers to estimating individual values. Naturally, this is generally accomplished with quantitative models. For example, a model might use a child's gender to estimate the child's score on a test of autism symptoms. Notice the change from making assertions about what is greater than what to estimating actual values. There is also a change from predicting overall trends to predicting individual cases; this makes little difference when the only input to a model is gender, but becomes important once many different inputs are available (say, age, family history, and blood levels of a hormone) and each case's combination of inputs is unique, meaning the model must integrate all the data it is given to maximize the accuracy of its predictions. Although prediction was originally the predominant focus of statistics, the rise of mathematical statistics led to the modern concern with testing models and estimating parameters (Geisser, 1993). From a predictive perspective, models and parameters are only a means to an end; the focus is on the dependent variable (DV) and how accurately it can be estimated (Arfer & Luhmann, 2015). In modern times, interest in prediction is keenest in the field of machine learning (e.g., Hastie, Tibshirani, & Friedman, 2009).

This may sound like a reaffirmation of the value of much of the autism research that has already has been conducted and is being conducted. However, research in autism, as elsewhere in psychology, tends not to do what would be needed to make good predictions and to tell us how good our predictions will be, even when the investigators themselves use the word "predict". The chief problem is when researchers train and test models with the same data; that is, use the same cases both to estimate model parameters (training) and to evaluate how well the model can predict the DV (testing). This is the statistical equivalent of asking a magician to guess what card

you're holding after he's already seen it, and leads to overfitting, and thus to an overly optimistic estimate of predictive accuracy (e.g., Wasserman, 2004, Theorem 13.15; Myung, 2000; Steyerberg et al., 2001; Hitchcock & Sober, 2004). An evaluation of this kind produces what may be called *association* (because it quantifies how closely one variable can be associated with other variables) or *training error* (that is, the error in fitting the model to its own training data). Examining association can be appropriate and useful when testing theories or investigating how a phenomenon can be explained or summarized, but association is distinct from predictive accuracy.

Consider previous applications of predictive methods to autism, specifically, to predict the diagnosis that would be made with a comprehensive procedure using only an abbreviated version of the procedure. This was the goal of Wall, Kosmicki, DeLuca, Harstad, and Fusaro (2012b) and Wall, Dally, Luyster, Jung, and DeLuca (2012a). Wall et al. (2012b) used a dataset of 627 people to reduce the 29 codes of Module 1 of the Autism Diagnostic Observation Schedule (ADOS). They found that alternating decision trees (Freund & Mason, 1999), a streamlined alternative to boosted sets of classification trees, could classify subjects into one of two ADOS diagnoses with accuracy near 100%, using only 8 of 29 codes. Wall et al. (2012a) used a dataset of 966 people assessed with the Autism Diagnostic Interview: Revised and again found that alternating decision trees could predict which of two diagnosis classes subjects belonged to with accuracy near 100%, this time using only 7 of 93 items. However, Bone et al. (2015) pointed out major conceptual and analytic issues with the Wall et al. studies. Most saliently, Wall et al. used extremely imbalanced data, in which only a few subjects were not diagnosed with autism (e.g., in Wall et al., 2012b, the training set had 891 positive cases and 75 negative cases [8% negative], and the two test sets together had 1,976 positive cases and 17 negative cases [less than 1% negative]). This means that a trivial classifier, predicting a positive diagnosis for every subject, could have achieved accuracy comparable to that of the alternating decision trees. Bone et al. went on to fail to replicate Wall et al.'s results for both instruments when using improved analytic methods.

Macari et al. (2012) and Chawarska et al. (2014) considered another predictive problem in autism: using measurements of children at one age to predict diagnoses at a later age. Macari et al. (2012) used classification trees to predict autism diagnoses at age 2 using the ADOS administered at age 1. They achieved an accuracy of 81%, a substantial improvement over the base rate of 60%. However, this accuracy was the best result of 14 separate cross-validation loops with different tree sizes, so it may be inflated by overfitting; accuracy on a new dataset, for which an optimal tree size has not been chosen, would likely be worse. Chawarska et al. (2014) used classification trees to predict autism diagnoses at age 3 using the ADOS administered at age 1½. They avoided repeating the mistake of Macari et al. (2012) by estimating accuracy on a separate validation set after choosing the tree size. The accuracy came out to 77.3%, identical to the base rate in the validation set, meaning that the fitted tree was no more accurate than simply predicting that no subjects at all would be diagnosed with autism.

The aforementioned studies represent the state of the art for predictive data analysis in the study of autism using data from traditional psychometric methods. Bone et al. (2015) conducted new analyses to check the claims of the Wall et al. studies, but did not succeed in finding a model that could accurately predict diagnoses using only a few items, as Wall et al. had hoped. Similarly, Chawarska et al. (2014) improved on the methods of Macari et al. (2012), but achieved accuracy no better than baseline.

More success has been achieved using high technology such as eye tracking and brain

imaging. For example, Campbell, Shic, Macari, and Chawarska (2014) clustered 20-month-old autistic children on the basis of eye tracking while watching videos. Subjects in clusters with greater attention to the depicted scene and the speaker's mouth were more likely to be verbal and high-functioning at age 3. Just, Cherkassky, Buchweitz, Keller, and Mitchell (2014) had autistic adults and matched controls think about 16 social scenarios under functional magnetic resonance imaging (fMRI). Naive Bayes classifiers assessed in leave-one-out cross-validation could correctly identify all but one subject's group on the basis of the fMRI data. Finally, Crippa et al. (2015) collected kinematic data from autistic children ages 2 and 4 and an equal number of controls while they reached for an object, picked it up, and dropped it. A support vector machine obtained leave-one-out classification accuracy of at least 75% for any number of features included in the model. However, notice that, like the aforementioned studies using the ADOS and interviews, these studies considered the prediction of diagnoses and symptoms exclusively. The present study is the first to focus on predicting treatment outcomes on the basis of pretreatment measures.

What, then, is the right way to estimate predictive accuracy? The key idea is to check what value a model produces for a case when it has *not* been trained using that very case. The simplest way to do this is to train the model on one group of cases and look at its predictions for a different group of cases. This is the technique used by Chawarska et al. (2014), who chose 20% of their sample to hold out as a validation set: the model was trained on the remaining 80%, then made predictions for the validation set. Another approach is cross-validation, which uses the same data for both training and testing, but not at the same time. Instead, the data is randomly partitioned into a number of subsets called folds, often 5 or 10 of them, and then each fold in turn is treated as a test set while the rest of the data is used for training, with the model-fitting being carried out independently each time. Cross-validation effectively gives one a much larger test set than setting aside a single validation set, although it is computationally slower and estimates a subtly different notion of test error (see Hastie et al., 2009, p. 254). Importantly, if one uses cross-validation to set a tuning parameter, as Macari et al. (2012) and Chawarska et al. (2014) did to choose tree size, one must use an outer round of cross-validation or a separate validation set to avoid having one's estimate of predictive accuracy be optimistically biased from any overfitting of the tuning parameter.

To be sure, there have been success stories in predictive data analysis, where the optimistic bias of overfitting was avoided and good accuracy was achieved. For example, Li, Wileyto, and Heitjan (2011) attempted to predict whether subjects still smoked 1 year after a 10-week smoking-cessation program. A logistic-regression model using sex, treatment condition, and a questionnaire test of nicotine dependence achieved a (bootstrap-corrected) area under the curve (AUC) of .50, which is no better than that of blind guessing. But when the authors added to this model a variable indicating whether the subject had been smoking at the end of treatment, the AUC increased to .74. Zhang, Wang, Zhou, Yuan, and Shen (2011) used variables from brain imaging and analysis of cerebrospinal fluid to distinguish people with Alzheimer's disease and mild cognitive impairment from controls. A support vector machine achieved a cross-validated accuracy of 93% for distinguishing Alzheimer's-affected from controls in a sample with a base rate of 50%. Some other examples of successful uses of neuroimaging and predictive methods to predict treatment response for mental disorder include Gong et al. (2011), which predicted response to antidepressants; Costafreda, Khanna, Mourao-Miranda, and Fu (2009), which predicted response to cognitive-behavioral therapy for depression; and Ball, Stein, Ramsawh, Campbell-Sills, and Paulus (2014), which predicted response to cognitive-behavioral therapy for

generalized anxiety disorder and panic disorder.

More broadly, predictive data analysis is not just the correct method to answer a special class of questions in applied scientific research, namely, where there is a preexisting interest in estimating what subjects will do ahead of time. It also represents an alternative, mostly untapped approach to science in general, especially when the object of study is an extremely large and complicated system, such as human behavior, as opposed to, say, a pendulum (see also the discussion in Arfer & Luhmann, 2015, and Arfer & Luhmann, 2016). A predictive approach to science suggests, instead of the usual goals of testing theories and elucidating internal mechanisms, building accurate predictive algorithms. There is an increased emphasis on observables, in preference to theoretical constructs. The focus changes from understanding how things work to understanding what things do, from the internal structure of an entity to its effects on observable features of the world.

Predictivism, as we will call this philosophy, is motivated by two basic ideas. One, an understanding of internal mechanisms may be neither necessary nor sufficient for prediction. In particular, the models or theories that are most tenable as the explanation of a system—that best describe the data-generating process—may not be the best at predicting the system's outputs. Even a model with known-false assumptions may be more predictively accurate than a more realistic model (Domingos & Pazzani, 1997). Intuitively, this should be more likely for systems with more internal complexity. Two, there is value in pursuing the sort of understanding that constitutes being able to, as Watson (1913) said, "predict and control" a system, not to be able to accurately explain how it works. If we can predict and control the behavior of organisms, but do not know how their brains and cognitive architectures actually translate inputs to outputs, then we are in a sense wiser (and practically better off) than if we knew all about brains and cognitive architectures, but the organisms remained unpredictable and uncontrollable.

Predictivism was originally discussed in Arfer and Luhmann (2015). This study considered the question of what models to use to predict behavior in a simple laboratory task involving intertemporal choice, that is, decisions about delayed rewards. While many models had previously been proposed and evaluated for intertemporal choice, they had mostly been evaluated on the basis of association or theoretical concerns. By considering predictive accuracy instead, we found that a wide variety of models are accurate, showing that longstanding modeling controversies in intertemporal choice are in fact somewhat overblown. In Arfer and Luhmann (2016), we took the idea of prediction of intertemporal choice further by using a similar laboratory task to generate predictor variables, but rather than predicting different trials of the same task, we tried to predict self-reports of behavior in real-world domains of intertemporal choices such as saving and debt. As we had argued in Arfer and Luhmann (2015), "After all, predicting behavior in laboratory tasks is only interesting insofar as laboratory behavior relates to real-life behavior." (p. 339) Arfer and Luhmann (2016) also differed from Arfer and Luhmann (2015) in that models were fit between subjects rather than only within subjects. The results were less positive, indicating that laboratory tasks of intertemporal choice had little to no utility for predicting real-world decisions. However, the ad-hoc and entirely self-reported DVs of Arfer and Luhmann (2016), and the use of only a single kind of predictor variable, may have made it difficult to find good predictive accuracy.

The present study

The present study reanalyzed previously conducted studies of social-skills training for

autism from a predictive perspective. The intention is to contribute to practical knowledge of the treatment of autism while also advancing the predictivist program of research described above. In fact, this study is the first to examine predictively the question of treatment outcomes for autism. It goes further than Arfer and Luhmann (2016) by considering a wider variety of predictors, by using previously established psychometrically validated tests, by including a randomly assigned treatment in some cases (allowing for the investigation of the prediction of casual effects), and considering a more applied setting and research question in general.

The study considered as outcome variables several measures of possible practical interest, from overall measures of symptomatology such as the Social Responsiveness Scale to specific measures such as the tendency to make hostile attributions in hypothetical situations. It employed several kinds of statistical models and make available all the measured variables possible as inputs to them, not out of theoretical considerations but to maximize the chances of finding a predictively useful model. (The study has favored behavioral measures over neuroimaging, hormone assays, and other physiological measures because they are less expensive and easier to use and therefore are more useful when they are predictively accurate.) The final results are in terms of predicted outcomes of the various treatments and the expected accuracy of these predictions, both for actual subjects who received a different treatment and for hypothetical populations of subjects. The analyses attend to issues raised by the findings of Bone et al. (2015) and Chawarska et al. (2014), such as the need to evaluate predictive accuracy relative to appropriate baseline measures. The methods improved upon the methods of Wall et al. (2012a), Wall et al. (2012b), Bone et al. (2015), Chawarska et al. (2014), and Macari et al. (2012) by considering a variety of models, beyond tree-based learners alone.

Data sources

Studies

The data for the analyses comes from five previously conducted empirical studies by Matthew D. Lerner, who provided the data. (Another study, called Norfolk, was planned for inclusion, but it turned out that no more than 4 subjects' worth of data was available for most measures, so this dataset was excluded.) Each study examined the effect of one or more interventions to increase social competence in children who were, in general, autistic. See the next section for full names and descriptions of the measurement instruments mentioned here. All samples were convenience samples to which various exclusion criteria, not discussed here, were applied. Instruments not used for analysis are not mentioned. Four of the the five studies examined treatments that spanned multiple sessions over the course of several weeks. In each of these studies, one of the interventions considered was socio-dramatic affective-relational intervention (SDARI; Lerner, Mikami, & Levine, 2011; Lerner & Mikami, 2012), a set of group activities that focus on practicing social skills. When a comparison intervention was present, it focused on explicit knowledge and direct instruction of social skills, rather than practice. Subjects participated in these interventions in groups, not always all in one group, and along with other autistic children of similar age who were not subjects of the same study. The fifth study, Knowledge or Performance, compared the effect of just 20 minutes of training between two types of training.

The study Spotlight 2007 (Lerner et al., 2011) examined children ages 11 to 17 who had been diagnosed with high-functioning autism by a physician, or had such a diagnosis included in an Individualized Education Program (IEP; 34 CFR § 300.320), a document that describes the educational needs of a disabled student. Of 17 subjects, 9 received SDARI during the study period and 8 did not, 4 of whom had received SDARI at an earlier time; these conditions were not randomly assigned. Before the study period, parents completed the DHF (from which, in the case of this study, only the genders and ages of subjects were available). At 7 measurement sessions at 3-week intervals, the DANVA-2, BDI-Y, EDI, SRS, and SSRS-P were administered. The CBCL was administered every other session. Treatment was administered to the treatment group in the middle 6 weeks of this 18-week span. Hence, for most instruments, there were 3 pretreatment administrations, 1 mid-treatment administration, and 3 posttreatment administrations, whereas for the CBCL, there were 2 pretreatment and 2 posttreatment administrations.

The studies Spotlight 2008 and Spotlight 2010 examined children ages 9 to 18 who were already enrolled in a summer program for improvement of social competence, and therefore had some kind of social deficit. All subjects participated in SDARI, with no comparison group. For this paper, data was obtained for 9 subjects from the 2008 wave and 30 subjects from the 2010 wave, not counting 12 subjects (9 from 2008, 3 from 2010) who were missing on all posttreatment measures, and 1 subject from 2010 who was missing on most pretreatment measures. Before treatment, parents completed the DHF. Before and after treatment, the SAS,

SRS, HAQ, SSRS-S, and SSRS-P were administered. There were 6 weeks of treatment.

The study Charlottesville examined children ages 9 to 14 diagnosed with Asperger's syndrome or high-functioning autism. Of the 13 subjects, 7 were randomly assigned to receive SDARI and 6 received skillstreaming. Skillstreaming is a treatment that differs from SDARI in that it focuses on the direct instruction of social skills, ignorance of which is hypothesized to be the cause of social deficits in autism (Goldstein & McGinnis, 1997). Before treatment, parents completed the DHF and SCQ. Before and after treatment, the SAS, SRS, HAQ, SSRS-S, SSRS-P, and PCToMM-E were administered. Treatment for both conditions comprised 5 weekly hour-and-a-half sessions. All subjects in each condition were in the same therapy group.

