Psychological Methods
2009, Vol. 14, No. 2, 81–100

1082-989X/09/$12.00 DOI: 10.1037/a0015914

Integrative Data Analysis: The Simultaneous Analysis of Multiple Data Sets

Patrick J. Curran and Andrea M. Hussong
University of North Carolina at Chapel Hill

There are both quantitative and methodological techniques that foster the development and maintenance of a cumulative knowledge base within the psychological sciences. Most noteworthy of these techniques is meta-analysis, which allows for the synthesis of summary statistics drawn from multiple studies when the original data are not available. However, when the original data can be obtained from multiple studies, many advantages stem from the statistical analysis of the pooled data. The authors define integrative data analysis (IDA) as the analysis of multiple data sets that have been pooled into one. Although variants of IDA have been incorporated into other scientific disciplines, the use of these techniques is much less evident in psychology. In this article the authors present an overview of IDA as it may be applied within the psychological sciences, discuss the relative advantages and disadvantages of IDA, describe analytic strategies for analyzing pooled individual data, and offer recommendations for the use of IDA in practice.

Keywords: integrative data analysis, pooled data analysis, longitudinal modeling, cumulative science

The cornerstone of any field of scientific inquiry is the pursuit of a body of cumulative knowledge, yet the psychological sciences have often fallen short of this goal (e.g., Gans, 1992; J. E. Hunter & Schmidt, 1996; Meehl, 1978; Schmidt, 1996). This is not for want of trying. Both quantitative and methodological techniques have been developed to help build a cumulative knowledge base. Most noteworthy of these techniques is meta-analysis, which allows for the synthesis of summary statistics drawn from multiple studies when the original data are not available (e.g., Cooper, Hedges, & Valentine, 2009; Glass, 1976; Rothstein, Sutton, & Borenstein, 2005; Smith & Glass, 1977). One of the original motivations for meta-analysis was the idea that these techniques would further support the creation of a cumulative knowledge within the social sciences, particularly in psychology (e.g., J. E. Hunter & Schmidt, 1996; Schmidt, 1984). There is no doubt that meta-analysis has substantially advanced our science toward this goal.

Because the focus in meta-analysis is on the synthesis of summary statistics drawn from multiple studies, this approach is ideal when the original individual data used in prior analyses are inaccessible or no longer existent. However, as we discuss in greater detail below, there are many advantages to fitting models directly to the original raw data instead of synthesizing the relevant summary statistics when the original individual data are available for analysis (e.g., Berlin, Santanna, Schmid, Szczech, & Feldman, 2002; Lambert, Sutton, Abrams, & Jones, 2002). Recent developments within the scientific community, such as greater expectations for data sharing and better options for electronic data storage and retrieval, have increased the potential for accessing original individual data for secondary analysis (i.e., the analysis of existing data). This potential in turn creates new opportunities for the development of alternative methods for integrating findings across studies by use of original individual data that could help overcome some of the unavoidable limitations of meta-analysis. (See Cooper & Patall, 2009, for a thoughtful comparison of the advantages and disadvantages of meta-analysis relative to the pooled analysis of raw data.)

Techniques for fitting models to pooled data go by various names, none of which have been broadly adopted within the social sciences. Simply to offer a starting point, we will

Patrick J. Curran and Andrea M. Hussong, Department of Psychology, University of North Carolina.

This work was partially supported by National Institute on Drug Abuse Grants DA013148, awarded to the Patrick J. Curran, and DA15398, awarded to Andrea M. Hussong. We are deeply indebted to Dan Bauer, Dave Flora, and all of the past and present members of the Cross Study group whose creativity, commitment, and exceedingly high frustration tolerance have made this project possible.

Correspondence concerning this article should be addressed to Patrick J. Curran, Department of Psychology, University of North Carolina, Chapel Hill, NC 27599-3270. E-mail: curran@unc.edu

refer to the fitting of models to data that have been pooled across multiple studies as integrative data analysis, or IDA.[1] We chose the term *integrative* over options such as pooled, simultaneous, unified, or concomitant to highlight our goal of creating "a whole by bringing all parts together," which is a common definition of *integrate* (e.g., American Heritage Dictionary of the English Language, 2006). IDA has been used in other areas of scientific inquiry for more than a decade. For example, IDA has been used to examine the efficacy of medications versus cognitive behavior therapy for severe depression (DeRubeis, Gelfand, Tang, & Simons, 1999); to evaluate clinical trial outcomes for treatment of Alzheimer's disease (Higgins, Whitehead, Turner, Omar, & Thompson, 2001); to examine the relation between fat intake and the risk of breast cancer (D. J. Hunter et al., 1996); to study the pharmacogenetics of tardive dyskinesia (Lerer et al., 2002); and to examine the relation among height, weight, and breast cancer risk (van den Brandt et al., 2000).

