Published in final edited form as:

Am J Intellect Dev Disabil. 2013 January ; 118(1): 3–15. doi:10.1352/1944-7558-118.1.3.

Establishing Equivalence: Methodological Progress in Group-Matching Design and Analysis

Sara T. Kover and

University of Wisconsin-Madison, Waisman Center, 1500 Highland Avenue, Madison, WI 53705

Amy K. Atwood

University of Wisconsin-Madison

Abstract

This methodological review draws attention to the challenges faced by intellectual and developmental disabilities researchers in the appropriate design and analysis of group comparison studies. We provide a brief overview of matching methodologies in the field, emphasizing group-matching designs utilized in behavioral research on cognition and language in neurodevelopmental disorders, including autism spectrum disorder, fragile X syndrome, Down syndrome, and Williams syndrome. The limitations of relying on p -values to establish group equivalence are discussed in the context of other existing methods: equivalence tests, propensity scores, and regression-based analyses. Our primary recommendation for advancing research on intellectual and developmental disabilities is the use of descriptive indices of adequate group matching: effect sizes (i.e., standardized mean differences) and variance ratios.

Keywords

matching; equivalence; methodology; autism; comparison; neurodevelopmental disorders

With the ultimate goal of understanding their causal effects on development, much of behavioral research on intellectual and developmental disabilities (IDDs) is designed to (1) characterize phenotypic strengths and weaknesses in behavior and cognition and/or (2) identify syndrome-specific aspects of these profiles. Such aims are often addressed with group-matching designs, in which statistical comparisons between nonrandomized groups (e.g., autism spectrum disorder [ASD] versus typical development) provide the basis for conclusions. Despite the considerable attention matching has received (Abbeduto, 2010; Burack, 2004), methodological issues in group matching remain at the forefront of concerns regarding the progress of behavioral research on neurodevelopmental disorders (Beeghly, 2006; Eigsti, de Marchena, Schuh, & Kelley, 2011).

The purpose of this article is to introduce methodological improvements to group-matching designs frequently used in IDD research. To that end, we discuss the pitfalls of common group-matching strategies and suggest metrics for establishing adequate group equivalence that are not novel, but are new to the field: effect sizes and variance ratios. Because our

primary goal is to provide a foundation from which informed decisions on research design, analysis, and interpretation can be made, we highlight several other study designs worthy of consideration. We conclude by emphasizing the need for thoughtful research questions and responsible use of equivalence thresholds.

Challenges of Group Matching in IDD Research

Frameworks for Causality

The ability to draw conclusions about causality has traditionally hinged upon random assignment of participants to the target group (e.g., treatment, intervention, diagnosis—in our case) and comparison group. Properly implemented, random assignment allows estimation of causal effects because it ensures, in the long run, that any differences between the target and comparison groups (i.e., bias or selection bias) aside from group assignment prior to the study are due to chance. One approach to causality, the Rubin Causal Model, defines the causal effect—that is, the effect of a manipulable treatment—in terms of potential outcomes: what the outcome would have been for participants in the comparison group had they received the treatment and what the outcome would have been for those in the treatment group had they not received it (Holland, 1986; Rubin, 1974). In quasi-experimental designs (e.g., regression discontinuity, interrupted time series), it is possible to test hypotheses about the effects of causes without random assignment. A nonequivalent control group design is one that seeks to remove the bias associated with nonrandom assignment by matching the target and comparison groups to establish equivalence, or balance (Shadish, Cook, & Campbell, 2002).

Methods in IDD Research

Although IDDs are attributable to neurodevelopmental disorders, those disorders can scarcely be considered manipulable causes. Research on IDDs is further constrained by ethical parameters (e.g., inability to randomly assign the circumstances that lead to a diagnosis of fetal alcohol spectrum disorder) and relatively small samples due to low prevalence. As such, the use of more desirable techniques, such as random assignment or sophisticated matching that relies on large datasets, is precluded. Instead, in the simplest and perhaps most common group-matching design in the field, two groups composed of participants with preexisting diagnoses are matched on a single variable, such as nonverbal cognitive ability, and then compared on some dependent variable of interest, such as vocabulary ability. These groups are selected in such a manner as to presume they are equivalent on a dimension of ability thought to be relevant to the dependent variable of interest. Differences between groups on the dependent variable are taken to indicate strengths or weaknesses on the construct of interest relative to the matching construct. How to select constructs and variables on which to match is discussed elsewhere and is beyond the current scope (see Burack, Iarocci, Bowler, & Motttron, 2002; Mervis & Klein-Tasman, 2004; Mervis & Robinson, 1999; Strauss, 2001). We focus here on a specific aspect of matching: establishing when groups are equivalent.

A customary group-matching procedure is to iteratively exclude participants from one or both groups until an independent samples *t*-test of the group mean difference on the

matching variable yields a sufficiently high p -value, showing that the groups do not significantly differ. This process is achieved by first testing the difference between the groups on the matching variable. For example, a hypothetical target group of 30 participants with a mean score of 60.10 on the matching variable would not be considered matched to a comparison group of 30 participants with a mean score of 71.70 because the p -value for the t -test on the matching variable is less than .05 (hypothetical data are given in the Appendix). Two matched groups might be attained by next removing all participants outside of the overlapping range of the groups or according to some other criterion, and then testing the group difference again, usually yielding iteratively higher p -values (Mervis & John, 2008). This procedure might be repeated an unreported number of times by a researcher and usually yields higher p -values as participants are removed.