The study Knowledge or Performance examined children ages 9 to 17 with a diagnosis of high-functioning autism listed in their IEPs or reported by their parents. The 40 subjects were split into 10 groups of 4 by age, gender, and intelligence; each group of 4 was split into two dyads; and each dyad was randomly assigned to one of two training conditions, knowledge training or performance training, such that each group of 4 had one dyad in each condition. Before any dyadic interaction, the DHF, ADOS (module 3 or 4), DANVA-2, CABS, DMQ, SCQ, SCT, SEL, and WISC were administered. (If subjects consented, they completed the DANVA-2 under electroencephalographic recording. However, electroencephalographic data was not analyzed in this study.) Then the subjects in each dyad met. They were given 10 minutes to interact freely, then received 20 minutes of training, then were given another 10 minutes of free interaction. In the knowledge condition, the trainer discussed basic social skills such as greeting and expressing emotions, and solicited answers to questions about hypothetical situations, but did not solicit any practice. In the performance condition, the trainer had subjects play social games such as group storytelling, but provided no instruction on social skills themselves. Behavior during the free-interaction periods was coded with the SIOS. Finally, subjects completed the DANVA-2 (without electroencephalography) and the SEL again.

Measures

Here is described each instrument present in the five datasets and how these instruments were used to create the specific variables used for analyses. They are in alphabetical order except that the basic demographic instrument is first. Most instruments, administered before treatment, were used only to create independent variables (IVs; the term is meant in the statistical sense, meaning "predictors", not in the experimental sense, meaning "manipulated variables"). The variables used for DVs are listed below in the section "DVs" for convenience.

Guardians of all subjects completed a developmental history form (DHF). This form included a variety of questions about the subject as well as the subject's parents, which were used to create IVs as follows:

- Gender was binary-coded.
- Age, in years, was left as-is.
- Race was ignored because 86 of 95 non-missing values were "Caucasian".
- Likewise, an item concerning the kind of guardian filling out the questionnaire was ignored, because 91 of 104 non-missing values were "Biological mother".
- Number of siblings was coded into two binary variables: whether the subject had exactly one sibling, and whether the subject had more than one sibling.
- Parent education was coded into a binary variable indicating whether at least one parent had an advanced graduate or professional degree.

- Parent employment status was coded into a binary variable indicating whether at least one parent was a stay-at-home parent.
- Household income was left in its original response format, which assigned "Less than $10,000" to 0, "$10,000 to $20,000" to 1, "$21,000 to 30,000" to 2, and so on up to "$91,000 to 100,000" to 9, then "$101,000 to 150,000" to 10, and "More than $150,000" to 11.
- Parent relationship status coded into a binary variable indicating whether the subject's parents were together.
- School type was coded into a binary variable indicating whether the subject attended public school.
- Number of medications taken by the subject was capped at 5, because of two apparent outliers at 8 and 16.
- Number of previous interventions received by the subject was used as-is.
- Family history of developmental disorders was coded into two binary variables: one indicating whether a biological parent or (half-)sibling was affected, and one indicating whether a blood relative other than those was affected.
- Diagnoses received by the subject were coded into 7 binary variables, which respectively indicated diagnoses of a learning disorder, an anxiety disorder, attention-deficit disorder, autism per se, Asperger syndrome, a pervasive developmental disorder not otherwise specified by the DSM (PDD-NOS), and anything else.

The Autism Diagnostic Observation Schedule (ADOS; Lord et al., 2000), 1st edition, is an interview procedure used primarily for diagnosing autism. It comprises four different modules (versions) to suit different ages and levels of language ability. The procedure takes about half an hour. Depending on the module, subjects may be asked to play with toys, tell a story, or answer questions about their emotions and social lives. Observations on several dimensions of behavior, such as "unusual eye contact" and "emphatic or emotional gestures", are coded from 0 to 3 indicating the degree of evidence of autism-related abnormality, and these codes are combined to form integer ratings on three subscales: communication, social skills, and stereotyped behavior and repetitive interests. The present study used these subscale scores as IVs.

The Beck Depression Inventory for Youth (BDI-Y) is a self-report measure of depression (Stapleton, Sander, & Stark, 2007). It includes 20 items, rated on 4-point scales from *never* to *always*, regarding negative views of oneself and the world, hopelessness, and physiological symptoms of depression. The respondent's score is simply the sum of the item responses. To reduce skew, the present study added 1 to scores and then took the square root.

The Children's Assertive Behavior Scale (CABS; Scanlon & Ollendick, 1985) is a self-report measure of the tendency to be assertive; that is, to be neither excessively passive and submissive nor excessively aggressive in social interactions. Each of the 27 items describes a situation such as "You feel insulted by something someone said to you" and asks what the respondent would usually do. The 5 response choices for each item range from very passive (coded as −2) to very aggressive (coded as 2). The sum of the absolute values of response codes measures unassertiveness (with lower score indicating higher assertiveness), and the negative and positive responses can be summed separately to yield indices of passivity and aggression. The present study used the passivity and aggression scores as IVs, dichotomizing aggression as zero versus nonzero.

The Child Behavior Checklist (CBCL) is an informant questionnaire (with all informants in our data being guardians) covering a wide spectrum of mental disorder (Ivanova et al., 2007).

There are 119 items, each describing a concrete symptom, which, contrary to the use of the term "checklist", are rated from 0 (*not true*) to 2 (*very true or often true*). The items are grouped into 8 syndromes: Anxious/Depressed, Withdrawn/Depressed, Somatic Complaints, Social Problems, Thought Problems, Attention Problems, Rule-Breaking Behavior, and Aggressive Behavior. The present study used the sum for each syndrome as an IV.

The Diagnostic Analysis of Nonverbal Accuracy 2 (DANVA-2; Nowicki & Carton, 1993; Nowicki & Duke, 1994) tests children's accuracy of reading basic emotions and expressing basic emotions through facial expressions, gestures, and paralanguage (nonverbal features of speech, such as tone of voice). There are 7 subtests, which respectively test interpreting facial expressions, interpreting whole-body postures, interpreting hand and arm gestures, interpreting paralanguage, producing facial expressions, producing hand and arm gestures, and producing paralanguage. In the interpretation tests, subjects have to identify which of a few different basic emotions (e.g., anger) is being portrayed. The subject's score is the number of items correctly identified. For IVs, the present study used scores on four domains of interpretation: adult voices, child voices, adult faces, and child faces.

The Dimensions of Mastery Questionnaire (DMQ; Morgan et al., 2015) is a measure of mastery motivation, "the intrinsic drive to explore and master one's environment" (p. 3). The version used in Knowledge or Performance was completed by guardians and had 6 questions rated from 1 (*not at all typical*) to 5 (*very typical*). The present study used as an IV the total score, which is the sum of the item ratings, with the last item reversed.

The Emory Dyssemia Index-Revised (EDI; Duke & Nowicki, 2005, p. 41) is an informant-report measure of nonverbal language deficient (dyssemia). The 42 items are grouped into 7 subscales of 6 items each: Gaze and Eye Contact, Space and Touch, Paralanguage, Facial Expression, Objectics, Social Rules/Norms, and Nonverbal Reciprocity. Each item describes an unusual behavior, such as "Seems tactless" or "Clothing is not fastened correctly", and is rated from 1 (*never*) to 5 (*very often*). The subscale scores and total score are sums of item ratings. The present study used the subscale scores as IVs.

The Hostile Attribution Questionnaire (HAQ) measures how children interpret ambiguous negative social situations, such as getting milk spilled on them. (Lerner, Calhoun, Mikami, & De Los Reyes, 2012, uses a similar but not identical instrument of the same name.) For each of 4 situations, the subject rates a statement making a hostile attribution ("that kid is a mean person"), a statement that something is wrong with the subject ("you are not well-liked"), and a neutral statement ("that kid wasn't looking and didn't see you") on a scale from 1 (*strongly disagree*) to 5 (*strongly agree*). The subject also rates how angry and sad they would be from 1 (*not at all*) to 5 (*very much*) in each situation. The ratings for each statement type and feeling are averaged across the 4 situations to yield overall agreement with each statement type and overall anger and sadness. The present study used these overall scores as IVs.

The Perception of Children's Theory of Mind Measure—Experimental Version (PCToMM-E; Hutchins, Bonazinga, Prelock, & Taylor, 2008) is a guardian-report measure of children's theory of mind, that is, their capacity to infer and reason about their own mental states and the mental states of others. For each of 33 statements, the guardian indicates agreement with a statement such as "My child can communicate to me that s/he wants something" by marking a horizontal line with endpoints labeled *definitely not* and *definitely*. The total score, which the present study used as an IV, is the mean of the distances of the respondent's marks from the left endpoints, with distances scaled such that the left endpoint is at 0 and the right endpoint is at 20.

The Social Anxiety Scale (SAS) is a self-report measure of anxious feelings in social

situations. There are two versions, which differ only in wording: one for children ages 7 to 13, the Social Anxiety Scale for Children—Revised (La Greca & Stone, 1993), and another for older children, the Social Anxiety Scale—Adolescents (La Greca & Lopez, 1998). The SAS has 18 items in 3 subscales (Fear of Negative Evaluation, Social Avoidance and Distress in General, and Social Avoidance and Distress Specific to New Peers or Situations) and 4 filler items, each rated from 1 (*not at all*) to 5 (*all the time*). Subscale scores, which the present study used as IVs, are computed as sums of item ratings.

The Social Communication Questionnaire (SCQ, formerly the Autism Screening Questionnaire; Berument, Rutter, Lord, Pickles, & Bailey, 1999) is an informant-report measure of autism symptoms, such as circumscribed interests and lack of interest in other children. The 40 items are each rated as *yes* or *no* and the number of *yes* answers is counted to yield an overall severity score, which the present study used as an IV.

The Social Creativity Tasks (SCT; Mouchiroud & Lubart, 2002) measure children's ability to think of original solutions to solving social problems. The subject considers two hypothetical social problems and is asked to produce as many original ideas for solving the problem as possible. One problem ("Peers") has the subject try to convince two other children to let the subject join a game. The other ("Dyad") has the subject try to convince a friend to play a game the subject likes instead of a different one. Each subject's score for each hypothetical social problem was the sum of originality scores for each idea they mentioned, with originality scores ranging from 1 to 7 and rated by condition-blind research assistants provided with anchor responses selected by 4 clinical experts. Two raters rated each subject, so the present study used the mean of the ratings as an IV.

The Stories from Everyday Life (SEL; Kaland et al., 2002) test children's theory of mind. In Knowledge or Performance, subjects heard one story of each of 4 kinds of stories, which respectively featured figurative speech, ironic speech, mistaken intentions, and contrary emotions. Subjects answered questions testing their basic understanding of the story, their ability to make a merely physical inference regarding the story (e.g., a character has to clean a kitchen floor daily because a dog has been getting it muddy), and their understanding of the concept being tested (e.g., figurative language). Each answer was rated 0 for incorrect, 1 for partly correct, or 2 for completely correct by two raters. The present study used the mean of the two ratings of concept understanding in each story type as IVs. The two SEL administrations tested the same concepts but used different stories.

The Social Interaction Observation System (SIOS) is a procedure for evaluating observed social interaction. Behavior is grouped into three categories: positive (e.g., making eye contact or saying hello), negative (e.g., looking away or punching), and low-level (e.g., looking without making eye contact or standing very close without saying anything). Raters estimate what proportion of the observation period each subject spent engaging in behavior of each category. The present study ignored ratings of negative behavior, since these were 0 for most subjects, but used ratings of positive and low-level social behavior as IVs.

The Social Responsiveness Scale (SRS; Constantino et al., 2003) in an informant-report measure of autism symptoms, such as difficulty communicating feelings and inappropriate laughter. The 65 items are rated from 1 (*not true*) to 4 (*almost always true*). There are five subsets of items (Social Awareness, Social Cognition, Social Communication, Social Motivation, and Restricted Interests and Repetitive Behavior), but the test's authors discourage the use of subscale scores in favor of the overall severity score, which is simply the sum of the item ratings with some items reversed. The present study used the subscale scores as IVs.

The Social Skills Rating System (SSRS) measures social skills, problem behavior, and academic competence with forms varying by rater (self, SSRS-S; or parent, SSRS-P) and the grade level of the subject (Demaray et al., 1995). Social skills are divided into the subdomains of assertion, cooperation, self-control, empathy (self-report form only), and responsibility (parent form only). Problem behavior is divided into externalizing, internalizing, and hyperactivity (preschool and secondary levels only). The various forms have from 34 to 57 items, each of which is rated 0 (*never*), 1 (*sometimes*), or 2 (*very often*). Raw scores are computed by summing item ratings. For IVs, the present study used assertion, cooperation, self-control, and empathy from the SSRS-S, and assertion, cooperation, self-control, responsibility, externalizing, and internalizing from the SSRS-P.

The Wechsler Intelligence Scale for Children (WISC; Williams, Weiss, & Rolfhus, 2003) is a general-purpose intelligence test of which Knowledge or Performance used two subtests. The Vocabulary subtest has subjects produce names for pictures. The Matrix Reasoning subtest is a matrix-analogy task similar to Raven's Progressive Matrices. the present study used the sum of raw scores on both subtests as an IV.

Analyses

All analysis code can be found online at `http://arfer.net/projects/pelt`.

Models

For each combination of a DV with a set of IVs, the present study evaluated the accuracy with which six different models could predict the DV given the IV. These analyses are the *primary analyses*, presented in the section of that name below. Using several models helps to check whether obtained accuracies can be improved with a more complex model (due to underfitting) or a less complex model (due to overfitting). The first three of these models are *baseline models*, present for comparison, whereas the other three, the *critical models*, are the models of chief interest. The Python package scikit-learn (Pedregosa et al., 2011) was used to fit all models.

A trivial model is particularly important for comparison purposes. A "trivial model" refers to a model that, given a training set, guesses a constant value for the DV, ignoring all IVs. Here the appropriate constant value is the median of the DV, since predictions are evaluated with mean absolute error (MAE). The benefit of including a trivial model is that its predictive accuracy provides a baseline. If a nontrivial model is not substantially more accurate than the corresponding trivial model, then the nontrivial model and its set of IVs are not helping to predict the DV (see the discussion of Wall et al., 2012a, and Wall et al., 2012b, in the introduction).

Where possible, it had been planned to use ordinary least squares (OLS); that is, plain multiple linear regression. OLS is a general-purpose data-analytic technique that figures largely in research on autism, not to mention scientific research generally. However, when the data is wide (that is, when there are more IVs than cases, or equally as many), OLS is unidentifiable (or just identified), and the data turned out to be wide for every primary analysis described below. OLS was still used with subsets of the IVs, which included only:

- Pretreatment measures of the same instrument used as a DV for the analysis in question
- Treatment condition
- In the amalgamate analysis (Table 1), dummy variables indicating which study each subject participated in

This model, called OLS-Reduced, provides another baseline for predictive accuracy, showing how well a DV can be predicted with no pretreatment variables beyond grouping variables, the treatment, and the pretreatment administration of the same measure.

Penalized linear regression models arise from the insight that by biasing regression coefficients towards 0 (that is, "penalizing" values that are large in absolute value), test error can be reduced although training error is increased, by reducing overfitting of the coefficients. Penalized methods not only make it possible to use linear regression in problems with more IVs than cases, but also tend to achieve greater predictive accuracy than OLS. Two of the most popular penalized regression methods are ridge regression, which penalizes proportional to the

sum of squares of coefficients, and the lasso, which penalizes proportional to the sum of absolute values of coefficients. The present study used elastic-net regression, which is a generalization including both ridge regression and the lasso as special cases. A tuning parameter controls the relative strength of the ridge and lasso penalties. Three elastic-net models were used, distinguished by the IVs made available to them:

- ENet-Reduced, which uses the same reduced dataset as OLS-Reduced. This model, too, is included as a baseline.
- ENet-Main, which uses all IVs available for the problem, as main effects only.
- ENet-Interact, which uses all IVs available for the problem as main effects, plus every possible first-order interaction. This increases the number of regression terms dramatically (n main effects yield $n(n - 1)$ first-order interactions before redundant terms are removed), but not as much as including every possible interaction (n main effects yield $2^n - (n + 1)$ interactions).

Finally, the analyses include random forests (Breiman, 2001), which grow a decision tree for each of many bootstrap samples of the training data, then make predictions for new cases by aggregating the predictions of the trees. The use of decision trees makes random forests flexible, capable of exploiting nonlinear patterns that the aforementioned regression models cannot, whereas the use of bootstrapping counters against the tendency of single trees to overfit. The random-forest model for each problem, which uses all available IVs, is called RF. It uses 500 trees per forest and it uses a mean-squared-error criterion to choose splits.

Data processing and cross-validation scheme

All non-dichotomous IVs were standardized to have mean 0 and SD 1/2 (Gelman, 2008). Standardization is necessary so that, for example, the penalty term of an elastic-net model treats all IVs equally regardless of their original scales.

There was a small amount of missing data among the included subjects. Collapsing across subjects, 61 of 4,651 IV values (1 in 76) and 69 of 1,324 DV values (1 in 20) were missing. Subjects missing on a DV were simply excluded from each analysis using that DV. Missing IVs were imputed (after the standardization step) using the Soft-Impute matrix-completion algorithm of Mazumder, Hastie, and Tibshirani (2010) as implemented in the Python package fancyimpute; imputed values for dichotomous variables were rounded to 0 or 1. (The primary analyses were also run using the simpler imputation technique of replacing missing values with column medians. This produced similar results.)

It was checked that Soft-Impute was accurate for the data used in the primary analyses by randomly ablating the data and examining Soft-Impute's estimates, as follows. For each of the four full sets of IVs used in the primary analyses, 1 in 30 randomly selected non-missing values were set to missing, and then the dataset was imputed and the imputed values were compared to the true values. This procedure was repeated 1,000 times. On the four datasets, this procedure yielded root mean squared errors (RMSE) of .52, .54, .50, and .56, respectively. Since standardization put variables on a half-SD scale, this means that the RMSE of imputation was about a quarter of an SD.

To estimate predictive accuracy in a fashion unbiased by overfitting, the primary analyses used cross-validation. A problem of ordinary k-fold cross-validation in this context is the dependency between subjects on the basis of treatment group. For example, Knowledge or Performance subjects were run in pairs, and the 30 subjects from Spotlight 2010 were spread

among 8 therapy groups ranging in size (not counting patients who were not subjects of the study or not analyzed) from 1 to 7. (Another instance of grouping, mentioned earlier, is that Knowledge or Performance grouped subjects into quartets in the process of determining dyads and treatment assignments. It had been planned to choose cross-validation folds according to quartet as well, but no record of the quartets exists.) If subjects in the same treatment group appeared in more than one cross-validation fold, we would expect this dependency to inflate estimates of predictive accuracy. Thus, we want to choose folds that are as similar in size as possible but keep treatment groups together. This problem was approached with a brute-force algorithm, enumerating all possible arrangements of treatment groups into 10 nonempty folds and randomly choosing among those arrangements that minimized the sum of squared differences between the fold sizes and the "ideal" fold size (the number of subjects divided by 10). For the analyses of Charlottesville and Spotlight 2007, which had no treatment groups beyond treatment condition and had less than 20 subjects (and hence tenfold cross-validation would leave at least one fold with only one subject), leave-one-out cross-validation was used.

With a selection of folds in hand, each fold in turn was treated as a test set, training each model on the remainder of the data and then having it predict the DV values in the selected fold. The elastic-net models require two tuning parameters, namely, the balance between the lasso and ridge penalties (the L_1 ratio) and the strength of the penalty (α), so these were set with an inner round of fivefold cross-validation on the training set (i.e., on the training folds from the outer cross-validation). L_1 ratios were allowed to vary among {.01, .1, .25, .5, .75, .9, .95, .99, 1}, whereas α varied among 100 values set automatically by scikit-learn. Predicted values were compared to actual values with absolute error, which is less influenced by outliers than squared error.

DVs

In choosing which variables to use as DVs, DVs were selected that represent outcomes of likely practical interest. For example, the PCToMM-E is a measure of children's ability to reason about mental states, and one common reason autistic children may receive treatment is to improve such social skills. This practical approach is in keeping with the goal of the study, which is not to test a theory, but to make progress towards the effective use of pretreatment measures to make treatment decisions.

DVs were drawn from seven instruments, rescaling each to put the minimum of the scale on 0 and the maximum on 1. Here is a list of the instruments and clarification of the direction of scores:

- HAQ subscales: Higher scores mean greater agreement with the type of statement that corresponds to the subscale.
- PCToMM-E: Higher scores mean better theory-of-mind skills.
- SAS: Higher scores mean more anxiety.
- SEL subscales: Higher scores mean better ability to make mental inferences.
- SIOS subscales: Higher scores mean more of the behavior measured by the subscale.
- SRS: Higher scores mean more severe symptoms.
- SSRS subscales: Higher scores are better for the main subscales, but worse for internalizing and externalizing.

Primary analyses

Four families of predictive analyses were conducted. They are shown in the following tables. Each cell for the six models shows how accurately the model could predict the given DV; the rows under the heading "Improvement" show how much the predictive accuracy of the best critical model exceeded (or, when negative, undershot) that of the best baseline model, and hence, how much more accurately one could predict results of treatment given pretreatment variables beyond the pretreatment administration of the same instrument, the treatment condition, and any grouping variables.

- Table 1 uses an amalgamate dataset that combines data from three studies: Spotlight 2008, Spotlight 2010, and Charlottesville. This dataset includes only subjects who received SDARI. (Thus, treatment condition is not an IV.) It draws from the DHF, SAS, SRS, HAQ, SSRS-S, and SSRS-P for IVs, plus 2 dummy variables indicating study, for a total of 45 IVs.
- Table 2 uses data from Charlottesville. It differs from the amalgamate analysis of Table 1 in that it includes subjects who received skillstreaming rather than SDARI, and it includes some IVs and one DV that were not included in the amalgamate analysis because they were not present in Spotlights 2008 and 2010. It uses the same IVs as the amalgamate analysis (minus the study variables) plus the SCQ, the PCToMM-E, and treatment condition, for a total of 46 IVs.
- Table 3 uses data from Spotlight 2007. It draws from the DHF (gender and age only), BDI-Y, CBCL, DANVA-2, EDI, SRS, SSRS-S, and SSRS-P for a total of 37 IVs. Since each instrument (besides the DHF) was administered several times before and after treatment, this analysis used within-subject means of each IV and DV. Administrations during treatment were ignored.
- Table 4 uses data from Knowledge or Performance. It draws from the DHF, ADOS, DANVA-2, CABS, DMQ, SCQ, SCT, SEL, SIOS, SRS, and WISC for IVs, plus the treatment condition, for a total of 45 IVs.

Table 1. Mean absolute error (MAE) of prediction for each dependent variable and model with the amalgamate dataset (which combines subjects from Spotlight 2008, Spotlight 2010, and Charlottesville). Sample sizes vary because subjects were excluded from analyses for which they were missing on the dependent variable. "Improvement" compares the best MAE among the critical models to the best MAE among the baseline models; differences are positive when one of the critical models was more accurate than all the baseline models.

Dependent variable	SAS	SRS	HAQ					SSRS-S				SSRS-P					
			Hostile	Critical	Neutral	Angry	Sad	Cooperation	Assertion	Empathy	Self-control	Cooperation	Assertion	Responsibility	Self-control	Externalizing	Internalizing
Sample size	45	44	36	36	36	36	36	45	45	45	45	45	45	45	45	44	44
Baseline models																	
Trivial	0.104	0.097	0.189	0.206	0.222	0.185	0.224	0.130	0.172	0.158	0.166	0.101	0.144	0.148	0.140	0.201	0.176
OLS-Reduced	0.082	0.071	0.132	0.148	0.122	0.123	0.132	0.089	0.129	0.125	0.157	0.072	0.092	0.111	0.078	0.136	0.137
ENet-Reduced	0.089	0.068	0.143	0.157	0.130	0.120	0.133	0.091	0.128	0.125	0.159	0.072	0.092	0.109	0.078	0.134	0.142
Critical models																	
ENet-Main	0.097	0.074	0.136	0.172	0.150	0.120	0.124	0.109	0.138	0.147	0.171	0.069	0.103	0.118	0.082	0.146	0.145
ENet-Interact	0.095	0.078	0.134	0.183	0.150	0.140	0.157	0.111	0.139	0.153	0.167	0.078	0.107	0.119	0.086	0.152	0.153
RF	0.089	0.073	0.110	0.173	0.171	0.154	0.192	0.125	0.141	0.149	0.163	0.082	0.106	0.117	0.087	0.121	0.137
Improvement																	
Difference	−0.007	−0.006	+0.022	−0.025	−0.028	−0.000	+0.008	−0.021	−0.010	−0.022	−0.006	+0.003	−0.011	−0.009	−0.004	+0.013	+0.000
Ratio	1.079	1.086	0.836	1.166	1.225	1.003	0.938	1.232	1.078	1.176	1.039	0.962	1.122	1.078	1.053	0.900	0.999

Table 2. Mean absolute error of prediction for each dependent variable and model with data from Charlottesville.

Dependent variable	SAS	SRS	HAQ					SSRS-S				SSRS-P						PCToMM-E
			Hostile	Critical	Neutral	Angry	Sad	Cooperation	Assertion	Empathy	Self-control	Cooperation	Assertion	Responsibility	Self-control	Externalizing	Internalizing	
Sample size	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13
Baseline models																		
Trivial	0.120	0.106	0.221	0.279	0.292	0.225	0.292	0.154	0.123	0.148	0.179	0.102	0.110	0.192	0.204	0.244	0.154	0.097
OLS-Reduced	0.107	0.095	0.141	0.090	0.096	0.153	0.130	0.057	0.082	0.069	0.106	0.095	0.109	0.136	0.112	0.131	0.100	0.061
ENet-Reduced	0.102	0.084	0.133	0.094	0.101	0.146	0.138	0.061	0.079	0.075	0.111	0.094	0.117	0.149	0.114	0.115	0.104	0.056
Critical models																		
ENet-Main	0.139	0.097	0.208	0.150	0.108	0.204	0.264	0.076	0.146	0.109	0.173	0.101	0.138	0.195	0.122	0.179	0.187	0.106
ENet-Interact	0.137	0.102	0.149	0.177	0.136	0.178	0.262	0.073	0.134	0.101	0.172	0.097	0.107	0.157	0.109	0.178	0.191	0.112
RF	0.107	0.094	0.177	0.176	0.140	0.208	0.253	0.084	0.133	0.099	0.169	0.089	0.122	0.152	0.162	0.191	0.153	0.100
Improvement																		
Difference	−0.005	−0.010	−0.016	−0.060	−0.012	−0.032	−0.123	−0.016	−0.054	−0.031	−0.062	+0.005	+0.002	−0.017	+0.003	−0.063	−0.054	−0.044
Ratio	1.051	1.117	1.124	1.671	1.120	1.219	1.945	1.273	1.685	1.445	1.588	0.951	0.985	1.124	0.975	1.546	1.537	1.784

Table 3. Mean absolute error of prediction for each dependent variable and model with data from Spotlight 2007.

Dependent variable	SRS	SSRS-S				SSRS-P					
		Cooperation	Assertion	Empathy	Self-control	Cooperation	Assertion	Responsibility	Self-control	Externalizing	Internalizing
Sample size	17	17	17	17	17	17	17	17	17	17	17
Baseline models											
Trivial	0.107	0.119	0.160	0.175	0.154	0.119	0.143	0.119	0.121	0.181	0.168
OLS-Reduced	0.071	0.056	0.116	0.112	0.166	0.062	0.092	0.086	0.107	0.129	0.146
ENet-Reduced	0.074	0.055	0.121	0.122	0.160	0.061	0.095	0.091	0.107	0.138	0.156
Critical models											
ENet-Main	0.082	0.076	0.122	0.174	0.144	0.087	0.133	0.100	0.128	0.192	0.144
ENet-Interact	0.075	0.079	0.125	0.168	0.147	0.079	0.135	0.107	0.139	0.181	0.133
RF	0.086	0.081	0.114	0.168	0.135	0.088	0.148	0.117	0.121	0.171	0.175
Improvement											
Difference	−0.005	−0.021	+0.002	−0.055	+0.019	−0.018	−0.041	−0.014	−0.014	−0.042	+0.013
Ratio	1.068	1.390	0.979	1.494	0.874	1.304	1.450	1.165	1.129	1.326	0.913

Table 4. Mean absolute error of prediction for each dependent variable and model with data from Knowledge or Performance.

Dependent variable	SEL				SIOS	
	Figurative speech	Irony	Contrary emotions	Mistaken intentions	Positive	Low-level
Sample size	39	39	39	39	40	40
Baseline models						
Trivial	0.282	0.298	0.205	0.256	0.235	0.131
OLS-Reduced	0.318	0.273	0.246	0.309	0.117	0.137
ENet-Reduced	0.311	0.275	0.235	0.307	0.118	0.137
Critical models						
ENet-Main	0.320	0.305	0.233	0.319	0.126	0.137
ENet-Interact	0.318	0.295	0.216	0.309	0.135	0.129
RF	0.330	0.310	0.247	0.364	0.144	0.142
Improvement						
Difference	−0.036	−0.023	−0.011	−0.053	−0.009	+0.002
Ratio	1.127	1.083	1.055	1.205	1.076	0.983

Overall, we see that the critical models, with their variety of pretreatment variables, are not helpful for predicting treatment outcomes. Most improvement differences are negative and improvement ratios are greater than 1, meaning that adding pretreatment variables only worsened predictive accuracy. The best improvement is for the HAQ hostile attributions scale in Table 1, which has a difference of +.02 (about a fiftieth of the way along the zero-to-one scale) and has 84% of the error of the best baseline model.

Why did the critical models fail, in this sense? It is useful to compare the nontrivial baseline models, namely OLS-Reduced and ENet-Reduced, to Trivial. To the degree that the instrument used as a DV has good retest reliability, and the treatment condition (when there is a treatment IV) makes a difference, we should see at least one of OLS-Reduced and ENet-Reduced improve on Trivial. This is indeed the case for most DVs, although there are some exceptions (e.g., three of the four SEL subscales in Table 4), and the degree of improvement over Trivial varies substantially between DVs. Improvement of OLS-Reduced and ENet-Reduced over Trivial gives us confidence that the DV is not all noise, because it is connected to pretreatment scores and treatment condition in a systematic fashion. At the same time, OLS-Reduced and ENet-Reduced nowhere achieve perfect accuracy, so there is residual variability for the critical models to predict. Thus, the findings generally suggest that the wide variety of pretreatment variables considered in this study are not, in fact, useful for predicting treatment outcomes.

Secondary analyses

It is often worthwhile to compare predictive accuracy to strength of association. Such an exercise illustrates the gap between predictive accuracy and association despite how the latter is often mistaken or otherwise substituted for the former, and exemplifies how association can be strong even when predictive accuracy is poor. In the case of this study, all the datasets examined have as many IVs as cases or more, so association for every DV would be perfect or near perfect given a suitable model.

Association can still be examined on a variable-by-variable basis with simple correlation. Table 5 through Table 8 present all IV–DV correlations in each of the datasets used in the primary analyses. For simplicity, missing values have been dropped pairwise, rather than imputed. We see that, discounting relationships between two administrations of the same scale or between a scale and one of its own subscales, absolute correlations top out at .64, .93, .76, and .53, respectively. By simulating some bivariate standard normal data with these correlations, and dividing the MAE of the DV estimated with the IV by the MAE of the DV estimated with its median, we get a figure analogous to the improvement ratios in the bottom rows of Table 1

through Table 4. Correlations of .64 and .53 correspond to ratios of .85 and .97, meaning a reduction of no more than 15% of the absolute error. This suggests that the amalgamate and Knowledge or Performance datasets lack any univariate relationships strong enough to leverage for prediction. On the other hand, correlations of .93 and .76 correspond to ratios of .38 and .70, which is a reduction of 62% for Charlottesville and 30% for Spotlight 2007. Such improvements in predictive accuracy, especially the former, could indeed be useful in practice. Perhaps these correlations represent predictively useful relationships in the data that the predictive models failed to pick out. However, since cherry-picking strong correlations like this constitutes data dredging, and this analysis has not attempted to distinguish association from prediction, it is unlikely that these results would generalize.