P -value Thresholds

The most persuasive standard in the field for group matching has been generated by Mervis and colleagues (Mervis & Klein-Tasman, 2004, Mervis & Robinson, 1999), who drew important attention to the matching procedures used to study individuals with IDD. Mervis and colleagues highlighted that accepting groups as matched when a t -test on the matching variable yields a p -value greater than .05 is not sufficient (Mervis & Klein-Tasman, 2004; Mervis & Robinson, 1999). As such, they substantially improved upon the common practice of accepting the null hypothesis that population means are equal given any nonsignificant p -value. Mervis and Klein-Tasman (2004) suggested that when considering a p -value threshold for matching groups, "...it is important to show that the group distributions on the matching variable overlap strongly, as evidenced, we suggest, by a p -level of at least .50 on the test of mean differences for the control variable(s)," (p. 9). While p -values below .05 are taken as clear indication of a group difference, Mervis and colleagues (2004, 1999) proposed that p -values between .20 and .50 are ambiguous; p -values of .50 and above are sufficient evidence of equivalence. The .50 p -value threshold was based on Frick's (1995) "good-effort criterion" for accepting the null hypothesis, which included p -value thresholds in combination with a small effect.

According to Mervis' guideline, groups might be considered matched on a measure of cognitive ability only when the p -value for the test of the group difference on the matching variable is greater than or equal to .50. In our hypothetical example, a subset of participants ($n = 20$ from each group) could be selected to improve the overlap of the groups on the matching variable by removing the lowest scoring participants of the target group and the highest scoring participants of the comparison group, yielding a mean of 68.10 for the target group and a mean of 67.10 for the comparison group. The t -test on the matching variable for these subgroups is not significant, but instead gives a p -value of .55, which is a value that might be taken as evidence that the groups are matched.

The Trouble with P -value Thresholds

Mervis and colleagues (2004, 1999) were correct in emphasizing that groups ought not be considered matched solely on the basis of a p -value greater than .05. Most importantly, their recommendations increased awareness in the field that some p -values should lead to a

conclusion of *failing to reject* the null hypothesis that the population means are equal. Nonetheless the p -value threshold proposed by Mervis et al. (2004, 1999) and Frick (1995) is not without limitations (Edgell, 1995; Imai, King, & Stuart, 2008). Difficulties hinge around the interpretation of p -values and the role of power in hypothesis testing.

Interpretation of a P -value

A p -value is defined as the probability of observing the sampled data or data more extreme, given that the null hypothesis is true. The hypotheses in question in the traditional matching procedure are $H_0: \mu = 0$ and $H_1: \mu \neq 0$, where μ is the population mean difference on the matching variable. The p -value represents the probability of observing the sample mean difference (or one more extreme) when the population mean difference is zero. When there is no difference in the population on the matching variable, one would expect a p -value to be less than or equal to .05 exactly 5% of the time. As such, using $p > .05$ as a threshold for declaring groups to be matched will result in groups being considered matched 95% of the time when there is no true population mean difference. This can be seen in the first panel of Figure 1. Likewise, one would expect a p -value to be less than or equal to .20 exactly 20% of the time; a p -value of less than or equal to .50 would be expected 50% of the time when the populations are equivalent. Thus, using a matching threshold of $p \leq .50$, one would conclude that the group samples are matched just 50% of the time on average when the groups are truly matched in the population.

Failure to reject $H_0: \mu = 0$ does *not* allow one to conclude that the groups come from populations with the same mean because a p -value denotes nothing about the probability of the truth of H_0 given the observed data (Schervish, 1996). Null hypothesis significance testing allows for rejecting or failing to reject the null hypothesis; the option of accepting the null hypothesis simply does not exist. Thus, observing a p -value of .50 leads only to a conclusion that the groups are not significantly different (i.e., a failure to reject the null hypothesis).

Power and P -values

It is possible to observe a large p -value due to a lack of effect, due to a small effect, and/or due to the lack of power to detect that effect because of a small sample size. Eliminating participants until a test of the mean difference results in a large enough p -value may decrease the difference between the observed sample means for the matching variable, but also decreases the power (due to the decreasing sample size) to detect any effect at all, including on the dependent variable of interest. In other words, increasing p -values on the matching variable may have less to do with achieving equivalence between groups and more to do with a reduction in power, particularly when sample sizes are initially small. Mervis and John (2008) provide a substantive example of this for a sample of participants with Williams syndrome.

Impact of sample size on p -value threshold matching—To further illustrate these difficulties, we simulated the process of sampling groups of various sizes from populations with known mean differences in a Monte Carlo simulation using R version 2.10.1 (R Development Core Team, 2010). Over 10,000 iterations, we tracked the frequency with

which various p -value matching thresholds resulted in concluding that the groups were matched. We defined groups to be matched when a t -test comparing group means on the matching variable resulted in a p -value greater than p -value thresholds of .05, .20, and .50. We based the matching variable on a standardized assessment of cognitive ability, such as IQ ($M = 100$, $SD = 15$), because it is often used in this context and is readily interpretable. The true difference in populations was calculated in terms of the standardized mean difference effect size, ranging from 0 (i.e., no difference) to 0.5 (i.e., a medium effect size using Cohen's [1988] guidelines).