Table 5. Pearson correlations between all IVs and DVs in the amalgamate dataset (which combines subjects from Spotlight 2008, Spotlight 2010, and Charlottesville).

Dependent variable	SAS	SRS	HAQ					SSRS-S				SSRS-P					
			Hostile	Critical	Neutral	Angry	Sad	Cooperation	Assertion	Empathy	Self-control	Cooperation	Assertion	Responsibility	Self-control	Externalizing	Internalizing
Study: Spotlight 2008	0.194	-0.109	-0.040	0.074	0.151	-0.094	0.106	0.130	-0.114	-0.009	0.167	0.214	-0.036	0.089	0.099	-0.239	-0.156
Study: Spotlight 2010	0.017	0.067	0.007	-0.131	-0.014	-0.042	0.179	-0.043	0.190	-0.010	0.111	-0.153	0.133	0.006	0.079	0.019	0.187
Female	0.121	0.164	-0.221	-0.318	0.204	-0.188	-0.004	-0.133	-0.012	-0.254	0.162	0.077	-0.241	0.102	0.232	-0.099	0.310
Age	0.350	-0.116	-0.137	-0.146	-0.104	-0.189	-0.046	0.005	-0.418	0.112	-0.176	0.251	-0.187	0.316	0.431	-0.550	-0.121
One sibling	-0.022	-0.062	0.167	0.153	-0.181	0.319	0.144	0.014	-0.163	-0.056	0.017	0.156	0.052	0.106	-0.125	0.097	-0.056
Multiple siblings	-0.069	0.142	-0.307	-0.258	0.108	-0.268	-0.245	-0.048	-0.087	0.170	-0.143	-0.230	-0.065	-0.304	0.021	0.061	-0.015
Parent with grad. degree	0.055	-0.164	0.075	0.090	-0.131	0.311	0.046	0.026	-0.095	-0.027	-0.132	0.378	0.157	-0.088	-0.086	0.156	-0.148
Stay-at-home parent	-0.238	0.187	-0.184	-0.096	0.205	-0.235	-0.086	-0.028	0.075	0.020	-0.030	-0.282	-0.047	-0.191	-0.374	0.369	0.092
Income	-0.021	-0.065	0.124	0.069	-0.056	-0.098	0.286	0.301	0.007	0.189	0.134	0.228	-0.027	0.206	-0.075	-0.069	-0.019
Parents are together	-0.226	-0.016	-0.160	-0.108	0.010	-0.147	0.112	0.115	-0.133	0.132	0.001	0.225	-0.057	-0.137	-0.237	0.069	-0.074
Attends public school	0.315	0.099	-0.042	-0.174	0.116	0.068	0.156	0.234	0.011	0.093	0.165	0.104	-0.030	0.051	0.206	-0.146	-0.053
No. of medications	0.320	0.145	0.045	-0.055	0.015	0.228	0.161	-0.296	-0.068	-0.242	-0.259	-0.107	-0.202	-0.103	0.298	-0.080	0.219
No. of interventions	0.173	0.037	-0.052	-0.111	0.139	-0.106	0.080	0.100	0.043	0.063	0.045	0.107	0.147	0.296	0.139	-0.007	0.083
Affected parent or sibling	0.020	0.205	-0.226	-0.265	0.171	-0.159	-0.195	-0.013	0.148	0.012	0.073	-0.164	0.076	-0.296	0.035	0.102	0.101
Other affected relative	-0.068	-0.031	-0.206	-0.328	0.264	-0.150	-0.059	-0.075	-0.064	0.130	0.082	-0.070	-0.085	0.070	-0.015	-0.017	-0.033
Diagnosis: Learning disorder	-0.006	0.128	-0.172	0.065	0.043	0.029	-0.272	-0.086	0.136	-0.077	-0.080	0.080	0.048	0.077	0.224	-0.179	0.433
Diagnosis: Anxiety	-0.102	-0.213	0.166	0.011	0.017	0.045	0.152	-0.028	0.136	-0.111	0.044	-0.131	0.108	-0.008	0.140	-0.189	0.066
Diagnosis: ADD	0.032	0.368	-0.391	-0.273	0.061	-0.193	-0.176	-0.217	0.006	-0.149	-0.096	-0.288	-0.251	-0.412	0.014	-0.026	0.308
Diagnosis: Autism	0.180	-0.089	0.198	0.255	-0.084	-0.062	0.279	-0.005	0.026	-0.060	0.109	0.154	-0.155	-0.058	0.007	-0.006	-0.128
Diagnosis: Asperger	-0.104	0.099	0.018	0.054	-0.026	0.086	-0.169	-0.025	0.050	0.021	-0.087	-0.328	0.004	-0.166	-0.276	0.040	0.053
Diagnosis: PDD-NOS	-0.043	0.073	-0.163	-0.243	0.203	-0.289	-0.010	0.183	-0.124	0.103	0.131	0.303	-0.211	0.091	0.308	-0.170	-0.135
Diagnosis: Other	0.046	0.076	-0.145	0.076	-0.157	0.248	0.009	-0.175	-0.055	-0.165	-0.253	-0.126	0.226	-0.236	-0.010	0.200	0.101
SAS: FNE	0.624	-0.063	0.385	0.322	-0.256	0.165	0.352	0.062	-0.068	0.206	0.145	0.062	-0.063	0.239	0.059	-0.095	0.282
SAS: SAD New	0.485	0.233	0.114	0.006	-0.226	0.092	0.256	-0.137	-0.131	0.021	0.036	0.089	-0.288	-0.049	-0.089	0.075	0.148
SAS: SAD General	0.525	0.095	0.054	-0.060	-0.077	0.002	0.154	-0.048	-0.232	0.016	0.060	0.125	-0.205	0.173	0.157	-0.137	0.155
SRS: Awareness	-0.329	0.394	-0.008	-0.125	0.001	-0.167	-0.089	0.086	0.276	0.034	-0.004	-0.641	0.010	-0.526	-0.407	0.281	-0.004
SRS: Cognition	0.027	0.633	-0.083	-0.194	0.089	-0.224	0.026	-0.018	0.091	-0.057	-0.029	-0.365	-0.329	-0.447	-0.043	-0.150	0.379
SRS: Communication	-0.090	0.684	-0.051	-0.172	-0.121	-0.214	-0.276	-0.037	-0.124	-0.125	-0.285	-0.156	-0.571	-0.492	-0.358	0.116	0.394
SRS: Motivation	0.155	0.483	0.038	0.064	-0.172	-0.058	-0.035	-0.336	-0.194	-0.258	-0.312	-0.197	-0.612	-0.284	0.026	-0.114	0.556
SRS: Mannerisms	-0.016	0.545	0.029	-0.136	-0.120	-0.135	-0.002	-0.082	0.119	-0.006	-0.069	-0.357	-0.444	-0.361	-0.111	-0.012	0.477
HAQ: Hostile	0.376	-0.079	0.706	0.433	-0.285	0.366	0.377	-0.028	0.105	-0.175	-0.027	0.143	0.004	0.166	-0.249	0.148	-0.001
HAQ: Critical	0.235	-0.128	0.455	0.704	-0.561	0.459	0.148	-0.265	-0.040	-0.193	-0.253	0.168	0.141	0.162	-0.323	0.201	-0.031
HAQ: Neutral	-0.144	-0.101	-0.222	-0.442	0.814	-0.112	0.001	0.299	0.130	0.158	0.395	-0.136	0.147	0.114	0.309	-0.198	-0.016
HAQ: Angry	0.142	-0.253	0.535	0.573	-0.307	0.709	0.391	-0.104	-0.078	-0.292	-0.093	0.220	0.215	0.207	-0.161	0.223	-0.124
HAQ: Sad	0.354	-0.321	0.356	0.258	-0.018	0.353	0.799	0.196	0.133	0.160	0.291	0.223	0.164	0.339	0.099	-0.032	-0.046
SSRS-S: Cooperation	-0.020	-0.112	-0.154	-0.372	0.262	-0.157	0.231	0.719	0.473	0.493	0.562	0.236	0.176	0.334	0.132	-0.231	-0.127
SSRS-S: Assertion	-0.151	0.028	-0.019	-0.284	0.293	-0.149	0.130	0.538	0.700	0.435	0.457	-0.114	0.313	-0.011	-0.002	0.042	-0.031
SSRS-S: Empathy	0.035	-0.071	-0.166	-0.231	0.234	-0.130	0.153	0.454	0.458	0.606	0.510	-0.198	0.301	0.208	0.190	-0.248	0.036
SSRS-S: Self-control	0.035	-0.122	-0.231	-0.383	0.340	-0.363	0.080	0.550	0.407	0.504	0.562	0.042	0.166	0.250	0.227	-0.221	-0.080
SSRS-P: Cooperation	0.106	-0.341	0.237	0.190	-0.227	0.142	0.099	0.150	-0.229	0.024	-0.147	0.782	-0.020	0.310	0.164	-0.089	-0.230
SSRS-P: Assertion	-0.209	-0.415	-0.003	0.192	0.041	0.012	0.081	0.144	0.236	0.206	0.193	0.019	0.774	0.108	-0.069	0.117	-0.269
SSRS-P: Responsibility	0.313	-0.345	0.101	0.090	-0.017	0.162	0.143	0.103	-0.170	0.103	0.065	0.402	0.128	0.683	0.381	-0.377	-0.034
SSRS-P: Self-control	0.171	-0.140	-0.303	-0.435	0.387	-0.020	0.013	0.028	-0.146	0.054	0.181	0.329	-0.020	0.441	0.850	-0.529	0.043
SSRS-P: Externalizing	-0.288	0.222	0.034	0.084	-0.243	0.065	0.105	-0.032	0.160	-0.104	-0.195	-0.222	0.019	-0.437	-0.541	0.699	-0.254
SSRS-P: Internalizing	0.204	0.122	0.055	-0.008	0.067	0.048	0.254	0.042	0.340	0.035	0.322	-0.027	-0.174	0.102	0.148	-0.244	0.596

Table 6. Pearson correlations between all IVs and DVs in data from Charlottesville.

Dependent variable	SAS	SRS	HAQ					SSRS-S				SSRS-P						PCToMM-E
			Hostile	Critical	Neutral	Angry	Sad	Cooperation	Assertion	Empathy	Self-control	Cooperation	Assertion	Responsibility	Self-control	Externalizing	Internalizing	
Treatment	-0.276	-0.033	-0.212	-0.074	-0.083	0.108	-0.419	-0.035	-0.254	-0.090	-0.274	0.119	-0.336	0.207	-0.198	0.215	0.236	-0.123
Age	0.278	-0.226	0.636	0.542	-0.615	0.449	0.033	0.051	-0.056	-0.117	-0.155	0.462	-0.443	0.336	0.149	-0.236	0.112	0.217
One sibling	0.061	0.095	-0.138	-0.068	0.160	0.359	-0.301	0.417	0.273	0.264	0.135	-0.129	0.040	0.130	0.233	-0.163	0.462	0.340
Multiple siblings	-0.061	-0.095	0.138	0.068	-0.160	-0.359	0.301	-0.417	-0.273	-0.264	-0.135	0.129	-0.040	-0.130	-0.233	0.163	-0.462	-0.340
Parent with grad. degree	0.529	0.076	0.341	-0.148	0.122	0.198	0.157	0.173	0.445	0.000	0.061	0.077	0.265	-0.232	-0.257	0.390	0.276	-0.125
Stay-at-home parent	-0.281	-0.076	-0.124	-0.253	0.330	-0.365	-0.100	0.108	0.241	0.231	0.182	-0.293	0.255	-0.334	0.011	-0.120	0.108	0.120
Income	0.484	-0.471	-0.032	-0.209	0.276	0.153	0.537	0.268	0.170	0.095	0.419	0.264	0.333	0.314	0.563	-0.283	0.177	0.556
Parents are together	0.077	-0.230	0.000	-0.383	0.526	-0.416	0.215	0.385	0.214	0.249	0.623	-0.171	0.370	-0.074	0.098	-0.168	-0.093	-0.152
Attends public school	0.485	0.156	0.324	0.020	0.084	0.260	0.070	0.517	0.319	0.308	0.323	-0.051	-0.279	0.056	0.215	-0.191	0.150	-0.029
No. of medications	0.121	0.220	-0.131	0.044	-0.219	0.277	0.018	0.044	0.045	0.124	-0.188	0.231	-0.439	0.307	0.276	-0.162	0.111	-0.002
No. of interventions	0.272	-0.099	0.058	0.198	0.019	-0.025	-0.191	0.159	0.042	-0.018	-0.014	-0.027	0.124	-0.240	0.330	-0.349	0.650	0.296
Affected parent or sibling	-0.019	0.473	0.124	-0.012	0.054	0.003	0.013	0.249	0.536	0.277	0.320	-0.293	-0.265	-0.334	-0.235	-0.066	0.365	-0.403
Other affected relative	-0.179	-0.581	-0.286	-0.527	0.484	-0.315	0.362	-0.088	-0.274	-0.107	0.222	0.013	0.386	0.074	0.056	0.278	-0.472	0.226
Diagnosis: Learning disorder	-0.096	0.230	-0.334	0.166	-0.154	0.326	-0.215	-0.195	-0.081	-0.187	-0.155	-0.022	-0.020	-0.027	0.091	-0.123	0.525	-0.171
Diagnosis: ADD	-0.240	0.294	0.031	0.282	-0.347	0.103	-0.185	-0.173	-0.004	-0.092	-0.131	-0.006	-0.265	-0.070	-0.165	-0.120	0.172	-0.451
Diagnosis: Autism	-0.170	0.259	0.167	0.238	-0.154	0.101	-0.024	0.058	0.051	0.187	0.173	-0.216	-0.108	0.125	0.044	-0.269	-0.425	-0.112
Diagnosis: Asperger	0.243	0.392	-0.286	-0.121	0.038	0.006	0.161	-0.184	0.218	-0.107	-0.182	0.070	0.138	-0.074	-0.381	0.407	0.324	-0.110
Diagnosis: PDD-NOS	-0.077	-0.674	-0.042	-0.194	0.219	-0.215	0.015	0.312	-0.214	0.187	0.266	0.171	-0.020	0.278	0.707	-0.560	-0.080	0.347
Diagnosis: Other	0.235	-0.244	-0.072	0.214	-0.299	0.033	0.461	-0.630	-0.331	-0.587	-0.379	0.427	0.386	0.030	-0.025	0.216	-0.102	0.166
SAS: FNE	0.712	-0.240	0.616	0.177	-0.323	0.375	0.558	0.246	0.281	-0.029	0.235	0.656	-0.122	0.513	0.062	-0.071	0.097	-0.029
SAS: SAD New	0.541	0.122	0.261	0.056	-0.067	0.346	0.336	0.332	0.639	0.207	0.309	0.234	-0.033	0.156	0.040	-0.155	0.635	0.153
SAS: SAD General	0.567	0.108	0.310	-0.005	0.021	0.233	0.005	0.538	0.431	0.243	0.344	0.216	-0.044	0.236	0.113	-0.232	0.478	-0.318
SRS: Awareness	-0.156	0.771	0.125	0.041	0.110	-0.019	-0.309	-0.036	0.413	0.115	-0.053	-0.714	-0.012	-0.867	-0.572	0.392	0.086	-0.442
SRS: Cognition	0.202	0.602	0.054	0.010	0.057	0.042	0.123	0.228	0.659	0.294	0.274	-0.262	-0.146	-0.385	-0.096	-0.141	0.492	-0.325
SRS: Communication	0.152	0.688	0.252	0.297	-0.200	0.136	-0.459	0.187	0.382	0.132	-0.111	-0.257	-0.307	-0.390	-0.373	0.034	0.562	-0.460
SRS: Motivation	0.141	0.346	0.122	0.495	-0.487	0.501	-0.091	0.015	0.259	0.053	-0.106	0.131	-0.369	0.135	0.167	-0.406	0.614	0.115
SRS: Mannerisms	0.077	0.569	0.114	0.204	-0.090	0.220	-0.016	0.121	0.459	0.289	0.020	-0.356	-0.316	-0.321	-0.016	-0.098	0.230	0.214
HAQ: Hostile	0.409	0.191	0.876	0.456	-0.574	0.494	-0.029	0.180	0.336	0.004	-0.051	0.328	-0.301	0.174	-0.309	0.054	0.058	-0.392
HAQ: Critical	0.154	0.176	0.545	0.938	-0.929	0.685	-0.344	-0.493	-0.292	-0.568	-0.658	0.299	-0.084	0.081	-0.311	0.108	0.179	-0.034
HAQ: Neutral	-0.111	-0.274	-0.573	-0.922	0.952	-0.625	0.390	0.457	0.254	0.485	0.722	-0.263	0.285	-0.041	0.382	-0.163	-0.135	0.037
HAQ: Angry	0.483	0.074	0.557	0.606	-0.746	0.775	0.052	-0.321	-0.064	-0.521	-0.471	0.520	-0.173	0.172	-0.327	0.351	0.318	-0.089
HAQ: Sad	0.563	-0.287	0.018	-0.432	0.269	0.048	0.812	0.478	0.548	0.332	0.596	0.447	-0.137	0.379	0.422	-0.247	0.347	0.185
SSRS-S: Cooperation	0.222	0.081	0.075	-0.635	0.544	-0.290	0.277	0.932	0.802	0.876	0.790	-0.066	-0.393	0.109	0.373	-0.433	0.173	-0.225
SSRS-S: Assertion	0.208	0.473	-0.034	-0.558	0.586	-0.296	0.190	0.810	0.858	0.829	0.702	-0.381	-0.366	-0.226	0.139	-0.237	0.236	-0.297
SSRS-S: Empathy	-0.141	0.267	-0.388	-0.765	0.714	-0.429	0.192	0.767	0.700	0.916	0.738	-0.356	-0.384	0.018	0.325	-0.377	0.103	-0.098
SSRS-S: Self-control	-0.075	0.179	-0.419	-0.783	0.814	-0.510	0.305	0.774	0.628	0.847	0.907	-0.380	-0.199	0.007	0.390	-0.447	0.053	-0.193
SSRS-P: Cooperation	0.445	-0.132	0.190	0.065	-0.249	0.060	0.095	0.293	-0.017	0.141	-0.020	0.621	-0.268	0.732	0.270	-0.289	-0.031	-0.088
SSRS-P: Assertion	-0.242	-0.347	-0.661	-0.304	0.491	-0.542	0.142	-0.172	-0.373	-0.085	0.186	-0.172	0.629	0.036	0.279	-0.171	-0.228	0.341
SSRS-P: Responsibility	0.489	-0.281	0.180	0.044	-0.262	0.023	0.456	0.230	-0.045	0.131	0.080	0.714	-0.329	0.743	0.562	-0.462	-0.207	0.019
SSRS-P: Self-control	0.129	-0.333	-0.343	-0.382	0.232	-0.233	0.459	0.376	0.074	0.404	0.406	0.363	-0.239	0.545	0.886	-0.720	-0.041	0.132
SSRS-P: Externalizing	-0.132	0.087	-0.039	-0.124	0.122	-0.130	0.043	-0.448	-0.226	-0.412	-0.261	-0.226	0.320	-0.434	-0.714	0.864	-0.370	-0.190
SSRS-P: Internalizing	0.203	0.125	0.012	0.068	0.018	0.394	0.044	0.363	0.507	0.267	0.350	-0.032	-0.186	0.012	0.304	-0.423	0.839	0.210
SCQ	-0.076	-0.037	-0.003	-0.204	0.175	-0.294	-0.327	0.614	0.183	0.563	0.274	-0.006	-0.412	0.139	0.563	-0.694	0.160	-0.249
PCToMM-E	-0.114	-0.347	-0.077	0.241	-0.111	0.199	0.170	-0.473	-0.406	-0.379	-0.252	-0.030	0.340	0.004	0.074	0.160	-0.280	0.831

Table 7. Pearson correlations between all IVs and DVs in data from Spotlight 2007.

Dependent variable	SRS	SSRS-S				SSRS-P					
		Cooperation	Assertion	Empathy	Self-control	Cooperation	Assertion	Responsibility	Self-control	Externalizing	Internalizing
Treatment	-0.488	0.191	0.196	0.254	0.024	0.176	0.634	0.370	0.529	-0.114	-0.276
Female	0.481	-0.731	-0.262	-0.073	-0.422	-0.018	-0.131	-0.381	-0.018	0.008	0.244
Age	-0.373	0.380	0.341	0.140	0.437	0.136	0.024	0.254	0.096	-0.160	-0.233
BDI	0.508	-0.544	-0.517	-0.385	-0.759	-0.340	0.061	-0.364	-0.170	0.186	0.419
CBCL: Anxious/depressed	0.476	-0.216	-0.431	-0.132	-0.543	-0.132	0.121	-0.224	-0.111	0.026	0.385
CBCL: Withdrawn/depressed	0.632	-0.296	-0.131	-0.209	-0.215	-0.276	-0.666	-0.642	-0.609	0.354	0.675
CBCL: Somatic complaints	0.498	-0.396	-0.343	-0.030	-0.628	-0.170	-0.037	-0.538	-0.275	0.157	0.438
CBCL: Social problems	0.678	-0.430	-0.246	0.108	-0.401	-0.145	0.064	-0.297	-0.258	0.260	0.148
CBCL: Thought problems	0.601	-0.353	-0.228	-0.146	-0.658	-0.428	0.078	-0.452	-0.180	0.194	0.101
CBCL: Attention problems	0.385	-0.286	0.112	0.124	-0.199	-0.439	-0.081	-0.204	-0.225	0.057	-0.082
CBCL: Rule-breaking	0.377	-0.520	0.012	-0.364	-0.556	-0.397	-0.004	-0.256	-0.221	0.648	0.078
CBCL: Aggression	0.360	-0.558	-0.197	-0.609	-0.719	-0.529	-0.145	-0.218	-0.362	0.602	0.372
DANVA: Adult voices	-0.184	-0.052	0.252	0.238	0.255	0.281	0.153	0.159	0.453	-0.384	-0.249
DANVA: Child voices	-0.002	-0.258	0.057	-0.127	0.160	0.003	-0.378	-0.153	-0.103	0.007	0.183
DANVA: Adult faces	0.147	-0.125	-0.046	-0.045	0.054	-0.013	-0.383	-0.418	-0.134	-0.236	0.236
DANVA: Child faces	0.044	-0.222	-0.213	-0.153	-0.176	-0.132	0.034	-0.068	0.034	-0.022	-0.114
EDI: Gaze and eye contact	0.217	0.121	0.071	0.270	0.012	-0.274	0.231	-0.237	-0.142	0.195	-0.138
EDI: Space and touch	-0.054	0.095	0.460	0.473	0.194	-0.071	0.358	0.073	0.265	0.071	-0.208
EDI: Paralanguage	0.240	-0.132	0.108	0.195	-0.181	-0.182	0.324	-0.100	-0.073	0.289	-0.020
EDI: Facial expression	0.384	-0.157	0.138	-0.037	-0.317	-0.321	-0.073	-0.385	-0.425	0.299	0.285
EDI: Objectics	0.704	-0.433	-0.056	-0.170	-0.450	-0.361	-0.082	-0.244	-0.338	0.347	0.152
EDI: Social rules/norms	0.575	-0.339	-0.090	0.016	-0.311	-0.231	0.157	-0.040	-0.128	0.099	-0.111

Dependent variable	SRS	SSRS-S				SSRS-P					
		Cooperation	Assertion	Empathy	Self-control	Cooperation	Assertion	Responsibility	Self-control	Externalizing	Internalizing
EDI: Nonverbal reciprocity	0.554	-0.301	-0.296	-0.182	-0.444	-0.255	-0.005	-0.208	-0.464	0.389	0.190
SRS: Awareness	0.502	-0.372	-0.239	-0.216	-0.406	-0.164	0.075	0.072	-0.192	0.085	-0.140
SRS: Cognition	0.664	-0.426	-0.229	-0.128	-0.348	-0.100	-0.020	-0.092	-0.273	-0.055	-0.007
SRS: Communication	0.679	-0.568	-0.146	-0.066	-0.362	-0.179	-0.057	-0.195	-0.352	0.224	0.154
SRS: Motivation	0.638	-0.420	-0.329	-0.247	-0.177	-0.213	-0.597	-0.406	-0.581	0.230	0.670
SRS: Mannerisms	0.742	-0.375	-0.304	-0.159	-0.500	-0.241	0.034	-0.284	-0.229	0.031	-0.002
SSRS-S: Cooperation	-0.457	0.862	0.208	0.234	0.630	0.259	0.124	0.479	0.232	-0.359	-0.227
SSRS-S: Assertion	-0.276	0.386	0.802	0.479	0.478	0.124	0.257	0.246	0.189	-0.028	-0.340
SSRS-S: Empathy	0.012	0.274	0.139	0.657	0.351	0.422	0.338	0.186	0.283	-0.224	-0.256
SSRS-S: Self-control	-0.317	0.446	-0.055	0.196	0.461	0.541	0.124	0.363	0.401	-0.413	-0.333
SSRS-P: Cooperation	-0.121	-0.067	-0.157	0.088	0.183	0.873	0.175	0.245	0.440	-0.391	-0.193
SSRS-P: Assertion	-0.373	0.272	0.224	0.518	0.232	0.602	0.649	0.315	0.610	-0.482	-0.714
SSRS-P: Responsibility	-0.737	0.488	0.312	0.327	0.589	0.450	0.348	0.757	0.652	-0.432	-0.577
SSRS-P: Self-control	-0.741	0.377	0.384	0.271	0.556	0.675	0.357	0.465	0.618	-0.210	-0.431
SSRS-P: Externalizing	0.369	-0.370	0.077	-0.306	-0.272	-0.485	-0.332	-0.158	-0.480	0.758	0.380
SSRS-P: Internalizing	0.447	-0.144	-0.193	-0.116	-0.172	-0.384	-0.370	-0.335	-0.534	0.258	0.650

Table 8. Pearson correlations between all IVs and DVs in data from Knowledge or Performance.

Dependent variable	SEL				SIOS	
	Figurative speech	Irony	Contrary emotions	Mistaken intentions	Positive	Low-level
Treatment	-0.250	-0.111	0.074	0.166	0.082	-0.051
Female	0.081	0.072	0.140	0.265	-0.008	0.072
Age	0.348	0.370	0.074	0.065	-0.303	0.092
One sibling	-0.201	-0.256	-0.015	-0.281	0.054	-0.178
Multiple siblings	0.203	0.169	-0.082	0.144	-0.073	0.327
Parent with grad. degree	0.060	-0.031	0.135	-0.316	-0.094	0.113
Stay-at-home parent	-0.224	0.015	0.309	-0.213	-0.162	0.129
Income	-0.018	0.152	-0.085	-0.257	0.004	0.033
Parents are together	-0.130	0.118	-0.198	-0.248	-0.095	0.161
Attends public school	0.163	0.106	-0.201	-0.201	-0.121	-0.116
No. of medications	-0.234	-0.162	0.098	-0.231	0.126	-0.179
No. of interventions	-0.016	0.237	-0.150	-0.058	-0.223	-0.027
Affected parent or sibling	0.001	0.315	-0.117	-0.011	-0.109	0.201
Other affected relative	-0.040	-0.022	0.042	0.159	0.007	-0.359
Diagnosis: Learning disorder	-0.247	0.079	0.197	-0.009	0.101	-0.249
Diagnosis: Anxiety	0.155	0.403	0.047	-0.075	-0.260	-0.066
Diagnosis: ADD	-0.190	-0.032	-0.185	-0.128	0.006	-0.224
Diagnosis: Autism	-0.341	-0.025	-0.243	-0.052	0.049	0.229
Diagnosis: Asperger	0.091	0.067	0.015	0.053	0.099	-0.334
Diagnosis: PDD-NOS	0.031	-0.016	0.133	-0.059	0.060	-0.031
Diagnosis: Other	0.133	0.306	0.230	0.173	-0.164	0.159
ADOS: Communication	-0.080	0.078	-0.022	0.109	-0.082	0.111
ADOS: Social	-0.133	-0.061	-0.209	0.089	-0.102	0.094
ADOS: SBaRI	-0.231	-0.300	0.075	-0.130	0.006	-0.005
DANVA: Adult voices	-0.339	-0.319	-0.188	-0.181	-0.088	0.185
DANVA: Child voices	-0.349	-0.230	-0.527	-0.017	0.187	0.057
DANVA: Adult faces	0.017	-0.331	-0.049	0.031	0.090	-0.131
DANVA: Child faces	-0.104	-0.088	-0.319	0.278	0.146	0.017
CABS: Passive	0.033	-0.021	0.073	0.213	0.173	-0.096
CABS: Aggression	-0.208	-0.264	-0.096	-0.021	0.109	-0.167
DMQ	0.013	-0.157	0.065	0.288	0.125	-0.089
SCQ	-0.086	-0.005	-0.212	-0.123	-0.275	0.122
SCT	-0.126	-0.132	-0.154	-0.178	0.323	-0.119
SEL: Figurative speech	0.189	0.341	0.127	0.036	-0.320	0.078
SEL: Irony	0.268	0.491	0.304	-0.026	-0.202	0.010
SEL: Contrary emotions	0.171	0.099	-0.051	0.104	0.044	-0.094
SEL: Mistaken intentions	-0.051	-0.079	0.435	-0.082	0.180	0.022
SIOS: Positive	-0.009	-0.152	0.004	0.019	0.784	-0.136
SIOS: Low-level	-0.060	0.136	0.132	0.091	-0.462	0.387
SRS: Awareness	-0.055	0.022	-0.201	0.031	-0.101	-0.171
SRS: Cognition	-0.283	-0.196	-0.291	-0.068	-0.149	0.128
SRS: Communication	0.026	0.082	-0.211	0.149	-0.164	0.034
SRS: Motivation	0.051	0.096	-0.098	0.090	-0.370	-0.001
SRS: Mannerisms	-0.106	-0.206	-0.168	0.147	-0.131	0.042
WISC	0.337	0.263	0.305	0.040	-0.116	0.079

Below are presented follow-up analyses to the evaluations of predictive accuracy in which counterfactual predictions are examined. "Counterfactual prediction" refers to having models generate predictions of outcomes under scenarios such as subjects having received a different

treatment from what they did receive, or subjects having different values on one or more pretreatment measures. Of particular interest is simulating general populations of subjects and predicting treatment results for them, giving an idea of the consequences of making a given treatment the standard of care. As mentioned above, none of the critical models were particularly successful; they improved on the baseline models modestly at best. This means that complex counterfactual analyses of the sort just described will be at best modestly more accurate than estimating treatment outcomes on the basis of pretreatment state and treatment condition alone.

With the foregoing caveat, we can at least perform counterfactual analyses to provide possible predictive connections that future studies might try to verify and explore in more detail. Considered are three DVs with relatively good performance from the three largest of the four datasets:

- The externalizing subscale of the SSRS-P from the amalgamate dataset (in Table 1, improvement difference +.013, improvement ratio .900)
- The self-control subscale of the SSRS-S from Spotlight 2007 (Table 3, improvement difference +0.019, improvement ratio .874)
- Low-level socializing as measured on the SIOS from Knowledge or Performance (Table 4, improvement difference +0.002, improvement ratio .983)

SSRS-P externalizing.