The impact of sample size on the p -value threshold method for determining when groups are matched becomes apparent when the population mean difference is truly greater than zero. As seen in Figure 1, when the population mean difference is a medium effect of $d = 0.50$, using a threshold of $p = .50$, one concludes that the groups are matched between 5% and 35% of the time across sample sizes of 10 to 50. This notable discrepancy in rate of meeting the p -value threshold and concluding that the groups are matched is due to variability in sample size alone.

Summary—Inferential statistics should not be used in isolation for establishing equivalence because the results of a t -test hinge on both the observed mean difference between groups and the statistical power of the test, which has a direct relation to sample sizes—dropping participants reduces power and increases a p -value without respect to the mean difference (Imai et al., 2008).

Improved Equivalence Thresholds: Recommendations

Descriptive Statistics for Group Matching

Contemporary reviews of matching methodologies highlight descriptive statistics pre- and post-matching as an alternative to inferential statistics for determining the adequacy of group equivalence (Steiner & Cook, 2012; Stuart, 2010). As the basis for equivalence thresholds—a term from Stegner et al. (1996)—for IDD research, we suggest two descriptive metrics: effect sizes (i.e., standardized mean differences) and variance ratios. These metrics are used widely in quasi-experimental designs with propensity score analysis, which we describe below. Importantly, effect sizes and variance ratios yield interpretable estimates of group matching adequacy and reduce the influence of sample size (Breugh & Arnold, 2007; Imai et al., 2008).

Sometimes referred to as standardized bias, standardized mean differences are a simple and effective index of matching (Rosenbaum & Rubin, 1985; Rubin, 2001). Where $\bar{x}_t$ and $\bar{x}_c$ are the means of the target and comparison groups on the matching variable, respectively, and s^2

is the corresponding variance, Cohen's d should be calculated as $(\bar{x}_t - \bar{x}_c) / \sqrt{s_t^2 + s_c^2} / 2$, when population variances are assumed equal a priori and sample sizes are equal. Note that

Cohen's d is calculated as $(\bar{x}_t - \bar{x}_c) / \sqrt{[(n_t - 1)s_t^2 + (n_c - 1)s_c^2] / [n_t + n_c - 2]}$ for equal or unequal sample sizes, but the formula simplifies as above when sample sizes are equal.

When variances are not assumed equal and/or interpreting the mean difference with respect to the variance of the comparison group is preferred, Cohen's d can be calculated as

$(\bar{x}_t - \bar{x}_c) / \sqrt{s_c^2}$. For our hypothetical example of the $n = 20$ groups (see Appendix), Cohen's d is: $(68.10 - 67.10) / \sqrt{(36+20)/2} = .19$. Effect sizes should be reported as best practice for tests of the dependent variable (American Psychological Association, 2010; Bakeman, 2006), but also reporting standardized mean differences alongside p -values on the matching variable provides context to the comparison of groups. The strategy of reporting effect sizes has been utilized in investigations of language and cognitive abilities in boys with ASD to aid the reader in interpreting the equivalence between the target and comparison samples (Brown, Aczel, Jimenez, Kaufman, & Grant, 2010; Kover, McDuffie, Hagerman, & Abbeduto, under revision).

The weaknesses of matching groups on just one aspect of their distributions (i.e., means) has been noted (Facon, Magis, & Belmont, 2011); however, using p -value threshold tests on variance, skewness, and kurtosis may exacerbate the issues associated with using p -value thresholds for means alone. We do not recommend p -value threshold matching for variances in addition to means. Rather, we favor Rubin's (2001) guideline, which avoids use of an inferential statistic: reporting the ratio of the variance of the target group to the variance of the comparison group. The variance ratio should be calculated as: s_t^2 / s_c^2 . For our hypothetical example of the $n = 20$ groups, the variance ratio is: $36.00/20.20 = 1.78$.

Thresholds for Effects Sizes and Variance Ratios

Of course, the issue of where to set the equivalence thresholds for effect sizes and variance ratios remains (Shadish & Steiner, 2010). Researchers will need to decide on meaningful thresholds based on seminal substantive studies because general guidelines are not universally applicable and should be used only when other references are not available (Cohen, 1988):

"The terms 'small,' 'medium,' and 'large' are relative, not only to each other, but to the area of behavioral science or even more particularly to the specific content and research method being employed in any given investigation...In the face of this relativity, there is a certain risk inherent in offering conventional operational definitions for these terms for use in power analysis in as diverse a field of inquiry as behavioral science. This risk is nevertheless accepted in the belief that more is to be gained than lost by supplying a common conventional frame of reference which is recommended for use only when no better bases for estimating the ES [effect size] index is available" (p. 25).

An adequately small effect size for matched groups might be defined as the smallest value at which a difference in groups would be clinically meaningful (Piaggio et al., 2006). Rubin (2001) proposed that the standardized mean difference be close to zero (less than half a standard deviation apart; $d < .5$) and that the ratio of variances be near 1 (0.5 and 2 serve as endpoints that indicate very poor variance matches). Others have been more specific in defining equivalence as a standardized mean difference near zero such that it is within .1 standard deviation and a variance ratio greater than .8 and less than 1.25 (Steiner, Cook, Shadish, & Clark, 2010). In research on ASD, a Cohen's d of less than .20 has been described as trivial, but this threshold has yet to be evaluated in terms of group-matching

adequacy (Cicchetti et al., 2011). Steiner and Cook (2012) point out that a given effect size on the matching variable must be interpreted together with the expected effect size of the variable of interest (e.g., a Cohen's d of 0.15 on the matching variable would not be sufficiently small if the effect of interest was expected to be 0.20).