In the amalgamate dataset, all subjects received SDARI, so it was not possible to estimate how subjects would have fared with no treatment or a different treatment. It was still possible to estimate the distribution of posttreatment scores for a general population. To determine what IVs to consider, the best-performing critical model for this DV (RF) was to used to produce IV importances (measured as mean decrease in node impurity (variance of left branch plus variance of right branch) across all the trees in the random forest; Breiman, 1984), normalized to sum to 1 across all IVs. The most important IVs were the pretreatment measure of the DV, SSRS-P externalizing, at .49, and age, at .11; the remaining IVs had importances less than .05. The following algorithm was used to simulate 10,000 subjects:

- Choose a gender, male or female, with equal probability. (Although gender had a low importance of 0.003, it is needed to select SSRS-P scores later because the SSRS-P norms are stratified by gender.)
- Choose an age, an integer from 8 to 17 inclusive (the range of the data), with equal probability (since ages among living Americans seem to be about uniformly distributed in this interval; United States Census Bureau, 2014).
- Choose a problem-behaviors score using the gender-specific percentiles in the SSRS-P norms. Norms for externalizing specifically are not listed in the SSRS-P forms, but problem-behaviors scores are simply the sum of externalizing and internalizing, and a simple linear regression model was fit in the amalgamate dataset to infer (pretreatment) externalizing on the basis of problem behaviors. The root mean squared residual from this model is used as the SD of an error term which is then added to the output.
- Choose a posttreatment externalizing score by using a random forest (fit to age, gender, and pretreatment externalizing in the amalgamate dataset) and adding normally distributed error with SD set to the root mean squared error of prediction, and rounding the result.

Figure 1 shows the distributions of the simulated pretreatment and posttreatment

externalizing scores. The externalizing scores are shown on their original 0 to 12 scales rather than the 0 to 1 scales used for all DVs in the primary analyses, since the discrete nature of the data plays a role in the simulation. We see that treated children are generally expected to be higher in externalizing than untreated children. Indeed, looking within-subjects in the simulated data, 75% of subjects are higher in externalizing after than before treatment, and the expected change in externalizing from treatment is +2.3 points. In short, this analysis predicts that one consequence of use of SDARI in the population would be to moderately increase children's externalizing behavior, as perceived by their guardians.

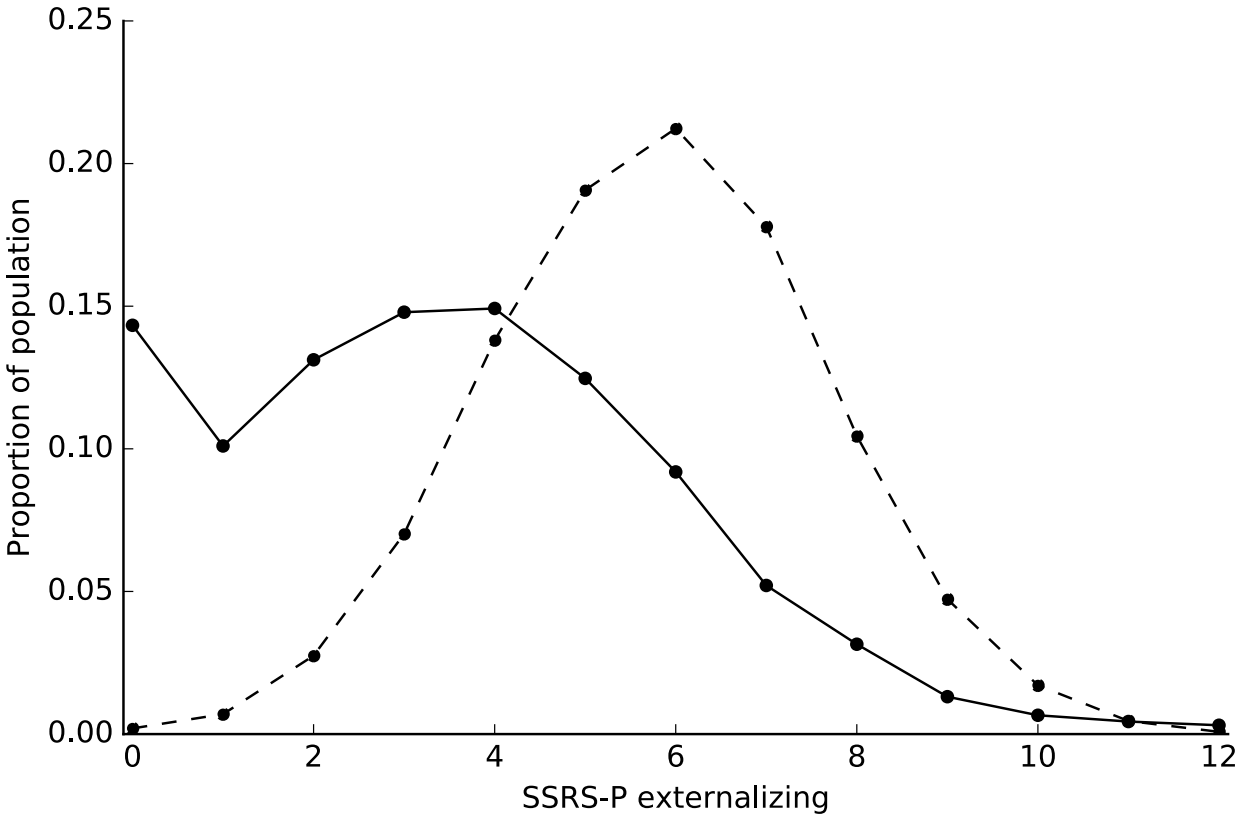

Figure 1. The distribution of simulated SSRS-P externalizing scores in a population of American children ages 8 to 17 before (solid line) and after (dashed line) treatment with SDARI.

SSRS-S self-control.

In Spotlight 2007, some subjects received treatment (namely, SDARI), whereas others did not. What outcomes would we have seen had each subject been assigned to the opposite condition? As before, the best-performing model was RF, so a random forest was used to make predictions. Figure 2 shows predictions alongside actual pretreatment and posttreatment scores. (Here, "posttreatment" refers only to a point in time, since the control condition was no treatment at all.) The figure portrays a complex picture where predicted posttreatment scores are usually (but not always) nearer the posttreatment scores for the other condition than they are to the pretreatment scores. Moreover, the model predicts SDARI as better than nothing for some

subjects, but worse for others. Of the 8 untreated subjects, 3 are predicted to have been better off with SDARI, and of the 9 treated subjects, 5 are predicted to have been better off with nothing. Across all subjects, regarding actual and predicted outcomes equally, no treatment led to a +0.37 change in self-control, while SDARI led to a +0.35 change in self-control. The model thus predicts that in general, no treatment improves self-reported self-control more than SDARI does, but the difference is very small. More important is the condition-neutral improvement over time.

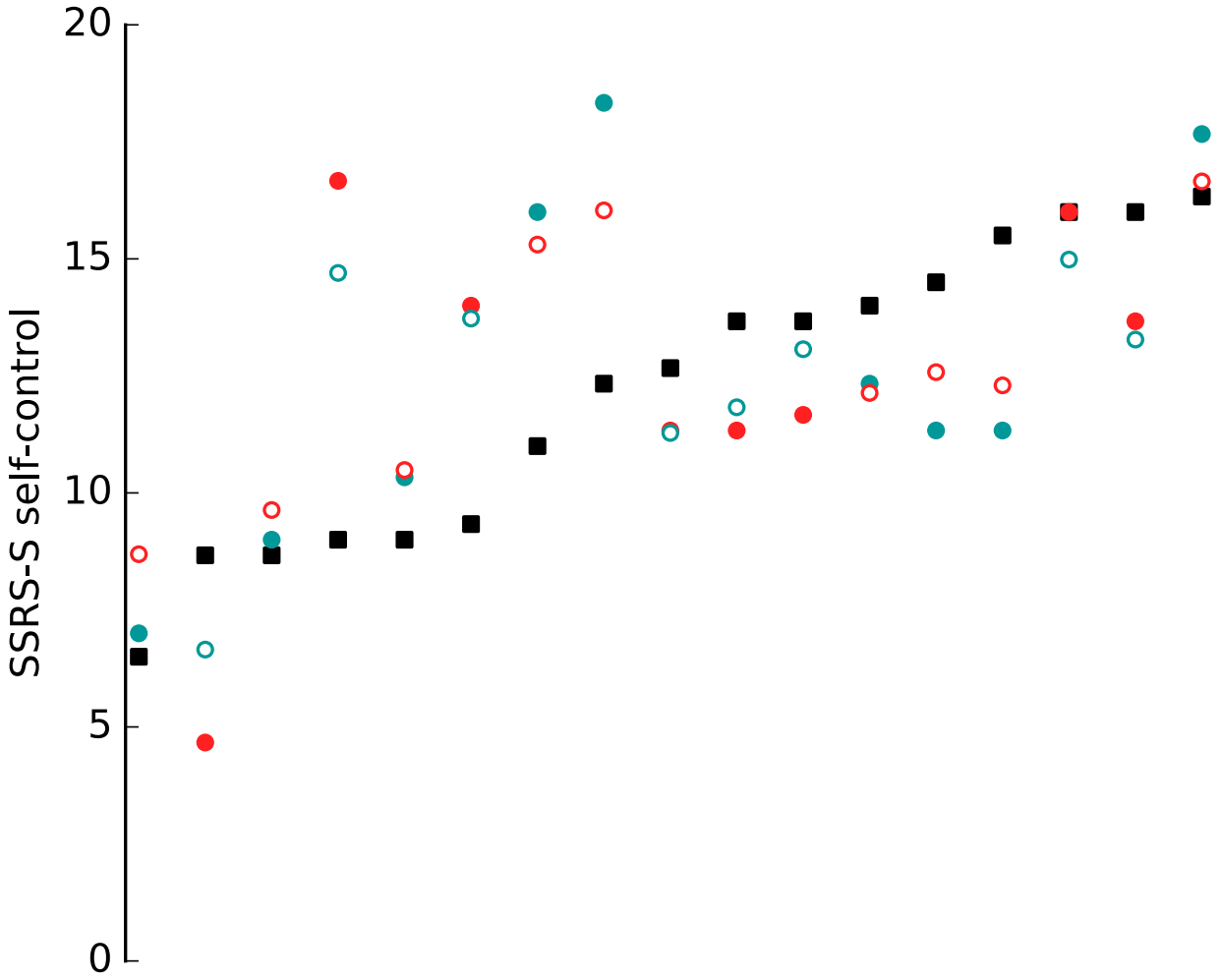

Figure 2. SSRS-S self-control scores before and after treatment. Subjects are sorted by pretreatment score. Solid squares and solid circles show observed pretreatment and posttreatment scores, whereas empty circles show predicted posttreatment scores. Blue circles indicate treatment with SDARI and red circles indicate no treatment other than the passage of time. Predictions have not been rounded.

This analysis, however, only considered the 17 subjects actually included in Spotlight 2007, with their particular combinations of IV values. Also presented here are simulated posttreatment scores for a general population. As with the previous population simulation, the process began by calculating IV importances according to the random forest. The most important IVs were BDI-Y (.145), CBCL thought problems (.105), and CBCL aggression (.092). The pretreatment SSRS-S

self-control (.066) was relatively unimportant, and the treatment condition (.001) was the least important IV. (The treatment having low importance is consistent with the previous analysis and also the fact that the critical models outperformed the baseline models the most for this DV.) The literature was able to provide means and SDs of the BDI-Y (Stapleton et al., 2007) and CBCL subscales (to be exact, percentile ranks for CBCL subscales; Mazefsky, Anderson, Conner, & Minshew, 2010) from samples larger than Spotlight 2007 itself, but not any reports on the relationship between CBCL subscales or between CBCL subscales and the BDI-Y. Hence, there was employed a simple simulation algorithm of drawing BDI-Y, CBCL thought problems percentile rank, and CBCL aggression percentile rank from a multivariate normal distribution with means and SDs from Stapleton et al. (2007) and Mazefsky et al. (2010) and a correlation matrix from Spotlight 2007.

Figure 3 shows the results of 10,000 simulated subjects. The distributions are very similar, indicating, again, a small treatment effect. This analysis, however, indicates that the overall treatment effect is beneficial. Within-subjects, 47% are better off with SDARI, 42% are better off with no treatment, and the remaining 11% have the same self-control score both ways; the mean difference is +.31 in favor of SDARI. Taken together, the two analyses of SSRS-S self-control scores suggest that no treatment would have been better (or roughly equivalent) for the children who happened to have been sampled for Spotlight 2007, but SDARI would be better for the general population.

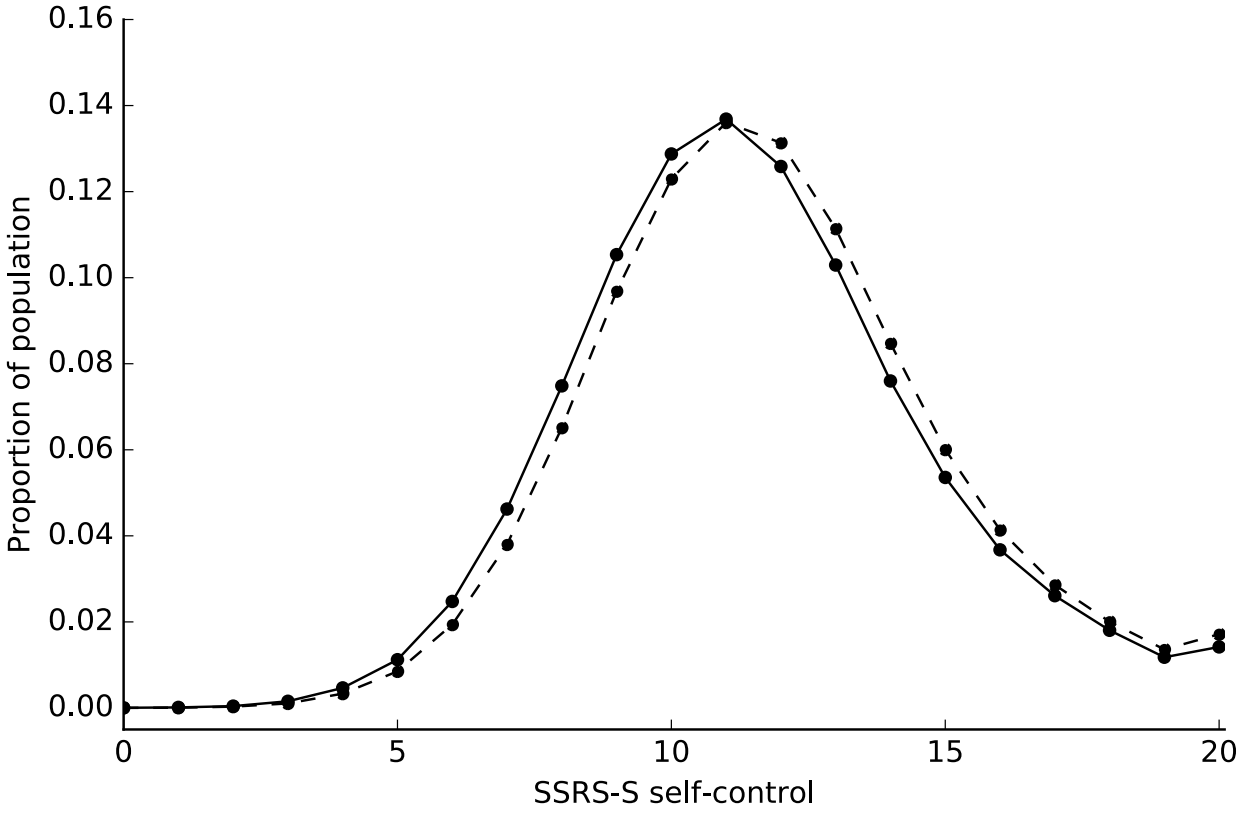

Figure 3. The distribution of simulated SSRS-S self-control scores after no treatment (solid line) or SDARI (dashed line).

SIOS low-level socializing.

It was intended to perform an analysis of counterfactual treatments for SIOS low-level socializing in Knowledge or Performance, as was done for SSRS-S self-control scores in Spotlight 2007. Here the best-performing model was ENet-Interact, so an elastic-net regression model with first-order interactions was fit to the Knowledge or Performance data. However, none of the 9 coefficients that the model chose to be nonzero (out of 1,032 coefficients total) were for the treatment term or an interaction with treatment. (Elastic-net regression, similar to the lasso, may shrink coefficients it decides are of low predictive value all the way to 0.) ENet-Main produced a similar result, with none of the 6 nonzero coefficients (out of 45) being the treatment condition. Hence, these models predicted no effect of treatment condition at all. In other words, the knowledge-training and performance-training interventions seemed to have the same effect on low-level socializing. This is consistent with how OLS-Reduced and ENet-Reduced underperformed Trivial for this DV: the treatment condition, along with pretreatment low-level socializing, had no predictive value for posttreatment low-level socializing. We can infer that SIOS low-level socializing had low retest reliability, suggesting that it could be not have been accurately predicted by anything.