We suggest that groups be considered adequately matched when they fall within the field's standards for *both* the absolute value of Cohen's d and the variance ratio. Table 1 lists a variety of effect sizes and variance ratios with illustrative corresponding means and variances on a matching variable for two groups. A Cohen's d of 0.00 reflects well-matched group means; a Cohen's d of 1.00 reflects poorly matched groups. A variance ratio of 1 indicates no difference in variances; a ratio of 2 reflects an unacceptable magnitude of difference in the spread of the distributions. For our hypothetical example, the effect size of .19 might be sufficiently small in some contexts for some researchers to conclude that the groups are matched, but taken together with the variance ratio of 1.78, it is unlikely that these two groups should be considered matched in most fields of study.

Although negotiating appropriate equivalence thresholds will be far from a trivial feat, these descriptive indices of group matching have several strengths. First, effect sizes are less directly affected by sample size than are p -values. Second, effect sizes and variance ratios can be used in combination with other metrics of equivalence, including visual inspection of plots and p -values from the t -test on the matching variable. Furthermore, because means and standard deviations are usually reported for matching variables in published studies, an interested reader can calculate effect sizes and variance ratios to aid in interpreting extant findings. In Table 2, we summarize the strengths and weaknesses of the indices of equivalence discussed, as well as the methods described in the next section.

Existing Methodologies Applied to IDD Research

Simple group-matching designs are ubiquitous in research on IDDs; however, other methodological options are available. We briefly describe three classes of methodologies with strengths and weaknesses that may be unfamiliar to the reader: equivalence tests, propensity score matching, and regression-based techniques.

Equivalence Tests

Often used in medical studies to demonstrate that the difference between two treatments does not exceed some clinically meaningful equivalence threshold, equivalence tests can also be applied to behavioral research (Rogers, Howard, & Vessey, 1993; Serlin & Lapsley, 1985; Stegner et al., 1996). Schuirmann (1987) suggested a "two one-sided tests" procedure wherein one may conclude that μ lies within the equivalence bounds $(-\delta_B, \delta_B)$ by simultaneously rejecting both $H_0: \mu - \delta_B$ and $H_0: \mu - \delta_B$. For Westlake's (1979) confidence interval method, equivalence is established if the confidence interval (constructed in the usual manner, but with coverage of 0.90) for $\hat{\mu}$, the observed mean difference, falls entirely within the equivalence bounds $(-\delta_B, \delta_B)$. Finally, the range equivalence confidence intervals proposed by Serlin and Lapsley (1985; 1993) stem from a good-enough principle and provide an additional alternative to "strawman" point null hypothesis testing. It is important to note that limited sample sizes may prevent equivalence methods from having

the necessary power to detect a truly ‘trivial’ effect, or else triviality may need to be set at a higher magnitude than would be desired. For example, Brown et al. (2010) concluded based on equivalence testing that implicit learning is unimpaired in individuals with ASD relative to typical development, though their choice of threshold value may have been unusually large.

Propensity Scores

The state-of-the-art for matching nonequivalent control groups in quasi-experimental design is propensity score analysis. With the goal of removing selection bias by modeling the probability of being in the target group, propensity scores are aggregated variables that predict group membership using logistic regression (Fraser & Guo, 2009; Shadish et al., 2002). Propensity score analysis involves creating a single score from many variables that could be related to group membership and then matching the groups on those propensity scores (Fraser & Guo, 2009). The nonequivalent control groups are often matched utilizing algorithms that, for example, select comparison participants who have scores within a defined absolute difference from a given target participant (i.e., caliper matching) or minimize the total difference between pairs of target and comparison participants (i.e., optimal matching; Rosenbaum, 1989).

Propensity scores are best suited to the analysis of large datasets in which it is reasonable to assume that all variables relevant to group membership have been measured and those with complete overlap between the groups on the range of propensity scores (Shadish et al., 2002). In addition, propensity score analysis may be no better than regression techniques unless the primary concern is the large number of matching variables included in the analysis (Shadish & Steiner, 2010). Despite the fact that these conditions are rarely met in IDD research, there are cases in which propensity score matching has been applied. For example, Blackford (2009) used propensity score matching for data from State of Tennessee administrative databases to test whether infants with Down syndrome have lower birth weight than those without. Unfortunately, such large databases are yet unavailable to answer many research questions relevant to neurodevelopmental disorders.

Importantly, matching groups on propensity scores escapes neither the problem of having a satisfactory way to determine when groups are adequately matched nor other problems associated with matching groups on a single variable. Even when using large samples and sophisticated matching algorithms, matching can be problematic when the populations of interest do not completely overlap in range. As such, group-matching procedures can lead researchers to analyze data from samples of participants that are not representative of the populations from which they are drawn or to which the researcher wishes to generalize (Shadish et al., 2002). Furthermore, when participants are chosen from the ends of their distributions due to matching criteria and when matching variables are measured with error, regression to the mean is of concern because participants selected for their extreme, apparently nonrepresentative scores are likely to have less extreme scores on the dependent variable and/or over time (Breaugh & Arnold, 2007; Marsh, 1998; Shadish et al., 2002). Thus, propensity scores are not a panacea for researchers interested in a single matching

construct or those with limited resources to collect large samples with all measurable variables relevant to group membership.

Regression-based Methods

ANCOVA—Analysis of covariance (ANCOVA) is sometimes used as an alternative to group matching. ANCOVA is ideal for reducing sampling bias due to variability between groups in experimental designs when unbalance occurs due to chance. Assumptions of ANCOVA include: group membership independent of the covariate, linearly related predictor and outcome, and identical slopes for the groups between the covariate and the dependent variable. When used with preexisting groups, a researcher can expect difficult interpretation, at best, and spurious findings, at worst, because ANCOVA attempts to adjust or control for part of what the group effect is thought to be (Brock, Jarrold, Farran, Laws, & Riby, 2007; Miller & Chapman, 2001). For neurodevelopmental disorders, the “selection bias” being removed is often integrally related to the causal effect of interest (e.g., background genes, maternal interaction styles, family stress, world experiences; see Newcombe, 2003 for an example related to children's socioemotional adjustment). In these cases, statistical adjustments between groups diminish true population differences that are attributes of the disorder, yielding uninterpretable results (Dennis et al., 2009; Miller & Chapman, 2001; Tupper & Rosenblood, 1984). A strong argument has been made in particular against the use of IQ as a covariate in studies of neurodevelopmental disorders because it is inseparable from the disorder itself (Dennis et al., 2009).

More generally, the process of choosing a matching variable or covariate should be deliberate. Preliminary tests of significance—including tests on the matching variable to decide whether it should be used as a covariate—are not recommended (Atwood, Swoboda, & Serlin, 2010; Zimmerman, 2004). Above all, the choice of covariate or matching variable is likely to have a greater impact on the conclusions drawn than the choice of analytic method and, thus, should be carefully theoretically justified (Breaugh & Arnold, 2007; Steiner et al., 2010).

Developmental trajectories and residuals—Distinct from ANCOVA, Thomas and colleagues (2009) have put forth a regression-based approach, termed cross-sectional developmental trajectories analysis, that allows testing within-group slope differences with respect to theoretically motivated predictors. From this perspective, trajectories are estimated for the dependent variable of interest relative to age and other predictors, such as nonverbal cognitive ability, and these trajectories are compared between a target group and a large comparison group. Conclusions can be drawn about group differences in intercepts (i.e., level of ability) and slopes (i.e., the relationship between a given predictor and the variable of interest). This approach has been applied to aspects of cognitive development in individuals with Williams syndrome (Karmiloff-Smith et al., 2004) and vocabulary ability in individuals with ASD (Kover et al., under revision). We refer the interested reader to the detailed substantive examples and thorough characterization of the approach provided by Thomas et al. (2009), which includes an online worksheet that walks through trajectory analyses step-by-step.

A special case of this type of analysis involves standardizing the performance of the target group based on the residual score (the difference between observed and predicted) from the trajectory of the comparison group (Jarrold & Brock, 2004). The *z*-scores (or alternatively, scores divided by the standard error of the regression estimate) of these residuals can be used to assess relative deficits on multiple tasks of interest that have been standardized using the same predictor (Jarrold, Baddeley, & Phillips, 2007; Jarrold & Brock, 2004). For example, Jarrold and colleagues (2007) examined the performance of individuals with Down syndrome and Williams syndrome on memory tasks with respect to multiple control variables (e.g., age, vocabulary ability), standardized against the performance of 110 typically developing children. By standardizing performance relative to these constructs, Jarrold et al. (2007) identified distinct relationships among abilities relative to the comparison group and differentiated the nature of the deficits in long-term memory in individuals with Down syndrome from those with Williams syndrome.

The developmental trajectories method carries fewer assumptions than ANCOVA because the regression with the matching variable is done for the comparison group alone, avoiding the potential to violate the assumption of independence between the covariate and group (Brock et al., 2007). While allowing simultaneous analysis and “comparison” of disparate participant groups, this procedure is not without limitations. First, a very large comparison group is required. Second, transformations of matching and dependent variables limit the extent to which results can be transparently interpreted. Finally, like other methods, this technique still requires linearity and complete overlap between the groups on the matching variable. As data sharing and access to national datasets (e.g., the National Database for Autism Research; NDAR) become more common, analytic techniques like the developmental trajectories approach will only become more valuable because of the availability of larger samples.

Summary of Recommendations for Researchers

Having brought attention to some of the methodological challenges in research on IDD, we close with comments on the relationship between research questions and design, and on the responsible use of effect sizes and variance ratios as descriptive equivalence thresholds.

Choose Productive Research Questions

Thoughtful research questions that yield interpretable results should drive study design. We have focused on the simplest type of group-matching design (i.e., two groups and one matching variable); however, many research questions call for other applications of nonequivalent comparison designs. For example, pair-wise matching on one or more control variables might ensure more closely matched groups, but it might also call into question the generalizability of the findings (Mervis & Robinson, 1999). In some cases, studies might be strengthened by including multiple comparison groups (Burack et al., 2002; Eigsti et al., 2011) or by matching that is conducted on control tasks that very closely align with the skill of interest (Jarrold & Brock, 2004). Another alternative is creating individual profiles of ability (e.g., case-study analysis), rather than group-level profiles that might fail to represent any individuals from the population from which the sample was drawn (Mervis & Robinson,

1999; Towgood, Meuwese, Gilbert, Turner, & Burgess, 2009). Regardless of the research question, reporting results based on multiple matching and analysis techniques will leave the reader informed and free to draw conclusions based on maximal information (Breugh & Arnold, 2007; Brock et al., 2007; Kover et al., under revision; Mervis & John, 2008).