Discussion

Four analyses spanning five datasets and dozens of variables attempted to predict the outcomes of behavioral treatments for autism. Treatment condition, and pretreatment administrations of the same instrument used for the DV, had some predictive value. But other variables had little to no predictive value, contrary to the goal of predicting individual treatment response. Some follow-up analyses were also performed although their conclusions were weakened by the poor results for predictive accuracy. These suggested that if a certain group psychotherapy, SDARI, were applied to general populations of children, treated children would have somewhat more self-control (in their own perceptions) but would externalize somewhat more (in their guardians' perceptions). A follow-up predictive analysis for low-level socializing seemed to be foiled by the low reliability of the instrument.

Potential explanations

Why did the present study fail to obtain good predictive accuracy? The most mundane possible explanation is sample size. The largest sample considered in the primary analyses had 45 subjects, which is large for a social-skills intervention study in autism, but not for typical behavioral research. Particularly troublesome is that in every primary analysis, the number of IVs met or exceeded the number of subjects. One of the advantages of elastic-net regression and random forests is that they can cope with wide datasets such as these, but wide data still makes them less effective than they otherwise would be. Intervention studies may need to expand substantially to investigate the predictive value of a large number of IVs.

Reliability of the DVs may also be to blame. For a surprisingly large number of DVs, both OLS-Reduced and ENet-Reduced failed to substantially improve on Trivial, and in some cases even performed worse. This means the DV had poor retest reliability, because knowing subjects' pretreatment was of little use for predicting their posttreatment scores. Although reliability traditionally receives much less attention in psychometric research than validity, it is an important quality of any test and should not be neglected (Thorndike & Thorndike-Christ, 2010).

A deeper possibility is that something is lacking in the conventional batteries of pretreatment measures given to subjects in outcome studies. So long as treatment can be kept relatively consistent and outcome measures are sufficiently reliable, it stands to reason that differences in treatment outcomes should lie in the many ways that subjects differ before the treatment begins. The tests used in this study, at least, seem to have missed the important differences. The amalgamate and Knowledge or Performance datasets, especially, seemed to lack any correlations that could have been strong enough for good predictive accuracy. New kinds of tests may be necessary, such as examining subjects' immediate reactions to a miniature form of the treatment, similar to the 20-minute training sessions used in Knowledge or Performance. Perhaps a detailed theory of the causes of variability in response to a treatment (including ideas such as Stahmer et al.'s (2011) suggestion that ability to initiate joint attention influences the effect of treatment modality) could be used to develop new measures.

The counterfactual predictions in this study suggested both good news and bad news for SDARI. On the one hand, according to the SSRS-S, the net effect of SDARI is to increase self-control, but on the other, according to the SSRS-P, the net effect is to increase externalizing. One way in which these results might fit together is that they show a potential double edge of giving previously asocial children practice with socializing and encouragement to socialize. Without previous experience with social norms, children may be unfamiliar with, for example, the fine line between assertion and aggression. If typically developing children learn such social norms through practice in socializing at younger ages, SDARI patients must do this learning during and after treatment. We should then expect that so long as benefits of SDARI to social awareness and competence are sustained, negative effects on externalizing will be ameliorated over the course of further development.

It can also be helpful to examine these dual effects as a kind of informant discrepancy. In the study of autism in children, versions of the same instrument for different informants—child, parent, and teacher—can yield systematically different results (Lerner et al., 2012; De Los Reyes, Lerner, Thomas, Daruwala, & Goepel, 2013). In particular, Lerner et al. (2012) observed that parents perceived their children as less socially competent than children perceived themselves, which agrees with the present study's findings for SSRS-S self-control versus SSRS-P externalizing. It makes sense that children with low social skills would overestimate their own abilities, because it is a general finding that incompetence coincides with overconfidence (Kruger & Dunning, 1999). Parents, in turn, might underestimate their children's social skills if they have a general bias to perceive their children as more vulnerable, analogous to Fessler, Holbrook, Pollack, and Hahn-Holbrook's (2014) finding that parents see threatening people as more threatening than non-parents do.

Implications

The findings of low predictive accuracy undermine the goal of using a variety of pretreatment measurements to predict treatment outcomes. It would be difficult to recommend administering a large number of tests to achieve an increase in predictive accuracy of the observed sizes. Like Arfer and Luhmann (2016), the present study poses difficulties for predictivism, which is the idea of focusing research on finding predictively accurate measures and statistical models. As Bone et al. (2015) and Chawarska et al. (2014) also found in their methodologically improved investigations of the themes pursued in Wall et al. (2012a), Wall et al. (2012b), and Macari et al. (2012), predictive accuracy can be elusive. Statistical significance and nonzero association are easy to come by—as stated by Meehl (1990) and illustrated by Standing, Sproule, and Khouzam (1991), "everything correlates to some extent with everything else." (p. 204)—but it may be that there are genuinely few good predictors for a given DV. In the realm of autism in particular, these failures of prediction should temper the present enthusiasm for individualized treatment. However popular individualized treatment, and however laudable the goal of maximizing individual outcomes, it seems unlikely that ad-hoc attempts to adapt treatment to individual circumstances, and thus predict how different patients will be affected by different treatments, will exceed these statistical attempts, given the overall inferiority of human prediction to formal methods (Dawes, Faust, & Meehl, 1989).

This is not to say that predictivism is a lost cause. There have certainly been studies that have succeeded in predicting variables including diagnosis of mental disorder and response to treatment, such as those described in the introduction. These studies suggest that more

technological and physiological methods such as neuroimaging and eye tracking, although more expensive and less convenient than questionnaire measures, may prove their worth by providing predictive value not otherwise available.

For autism researchers and treatment providers, perhaps the most important lesson to be learned from the present study is the difficulty of good individualization of treatment. As just mentioned, the failure to identify good statistical predictors of treatment response does not bode well for clinical prediction of treatment response, either. Hence, trying to individualize treatment may be unhelpful before future research identifies a specific effective strategy. Another important finding was that retest reliability can be poor even for measures that are well-established in autism research and treatment. Since reliability is necessary for validity, and a test with low validity for a given application is not useful for that application, it would be wise to view to take a critical look at such measures. Finally, predictive analysis of the kind conducted in the present study could be a useful way to evaluate autism-related tests. After all, in practice, the value of a test is that it tells the users of the test something they do not already know. Predictive accuracy is a direct measure of how accurate the test's claims about patients are.

Considerations for future work

Another infrequently mentioned issue regarding the choice of measures in research on autism treatment is that autism is not all bad news. In fact, it is not obvious that autism itself should be considered a disease to be cured rather than a non-pathological variation to be accepted (Pellicano & Stears, 2011; Kapp, Gillespie-Lynch, Sherman, & Hutman, 2013). After all, autism is not primarily defined in terms of negative affect (as depression is) or lack of ability (as mental retardation is) but in terms of habits, and the habits in question are not widely regarded as antisocial (as in the case of conduct disorder). Moreover, autism is associated with some cognitive benefits, such as improved performance in some visuospatial reasoning tasks (e.g., Ropar & Mitchell, 2001; Pellicano, Maybery, Durkin, & Maley, 2006) and better memory for the pitch of musical notes (e.g., Bonnel et al., 2003; Heaton, 2003). A small minority of autistic people, called savants, show one or more extraordinary talents, generally including a vast memory for a specialized topic (Treffert, 2009; Howlin, Goode, Hutton, & Rutter, 2009). These autism-related strengths might be useful for predicting treatment outcomes, particularly if the strengths can be leveraged to improve outcomes. It is also worth considering effects on areas of strength as treatment outcomes. If a treatment, in the process of increasing a person's social abilities to a more typical level, can also *decrease* his or her cognitive strengths to a more typical level, we must be careful not to throw the baby out with the bathwater. SDARI has some promise of this considering that it focuses on providing opportunities to practice socializing without trying to force patients to act more typical. Indeed, the finding of increased externalizing but also increased self-perceived self-control suggests that patients are trying out more socially oriented behavior on their own terms.

This said, none of the benefits of autism undo the social costs of autism discussed in the introduction; to simply not provide treatment to anybody would be to ignore a major source of disability. Rather, the benefits of autism are yet another factor that society as a whole, and clinicians and researchers individually, must be mindful of in order to meet the challenges of autism.

References

Arfer, K. B., & Luhmann, C. C. (2015). The predictive accuracy of intertemporal-choice models. *British Journal of Mathematical and Statistical Psychology, 68*(2), 326–341. doi:10.1111/bmsp.12049. Retrieved from http://arfer.net/projects/builder/paper

Arfer, K. B., & Luhmann, C. C. (2016). *Time-preference tests fail to predict self-control behavior*. Retrieved from http://arfer.net/projects/rickrack/paper

Attwood, T. (2000). Strategies for improving the social integration of children with Asperger syndrome. *Autism , 4*(1), 85–100.

Autism and Developmental Disabilities Monitoring Network. (2014). Prevalence of autism spectrum disorder among children aged 8 years—Autism and Developmental Disabilities Monitoring Network, 11 sites, United States, 2010. *Morbidity and Mortality Weekly Report, 63*(2), 1–21. Retrieved from http://www.cdc.gov/mmwr/preview/mmwrhtml/ss6302a1.htm

Ball, T. M., Stein, M. B., Ramsawh, H. J., Campbell-Sills, L., & Paulus, M. P. (2014). Single-subject anxiety treatment outcome prediction using functional neuroimaging. *Neuropsychopharmacology, 39*(5), 1254–1261. doi:10.1038/npp.2013.328

Baron-Cohen, S. (2002). The extreme male brain theory of autism. *Trends in Cognitive Sciences, 6*(6), 248–254. doi:10.1016/S1364-6613(02)01904-6

Baron-Cohen, S., Wheelwright, S., & Jolliffe, T. (1997). Is there a "language of the eyes"? Evidence from normal adults, and adults with autism or Asperger Syndrome. *Visual Cognition, 4*(3), 311–331. doi:10.1080/713756761

Berument, S. K., Rutter, M., Lord, C., Pickles, A., & Bailey, A. (1999). Autism Screening Questionnaire: Diagnostic validity. *British Journal of Psychiatry, 175*, 444–451. doi:10.1192/bjp.175.5.444

Bone, D., Goodwin, M. S., Black, M. P., Lee, C.-C., Audhkhasi, K., & Narayanan, S. (2015). Applying machine learning to facilitate autism diagnostics: Pitfalls and promises. *Journal of Autism and Developmental Disorders, 45*(5), 1121–1136. doi:10.1007/s10803-014-2268-6

Bonnel, A., Mottron, L., Peretz, I., Trudel, M., Gallun, E., & Bonnel, A.-M. (2003). Enhanced pitch sensitivity in individuals with autism: A signal detection analysis. *Journal of Cognitive Neuroscience, 15*(2), 226–235. doi:10.1162/089892903321208169

Bottema-Beutel, K., Mullins, T. S., Harvey, M. N., Gustafson, J. R., & Carter, E. W. (2016). Avoiding the "brick wall of awkward": Perspectives of youth with autism spectrum disorder on social-focused intervention practices. *Autism, 20*(2), 196–206. doi:10.1177/1362361315574888

Breiman, L. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth International Group. ISBN 978-0-534-98054-2.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Campbell, D. J., Shic, F., Macari, S., & Chawarska, K. (2014). Gaze response to dyadic bids at 2 years related to outcomes at 3 years in autism spectrum disorders: A subtyping analysis. *Journal of Autism and Developmental Disorders, 44*(2), 431–442. doi:10.1007/s10803-013-1885-9

Cavagnaro, A. T. (2009). *Autism spectrum disorders: Changes in the California caseload*. California Health and Human Services Agency. Retrieved from http://www.dds.ca.gov/autism/docs/AutismReport_2007.pdf

Chawarska, K., Shic, F., Macari, S., Campbell, D. J., Brian, J., Landa, R., … Bryson, S. (2014). 18-month predictors of later outcomes in younger siblings of children with autism spectrum disorder: A Baby Siblings Research Consortium study. *Journal of the American Academy of Child and Adolescent Psychiatry, 53*(12), 1317–1327. doi:10.1016/j.jaac.2014.09.015

Choque Olsson, N., Rautio, D., Asztalos, J., Stoetzer, U., & Bölte, S. (2016). Social skills group training in high-functioning autism: A qualitative responder study. *Autism*. Advance online publication. doi:10.1177/1362361315621885

Constantino, J. N., Davis, S. A., Todd, R. D., Schindler, M. K., Gross, M. M., Brophy, S. L., … Reich, W. (2003). Validation of a brief quantitative measure of autistic traits: Comparison of the Social Responsiveness Scale with the Autism Diagnostic Interview—Revised. *Journal of Autism and Developmental Disorders, 33*(4), 427–433. doi:10.1023/A:1025014929212

Costafreda, S. G., Khanna, A., Mourao-Miranda, J., & Fu, C. H. Y. (2009). Neural correlates of sad faces predict clinical remission to cognitive behavioural therapy in depression. *NeuroReport, 20*(7), 637–641. doi:10.1097/WNR.0b013e3283294159

Crippa, A., Salvatore, C., Perego, P., Forti, S., Nobile, M., Molteni, M., & Castiglioni, I. (2015). Use of machine learning to identify children with autism and their motor abnormalities. *Journal of Autism and Developmental Disorders, 45*(7), 2146–2156. doi:10.1007/s10803-015-2379-8

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243*(4899), 1668–1674. doi:10.1126/science.2648573

Dawson, G., Rogers, S., Munson, J., Smith, M., Winter, J., Greenson, J., … Varley, J. (2010). Randomized, controlled trial of an intervention for toddlers with autism: The Early Start Denver Model. *Pediatrics, 125*(1), e17–e23. doi:10.1542/peds.2009-0958

De Los Reyes, A., Lerner, M. D., Thomas, S. A., Daruwala, S., & Goepel, K. (2013). Discrepancies between parent and adolescent beliefs about daily life topics and performance on an emotion recognition task. *Journal of Abnormal Child Psychology, 41*(6), 971–982. doi:10.1007/s10802-013-9733-0

Demaray, M. K., Ruffalo, S. L., Carlson, J., Busse, R. T., Olson, A. E., McManus, S. M., & Leventhal, A. (1995). Social skills assessment: A comparative evaluation of six published rating scales. *School Psychology Review, 24*(4), 648–671.

Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning, 29*(1, 3), 103–130. doi:10.1023/A:1007413511361

Duke, M. P., & Nowicki, S., Jr. (2005). The Emory Dyssemia Index. In V. L. Manusov (Ed.), *The sourcebook of nonverbal measures: Going beyond words* (pp. 35–46). Mahwah, NJ: Lawrence Erlbaum. ISBN 978-0-8058-4746-8.

Elsabbagh, M., Divan, G., Koh, Y.-J., Kim, Y. S., Kauchali, S., Marcín, C., … Fombonne, E. (2012). Global prevalence of autism and other pervasive developmental disorders. *Autism Research, 5*(3), 160–179. doi:10.1002/aur.239

Estes, A., Munson, J., Rogers, S. J., Greenson, J., Winter, J., & Dawson, G. (2015). Long-term outcomes of early intervention in 6-year-old children with autism spectrum disorder. *Journal of the American Academy of Child and Adolescent Psychiatry, 54*(7), 580–587. doi:10.1016/j.jaac.2015.04.005

Fessler, D. M. T., Holbrook, C., Pollack, J. S., & Hahn-Holbrook, J. (2014). Stranger danger: Parenthood increases the envisioned bodily formidability of menacing men. *Evolution and Human Behavior, 35*(2), 109–117. doi:10.1016/j.evolhumbehav.2013.11.004

Fisch, G. S. (2012). Nosology and epidemiology in autism: Classification counts. *American Journal of Medical Genetics, 160*(2), 91–103. doi:10.1002/ajmg.c.31325

Freund, Y., & Mason, L. (1999). The alternating decision tree learning algorithm. In *Proceedings of the Sixteenth International Conference on Machine Learning* (pp. 124–133). Retrieved from http://cseweb.ucsd.edu/~yfreund/papers/atrees.pdf

Ganz, M. L. (2007). The lifetime distribution of the incremental societal costs of autism. *Archives of Pediatrics and Adolescent Medicine, 161*(4), 343–349. doi:10.1001/archpedi.161.4.343

Geisser, S. (1993). *Predictive inference: An introduction*. New York, NY: Chapman & Hall. ISBN 978-0-412-03471-8.

Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine, 27*(15), 2865–2873. doi:10.1002/sim.3107

Goldstein, A. P., & McGinnis, E. (1997). *Skillstreaming the adolescent: New strategies and perspectives for teaching prosocial skills*. Champaign, IL: Research Press. ISBN 978-0-87822-369-5.

Gong, Q., Wu, Q., Scarpazza, C., Lui, S., Jia, Z., Marquand, A., … Mechelli, A. (2011). Prognostic prediction of therapeutic response in depression using high-field MR imaging. *NeuroImage, 55*(4), 1497–1503. doi:10.1016/j.neuroimage.2010.11.079

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York, NY: Springer. Retrieved from http://www-stat.stanford.edu/~tibs/ElemStatLearn

Heaton, P. (2003). Pitch memory, labelling and disembedding in autism. *Journal of Child Psychology and Psychiatry, 44*(4), 543–551. doi:10.1111/1469-7610.00143

Hitchcock, C., & Sober, E. (2004). Prediction versus accommodation and the risk of overfitting. *British Journal for the Philosophy of Science , 55*(1), 1–34. Retrieved from http://www.researchgate.net/profile/Elliott_Sober/publication/31226577/links/5416de1c0cf2fa878ad430ac.pdf

Howlin, P., Goode, S., Hutton, J., & Rutter, M. (2009). Savant skills in autism: Psychometric approaches and parental reports. *Philosophical Transactions of the Royal Society B, 364*(1522), 1359–1367. doi:10.1098/rstb.2008.0328

Hutchins, T. L., Bonazinga, L. A., Prelock, P. A., & Taylor, R. S. (2008). Beyond false beliefs: The development and psychometric evaluation of the Perceptions of Children's Theory of Mind Measure—Experimental Version (PCToMM-E). *Journal of Autism and*

Developmental Disorders, 38(1), 143–155. doi:10.1007/s10803-007-0377-1

Ivanova, M. Y., Achenbach, T. M., Dumenci, L., Rescorla, L. A., Almqvist, F., Weintraub, S., … Verhulst, F. C. (2007). Testing the 8-syndrome structure of the Child Behavior Checklist in 30 societies. *Journal of Clinical Child and Adolescent Psychology, 36*(3), 405–417. doi:10.1080/15374410701444363

Just, M. A., Cherkassky, V. L., Buchweitz, A., Keller, T. A., & Mitchell, T. M. (2014). Identifying autism from neural representations of social interactions: Neurocognitive markers of autism. *PLOS ONE*. doi:10.1371/journal.pone.0113879

Järbrink, K., & Knapp, M. (2001). The economic impact of autism in Britain. *Autism, 5*(1), 7–22. doi:10.1177/1362361301005001002

Kaland, N., Møller-Nielsen, A., Callesen, K., Mortensen, E. L., Gottlieb, D., & Smith, L. (2002). A new "advanced" test of theory of mind: Evidence from children and adolescents with Asperger syndrome. *Journal of Child Psychology and Psychiatry, 43*(4), 517–528. doi:10.1111/1469-7610.00042

Kapp, S. K., Gillespie-Lynch, K., Sherman, L. E., & Hutman, T. (2013). Deficit, difference, or both? Autism and neurodiversity. *Developmental Psychology, 49*(1), 59–71. doi:10.1037/a0028353

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121–1134. doi:10.1037/0022-3514.77.6.1121

La Greca, A. M., & Lopez, N. (1998). Social anxiety among adolescents: Linkages with peer relations and friendships. *Journal of Abnormal Child Psychology, 26*(2), 83–94. doi:10.1023/A:1022684520514

La Greca, A. M., & Stone, W. L. (1993). Social Anxiety Scale for Children—Revised: Factor structure and concurrent validity. *Journal of Clinical Child Psychology, 22*(1), 17–27. doi:10.1207/s15374424jccp2201_2

Lai, M.-C., Lombardo, M. V., & Baron-Cohen, S. (2014). Autism. *Lancet, 383*(9920), 896–910. doi:10.1016/S0140-6736(13)61539-1

Leigh, J. P., & Du, J. (2015). Brief report: Forecasting the economic burden of autism in 2015 and 2025 in the United States. *Journal of Autism and Developmental Disorders*. doi:10.1007/s10803-015-2521-7

Lerner, M. D., Calhoun, C. D., Mikami, A. Y., & De Los Reyes, A. (2012). Understanding parent–child social informant discrepancy in youth with high functioning autism spectrum disorders. *Journal of Autism and Developmental Disorders, 42*(12), 2680–2692. doi:10.1007/s10803-012-1525-9

Lerner, M. D., & Mikami, A. Y. (2012). A preliminary randomized controlled trial of two social skills interventions for youth with high-functioning autism spectrum disorders. *Focus on Autism and Other Developmental Disabilities, 27*(3), 147–157. doi:10.1177/1088357612450613

Lerner, M. D., Mikami, A. Y., & Levine, K. (2011). Socio-dramatic affective-relational intervention for adolescents with Asperger syndrome & high functioning autism: Pilot study. *Autism, 15*(1), 21–42. doi:10.1177/1362361309353613

Lerner, M. D., White, S. W., & McPartland, J. C. (2012). Mechanisms of change in psychosocial

interventions for autism spectrum disorders. *Dialogues in Clinical Neuroscience, 14*(3), 307–318.

Li, Y., Wileyto, E. P., & Heitjan, D. F. (2011). Prediction of individual long-term outcomes in smoking cessation trials using frailty models. *Biometrics, 67*(4), 1321–1329. doi:10.1111/j.1541-0420.2011.01578.x

Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Jr., Leventhal, B. L., DiLavore, P. C., … Rutter, M. (2000). The Autism Diagnostic Observation Schedule—Generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of Autism and Developmental Disorders, 30*(3), 205–223. doi:10.1023/A:1005592401947

Macari, S. L., Campbell, D., Gengoux, G. W., Saulnier, C. A., Klin, A. J., & Chawarska, K. (2012). Predicting developmental status from 12 to 24 months in infants at risk for autism spectrum disorder: A preliminary report. *Journal of Autism and Developmental Disorders, 42*(12), 2636–2647. doi:10.1007/s10803-012-1521-0

Maglione, M. A., Gans, D., Das, L., Timbie, J., & Kasari, C. (2012). Nonmedical interventions for children with ASD: Recommended guidelines and further research needs. *Pediatrics, 130*(Suppl. 2), S169–S178. doi:10.1542/peds.2012-0900O

Mazefsky, C. A., Anderson, R., Conner, C. M., & Minshew, N. (2010). Child behavior checklist scores for school-aged children with autism: Preliminary evidence of patterns suggesting the need for referral. *Journal of Psychopathology and Behavioral Assessment, 33*(1), 31–37. doi:10.1007/s10862-010-9198-1

Mazumder, R., Hastie, T., & Tibshirani, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research, 11*, 2287–2322. Retrieved from http://www.jmlr.org/papers/v11/mazumder10a.html

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*(4), 359–372.

McGrath, J., Saha, S., Chant, D., & Welham, J. (2008). Schizophrenia: A concise overview of incidence, prevalence, and mortality. *Epidemiologic Reviews, 30*, 67–76. doi:10.1093/epirev/mxn001

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 66*(1), 195–244. doi:10.2466/PR0.66.1.195-244

Miller, A., Vernon, T., Wu, V., & Russo, K. (2014). Social skill group interventions for adolescents with autism spectrum disorders: A systematic review. *Review Journal of Autism and Developmental Disorders, 1*(4), 254–265. doi:10.1007/s40489-014-0017-6

Morgan, G. A., Wang, J., Barrett, K. C., Liao, H.-F., Wang, P.-J., Huang, S.-Y., & Jozsa, K. (2015). *The revised Dimensions of Mastery Questionnaire (DMQ 18)*. Retrieved from http://web.archive.org/web/20160505140617/https://1-s-sites.googlegroups.com/a/rams.colostate.edu/georgemorgan/mastery-motivation/DMQ18Manual.pdf

Mouchiroud, C., & Lubart, T. (2002). Social creativity: A cross-sectional study of 6- to 11-year-old children. *International Journal of Behavioral Development, 26*(1), 60–69.

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology, 44*(1), 190–204. doi:10.1006/jmps.1999.1283

Nowicki, S., Jr., & Carton, J. (1993). The measurement of emotional intensity from facial expressions. *Journal of Social Psychology, 133*(5), 749–750. doi:10.1080/00224545.1993.9713934

Nowicki, S., Jr., & Duke, M. P. (1994). Individual differences in the nonverbal communication of affect: The Diagnostic Analysis of Nonverbal Accuracy Scale. *Journal of Nonverbal Behavior, 18*(1), 9–35. doi:10.1007/BF02169077

Otero, T. L., Schatz, R. B., Merrill, A. C., & Bellini, S. (2015). Social skills training for youth with autism spectrum disorders: A follow-up. *Child and Adolescent Psychiatric Clinics of North America, 24*(1), 99–115. doi:10.1016/j.chc.2014.09.002

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Braun, M. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Pellicano, E., Maybery, M., Durkin, K., & Maley, A. (2006). Multiple cognitive capabilities/deficits in children with an autism spectrum disorder: "Weak" central coherence and its relationship to theory of mind and executive control. *Development and Psychopathology, 18*(1), 77–98. doi:10.1017/S0954579406060056. Retrieved from http://www.researchgate.net/profile/Murray_Maybery/publication/7297570/links/53f2848d0cf2f2c3e8014b7a.pdf

Pellicano, E., & Stears, M. (2011). Bridging autism, science and society: Moving toward an ethically informed approach to autism research. *Autism Research, 4*(4), 271–282. doi:10.1002/aur.201

Ropar, D., & Mitchell, P. (2001). Susceptibility to illusions and performance on visuospatial tasks in individuals with autism. *Journal of Child Psychology and Psychiatry, 42*(4), 539–549. doi:10.1111/1469-7610.00748

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*(6), 417–438. doi:10.1352/0895-8017(2005)110[417:IBTFCW]2.0.CO;2

Scanlon, E. M., & Ollendick, T. H. (1985). Children's assertive behavior: The reliability and validity of three self-report measures. *Child and Family Behavior Therapy, 7*(3), 9–21. doi:10.1300/J019v07n03_02

Seida, J. K., Ospina, M. B., Karkhaneh, M., Hartling, L., Smith, V., & Clark, B. (2009). Systematic reviews of psychosocial interventions for autism: An umbrella review. *Developmental Medicine and Child Neurology, 51*(2), 95–104. doi:10.1111/j.1469-8749.2008.03211.x

Smith, R. S., & Sharp, J. (2013). Fascination and isolation: A grounded theory exploration of unusual sensory experiences in adults with Asperger syndrome. *Journal of Autism and Developmental Disorders, 43*(4), 891–910. doi:10.1007/s10803-012-1633-6

Smith, T., Groen, A. D., & Wynn, J. W. (2000). Randomized trial of intensive early intervention for children with pervasive developmental disorder. *American Journal on Mental Retardation, 105*(4), 269–285. doi:10.1352/0895-8017(2000)105<0269:RTOIEI>2.0.CO;2

Soorya, L. V., Siper, P. M., Beck, T., Soffes, S., Halpern, D., Gorenstein, M., … Wang, A. T. (2014). Randomized comparative trial of a social cognitive skills group for children with

autism spectrum disorder. *Journal of the American Academy of Child and Adolescent Psychiatry, 54*(3), 208–216. doi:10.1016/j.jaac.2014.12.005

Stahmer, A. C., Schreibman, L., & Cunningham, A. B. (2011). Toward a technology of treatment individualization for young children with autism spectrum disorders. *Brain Research, 1380*, 229–239. doi:10.1016/j.brainres.2010.09.043

Standing, L., Sproule, R., & Khouzam, N. (1991). Empirical statistics: IV. Illustrating Meehl's sixth law of soft psychology: Everything correlates with everything. *Psychological Reports, 69*(1), 123–126. doi:10.2466/PR0.69.5.123-126

Stapleton, L. M., Sander, J. B., & Stark, K. D. (2007). Psychometric properties of the Beck Depression Inventory for Youth in a sample of girls. *Psychological Assessment, 19*(2), 230–235. doi:10.1037/1040-3590.19.2.230

Steyerberg, E. W., Harrell, F. E., Jr., Borsboom, G. J., Eijkemans, M. J., Vergouwe, Y., & Habbema, J. D. (2001). Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. *Journal of Clinical Epidemiology, 54*(8), 774–781. doi:10.1016/S0895-4356(01)00341-9

Thorndike, R. M., & Thorndike-Christ, T. (2010). *Measurement and evaluation in psychology and education* (8th ed.). Boston, MA: Pearson. ISBN 0-13-240397-8.

Treffert, D. A. (2009). The savant syndrome: An extraordinary condition. A synopsis: Past, present, future. *Philosophical Transactions of the Royal Society B, 364*(1522), 1351–1357. doi:10.1098/rstb.2008.0326

United States Census Bureau. (2014). *2014 American Community Survey 1-year estimates*. Retrieved from http://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_14_1YR_S0101

Wall, D. P., Dally, R., Luyster, R., Jung, J.-Y., & DeLuca, T. F. (2012a). Use of artificial intelligence to shorten the behavioral diagnosis of autism. *PLOS ONE*. doi:10.1371/journal.pone.0043855

Wall, D. P., Kosmicki, J., DeLuca, T. F., Harstad, E., & Fusaro, V. A. (2012b). Use of machine learning to shorten observation-based screening and diagnosis of autism. *Translational Psychiatry, 2*, e100. doi:10.1038/tp.2012.10

Wasserman, L. (2004). *All of statistics: A concise course in statistical inference*. New York, NY: Springer. ISBN 978-0-387-40272-7.

Watson, J. B. (1913). Psychology as the behaviorist views it. *Psychological Review, 20*(2), 158–177. doi:10.1037/h0074428

Williams, P. E., Weiss, L. G., & Rolfhus, E. (2003). *WISC–IV technical report #1: Theoretical model and test blueprint*. Psychological Corporation. Retrieved from https://web.archive.org/web/20131204041019/http://images.pearsonclinical.com/images/pdf/wisciv/WISCIVTechReport1.pdf

Wong, C., Odom, S. L., Hume, K. A., Cox, A. W., Fettig, A., Kucharczyk, S., … Schultz, T. R. (2015). Evidence-based practices for children, youth, and young adults with autism spectrum disorder: A comprehensive review. *Journal of Autism and Developmental Disorders, 45*(7), 1951–1966. doi:10.1007/s10803-014-2351-z

Woolfenden, S., Sarkozy, V., Ridley, G., Coory, M., & Williams, K. (2012). A systematic review of two outcomes in autism spectrum disorder—epilepsy and mortality. *Developmental*

Medicine and Child Neurology, 54(4), 306–312. doi:10.1111/j.1469-8749.2012.04223.x

Yoder, P., & Stone, W. L. (2006). Randomized comparison of two communication interventions for preschoolers with autism spectrum disorders. *Journal of Consulting and Clinical Psychology, 74*(3), 426–435. doi:10.1037/0022-006X.74.3.426

Zhang, D., Wang, Y., Zhou, L., Yuan, H., & Shen, D. (2011). Multimodal classification of Alzheimer's disease and mild cognitive impairment. *NeuroImage, 55*(3), 856–867. doi:10.1016/j.neuroimage.2011.01.008