Shifting focus towards understanding individual variability avoids some difficulties associated with group matching, while also leading researchers closer to understanding the sources of difficulty that result in phenotypic strengths and weaknesses. Comparing unrepresentative samples provides little advantage over studying the entire range of variability within a given phenotype and identifying foundational cognitive skills that account for individual variation (Beeghly, 2006; Eigsti et al., 2011). Adopting an individual differences approach can highlight phenotypic variability and emphasizes the prerequisite skills necessary for development, ultimately supporting research that emphasizes learning mechanisms rather than outcomes. Of course, some research questions will nonetheless necessitate group comparisons.

Use Effect Sizes and Variance Ratios for Equivalence Thresholds

Group-matching studies that appropriately compare groups presumed to be equivalent on a single matching variable have the potential to provide the groundwork for stronger, well-controlled studies of greater scope. Researchers will benefit from including as many sources of information as possible for establishing group equivalence: plots of the distributions, effect sizes, variance ratios, etc. Given the complexities faced by IDD researchers, our recommendation is that groups be considered adequately matched when both the effect size (e.g., Cohen's d) and variance ratio fall within acceptable ranges for a particular area of research. We have provided a table of effect sizes and variance ratios that demonstrates how this technique can be applied to decision making regarding group matching adequacy; however, this table is meant to be thought-provoking, not prescriptive. In published reports, best practice would be to report effect sizes and variance ratios in all cases—for the matching variable and the dependent variable of interest—to allow the reader to interpret where meaningful differences exist.

Conclusions and Future Directions

We have discussed the limitations of p -value thresholds and ways in which using descriptive diagnostics (effect sizes and variance ratios) as equivalence thresholds will benefit research on neurodevelopmental disorders. Drawing the interest of methodological specialists to the study of IDDs will also be key to advancing the field. Open dialogue concerning current practices, paired with the development of improved methods for defining and testing meaningful differences, will significantly improve the design and implementation of research on IDDs.

Acknowledgments

This work was supported in part by NIH P30 HD03352 to the Waisman Center and NIH T32 DC05359. We thank Peter Steiner for his comments on an earlier draft. Following Strauss (2001), we chose to maintain a methodological focus and avoided citing substantive studies as examples, with the exception of those that have utilized methodologies likely to be lesser known to the reader.

References

- Abbeduto L. Editorial. *American Journal on Intellectual and Developmental Disabilities*. 2010; 115(1): 1–2.
- American Psychological Association. Publication manual of the American Psychological Association. Sixth. Washington, D.C: Author; 2010.
- Atwood, AK., Swoboda, CM., Serlin, RC. The impact of selection procedures for nonnormal covariates on the Type I error rate and power of ANCOVA. Paper presented at the the annual meeting of the American Educational Research Association; Denver, CO. 2010. Paper retrieved from
- Bakeman R. VII. THE PRACTICAL IMPORTANCE OF FINDINGS. *Monographs of the Society for Research in Child Development*. 2006; 71(3):127–145.
- Beeghly M. Translational research on early language development: Current challenges and future directions. *Development and Psychopathology*. 2006; 18(03):737–757. [PubMed: 17152398]
- Blackford JU. Propensity Scores: Method for Matching on Multiple Variables in Down Syndrome Research. *Intellectual and Developmental Disabilities*. 2009; 47(5):348–357. [PubMed: 19842739]
- Breaugh JA, Arnold J. Controlling nuisance variables by using a matched-groups design. *Organizational Research Methods*. 2007; 10(3):523–541.
- Brock J, Jarrold C, Farran EK, Laws G, Riby DM. Do children with Williams syndrome really have good vocabulary knowledge? Methods for comparing cognitive and linguistic abilities in developmental disorders. *Clinical Linguistics & Phonetics*. 2007; 21(9):673–688. [PubMed: 17701755]
- Brown J, Aczel B, Jimenez L, Kaufman SB, Grant KP. Intact implicit learning in autism spectrum conditions. *Quarterly Journal of Experimental Psychology (Hove)*. 2010; 63(9):1789–1812.
- Burack J. Editorial Preface. *Journal of Autism and Developmental Disorders*. 2004; 34(1):3–5.
- Burack JA, Iarocci G, Bowler D, Motttron L. Benefits and pitfalls in the merging of disciplines: The example of developmental psychopathology and the study of persons with autism. *Development and Psychopathology*. 2002; 14(2):225–237. [PubMed: 12030689]
- Cicchetti DV, Koenig K, Klin A, Volkmar FR, Paul R, Sparrow S. From Bayes through marginal utility to effect sizes: a guide to understanding the clinical and statistical significance of the results of autism research findings. *Journal of Autism and Developmental Disorders*. 2011; 41(2):168–174.
- Cohen, J. Statistical power analysis for the behavioral sciences. 2. Hillsdale, N.J: L. Erlbaum Associates; 1988.
- Dennis M, Francis DJ, Cirino PT, Schachar R, Barnes MA, Fletcher JM. Why IQ is not a covariate in cognitive studies of neurodevelopmental disorders. *Journal of the International Neuropsychological Society*. 2009; 15(03):331–343. [PubMed: 19402919]
- Edgell SE. Commentary on "Accepting the null hypothesis". *Memory & Cognition*. 1995; 23(4):525–526. [PubMed: 7666764]
- Eigsti IM, de Marchena AB, Schuh JM, Kelley E. Language acquisition in autism spectrum disorders: A developmental review. *Research in Autism Spectrum Disorders*. 2011; 5(2):681–691.
- Facon B, Magis D, Belmont JM. Beyond matching on the mean in developmental disabilities research. *Research in Developmental Disabilities*. 2011; 32(6):2134–2147. [PubMed: 21856117]
- Fraser, MW., Guo, S. Propensity Score Analysis: Statistical Methods and Applications. SAGE Publications; 2009.
- Frick RW. Accepting the Null Hypothesis. *Memory & Cognition*. 1995; 23(1):132–138. [PubMed: 7885262]
- Holland PW. Statistics and Causal Inference. *Journal of the American Statistical Association*. 1986; 81(396):945–960.
- Imai K, King G, Stuart EA. Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society Series a-Statistics in Society*. 2008; 171:481–502.

- Jarrold C, Baddeley AD, Phillips C. Long-term memory for verbal and visual information in Down syndrome and Williams syndrome: performance on the Doors and People test. *Cortex*. 2007; 43(2):233–247. [PubMed: 17405669]
- Jarrold C, Brock J. To match or not to match? Methodological issues in autism-related research. *Journal of Autism and Developmental Disorders*. 2004; 34(1):81–86. [PubMed: 15098961]
- Karmiloff-Smith A, Thomas M, Annaz D, Humphreys K, Ewing S, Brace N, Campbell R. Exploring the Williams syndrome face-processing debate: the importance of building developmental trajectories. *Journal of Child Psychology and Psychiatry*. 2004; 45(7):1258–1274. [PubMed: 15335346]
- Kover S, McDuffie A, Hagerman R, Abbeduto L. Receptive vocabulary in boys with autism spectrum disorder: Cross-sectional developmental trajectories. (under revision).
- Marsh HW. Simulation study of nonequivalent group-matching and regression-discontinuity designs: Evaluations of gifted and talented programs. *Journal of Experimental Education*. 1998; 66(2):163–192.
- Mervis CB, John AE. Vocabulary abilities of children with Williams syndrome: strengths, weaknesses, and relation to visuospatial construction ability. *Journal of Speech, Language, and Hearing Research*. 2008; 51(4):967–982.
- Mervis CB, Klein-Tasman BP. Methodological Issues in Group-Matching Designs: α Levels for Control Variable Comparisons and Measurement Characteristics of Control and Target Variables. *Journal of Autism and Developmental Disorders*. 2004; 34(1):7–17. [PubMed: 15098952]
- Mervis CB, Robinson BF. Methodological issues in cross-syndrome comparisons: Matching procedures, sensitivity (Se) and specificity (Sp). *Monographs of the Society for Research in Child Development*. 1999; 64(1):115–130. [PubMed: 10412223]
- Miller GA, Chapman JP. Misunderstanding analysis of covariance. *Journal of Abnormal Psychology*. 2001; 110(1):40–48. [PubMed: 11261398]
- Newcombe NS. Some controls control too much. *Child Development*. 2003; 74(4):1050–1052. [PubMed: 12938700]
- Piaggio G, Elbourne DR, Altman DG, Pocock SJ, Evans SW, f CG. Reporting of noninferiority and equivalence randomized trials: An extension of the consort statement. *Journal of the American Medical Association*. 2006; 295(10):1152–1160. [PubMed: 16522836]
- R Development Core Team. R: A Language and Environment for Statistical Computing (Version 2.10.1). Vienna, Austria: R Foundation for Statistical Computing; 2010. Retrieved from <http://www.R-project.org>
- Rogers JL, Howard KI, Vessey JT. Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin*. 1993; 113(3):553–565. [PubMed: 8316613]
- Rosenbaum PR. Optimal Matching for Observational Studies. *Journal of the American Statistical Association*. 1989; 84(408):1024–1032.
- Rosenbaum PR, Rubin DB. Constructing a Control-Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score. *American Statistician*. 1985; 39(1):33–38.
- Rubin DB. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*. 1974; 66(5):688–701.
- Rubin DB. Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation. *Health Services and Outcomes Research Methodology*. 2001; 2(3):169–188.
- Schervish MJ. P values: What they are and what they are not. *American Statistician*. 1996; 50(3):203–206.
- Schuirman DJ. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*. 1987; 15(6):657–680. [PubMed: 3450848]
- Serlin RC, Lapsley DK. Rationality in psychological research: The good-enough principle. *American Psychologist*. 1985; 40(1):73–83.
- Serlin, RC., Lapsley, DK. Rational appraisal of psychological research and the good-enough principle. In: Keren, G., Lewis, C., editors. *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues*. Hillsdale, NJ: Erlbaum; 1993. p. 199–228.

- Shadish, WR., Cook, TD., Campbell, DT. Experimental and quasi-experimental designs for generalized causal inference. Houghton Mifflin: 2002.
- Shadish WR, Steiner PM. A Primer on Propensity Score Analysis. *Newborn and Infant Nursing Reviews*. 2010; 10(1):19–26.
- Stegner BL, Bostrom AG, Greenfield TK. Equivalence testing for use in psychosocial and services research: An introduction with examples. *Evaluation and Program Planning*. 1996; 19(3):193–198.
- Steiner, PM., Cook, DL. Matching and propensity scores. In: Little, TD., editor. *Oxford Handbook of Quantitative Methods*. New York: Oxford University Press; 2012.
- Steiner PM, Cook TD, Shadish WR, Clark MH. The Importance of Covariate Selection in Controlling for Selection Bias in Observational Studies. *Psychological Methods*. 2010; 15(3):250–267. [PubMed: 20822251]
- Strauss ME. Demonstrating specific cognitive deficits: A psychometric perspective. *Journal of Abnormal Psychology*. 2001; 110(1):6–14. [PubMed: 11261401]
- Stuart EA. Matching Methods for Causal Inference: A Review and a Look Forward. *Statistical Science*. 2010; 25(1):1–21. [PubMed: 20871802]
- Thomas MS, Annaz D, Ansari D, Scerif G, Jarrold C, Karmiloff-Smith A. Using developmental trajectories to understand developmental disorders. *Journal of Speech, Language, and Hearing Research*. 2009; 52(2):336–358.
- Towgood KJ, Meuwese JDI, Gilbert SJ, Turner MS, Burgess PW. Advantages of the multiple case series approach to the study of cognitive deficits in autism spectrum disorder. *Neuropsychologia*. 2009; 47(13):2981–2988. [PubMed: 19580821]
- Tupper DE, Rosenbloom LK. Methodological considerations in the use of attribute variables in neuropsychological research. *Journal of Clinical Neuropsychology*. 1984; 6(4):441–453. [PubMed: 6501582]
- Westlake WJ. Statistical Aspects of Comparative Bioavailability Trials. *Biometrics*. 1979; 35(1):273–280. [PubMed: 583027]
- Zimmerman DW. A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology*. 2004; 57(1):173–181. [PubMed: 15171807]

Appendix

Hypothetical Scores on a Matching Variable from Two Groups

Participant Count	Target Group Scores	Comparison Group Scores
1	36	92
2	36	88
3	36	87
4	42	85
5	46	82
6	47	76
7	48	75
8	49	75
9	50	75
10	51	74
11	52	72
12	60	72
13	61	72
14	62	72
15	64	71

Participant Count	Target Group Scores	Comparison Group Scores
16	67	71
17	67	70
18	68	70
19	68	70
20	69	69
21	69	67
22	70	67
23	70	66
24	71	65
25	72	64
26	72	63
27	73	62
28	74	61
29	75	60
30	78	58
<i>n</i> = 30 Mean (SD)	60.10 (12.94)	71.70 (8.42)
<i>n</i> = 20 Mean (SD)	68.10 (6.00)	67.10 (4.49)

Note. The shaded cells show the subset of 20 participants in each group who remained in the analysis during the process of obtaining a higher *p*-value. They were chosen simply to demonstrate the calculation of effect size and variance ratio, not to demonstrate adequate equivalence.

Figure 1.

Proportion of samples considered to be adequately matched according to p -value thresholds by sample size and population effect size (Cohen's d).

Table 1
Example Standardized Mean Differences (Cohen's *d*) and Variance Ratios as Thresholds

<i>d</i>	Means			Variances		
	Target Group	Comparison Group	Adequacy	Ratio	Target Group	Comparison Group Adequacy
0.00	100	100	Matched	1.00	225	Matched
0.05	99.25	100	Matched	1.10	247.5	Matched
0.13	98	100	Matched	1.20	270	Matched
0.20	97	100	Not Matched	1.33	299.25	Not Matched
0.33	95	100	Not Matched	1.5	337.5	Not Matched
0.50	92.5	100	Not Matched	1.75	394	Not Matched
1.00	85	100	Not Matched	2.00	450	Not Matched

Note. Matched-group adequacy should be evaluated with respect to both means (i.e., absolute value of the effect size) and variances. We emphasize that decisions regarding the adequacy of group matches must be reached through consensus within individual fields; we merely provide starting points based on Rubin (2001) and Steiner and Cook (2012). Sample statistics reflect a matching variable with $M = 100$ and $SD = 15$.

Table 2

Brief Summary of Strengths and Weaknesses of Methodologies for IDD Research

Method	Strength	Weakness
Single-variable matching		
<i>p</i> -value threshold	Widely used in IDD research	Violates underlying logic of hypothesis testing; matching conclusions depend on sample size
Effect size threshold	Avoids direct influence of sample size on matching decision	Not sufficient without other evidence (e.g., variance ratio, plot)
Variance ratio threshold	Descriptive criterion for width of distributions	Not sufficient without other evidence (e.g., effect size, plot)
Equivalence testing	Allows appropriate logical use of <i>p</i> -values to indicate statistical equivalence	Large sample sizes usually required
Propensity score matching	Sophisticated analysis for causal inference	Many participants and many measures usually required; establishing balance is still subjective
ANCOVA	Simple implementation	Assumptions may be violated for IDD participant samples; complex interpretations of findings
Developmental trajectories (Thomas et al., 2009; Jarrold & Brock, 2004)	Accessible, theoretically-motivated analysis	Large comparison group required