Despite the broader use of IDA techniques in other disciplines, such applications are relatively novel within the behavioral sciences in general and within psychology in particular (for notable exceptions, see Lorenz et al., 1997; McArdle, Hamagami, Meredith, & Bradway, 2000; McArdle, Prescott, Hamagami, & Horn, 1998). One reason behind the slow adoption of these techniques may be the significant challenges that psychologists face in pooling across studies that are highly heterogeneous in their methodology, even when these studies examine the same topic. Differences between studies in sampling techniques and frame, historical timing, design characteristics, and measurement create seeming barriers to study comparison and integration. However, if we incorporate information about such between-study heterogeneity into our techniques for study integration, our conclusions may be more generalizable and our progress as a science may be more cumulative. Thus, use of IDA capitalizes upon such between-study heterogeneity not only to promote better understand findings across existing studies (i.e., study integration) but also to probe meaningful sources of between-study variability that may contribute to, and thus inform theories about, key psychological phenomena (i.e., study comparison).

The topics that underlie IDA are both broad and complex, and a comprehensive treatment is beyond the scope of any single article. Thus, our intent here is rather modest. We offer a general discussion of the core issues that typically arise in applications of IDA for study integration in the psychological sciences. These topics and our guiding perspective on IDA are largely culled from our experience in using these techniques on a project that we call Cross Study. Cross Study involves the integrated analysis of three independent longitudinal studies of children of alcoholic parents and matched controls. These data sets are unique in their excellent retention, breadth of measurement, and sampling of nontreatment samples. Nonetheless, the three studies differ in many respects (e.g., geographic location, developmental coverage, measurement, assessment modality). Because applications of IDA are necessarily idiosyncratic to the theoretical questions and sample characteristics at hand, we wholly acknowledge that our experiences on Cross Study have shaped our views of IDA and that this fact in turn is reflected throughout our work here. However, this same sensitivity to the specific theoretical and methodological context makes IDA both a broad topic that eludes simple description and a flexible, informative set of techniques that is critically needed in our field.

In this article, we build on our work with Cross Study in an effort to further establish IDA as a potential tool for pursuing and fostering a cumulative knowledge base in our field. We begin with a discussion of what IDA is and what advantages IDA offers when appropriate data are available for analysis. Next we detail potential influences on between-sample heterogeneity that may serve either as nuisance factors when study integration is the goal or, in many instances, as sources of variance that offer novel insights about why a phenomenon may show study-to-study differences. We then explore general analytic strategies that address between-study heterogeneity. We conclude with future directions for research and recommendations for the use of these techniques in practice.

What Exactly Is IDA?

Because methods for pooling existing data can vary across discipline, we begin by offering a specific definition of IDA within the psychological sciences. IDA is the statistical analysis of a single data set that consists of two or more separate samples that have been pooled into one. What constitutes "separate" sometimes can be unambiguously determined and other times cannot. In some cases, minor design differences between samples may be present. For example, separate samples may be collected within a multisite or rolling recruitment single-site design, in which key design characteristics are held constant (e.g., recruitment, procedures, measurement) yet each study is conducted in a different setting (e.g., different hospitals or regions of country) or across different time periods (e.g., as recruitment rolls across different school years or birth cohorts). These separate samples are then pooled for analysis, with some control for site or cohort differences (e.g., Kaplow, Curran, Dodge, and the Conduct Problems Prevention Research Group, 2002; Stark et al., 2005). In other cases, many design differences between samples may be present. For example, multiple separate samples may each be collected as part of

[1] We realize that psychology needs another acronym like it needs a hole in the head, but we also believe that the set of techniques we explore here is in need of some shared terminology (thus, IDA).

a different independent study that was conducted at a different historical time with different sampling mechanisms, experimental procedures, and psychometric instruments. Thus, what constitutes a "separate" sample ultimately resides on a continuum ranging from single-study multisite designs to the aggregation of multiple independent data sets.

Although we are cognizant of the continuum of designs to which IDA may be applied, our focus here is explicitly on the latter situation, namely, one in which multiple samples are drawn from independent existing studies and pooled into a single data set for subsequent analyses. This was precisely our experience in Cross Study, in which our focus was on data pooled from samples that were drawn from three independent studies whose participants differed from one another in ways that were both theoretically meaningful (e.g., status of family psychopathology) and methodologically meaningful (e.g., developmental level, measurement, recruitment strategies). We believe that the broadest potential for future applications of IDA in psychological research relates to the pooling of data that are drawn from two or more existing studies. For this reason, we focus the remainder of our discussion on this topic. To highlight the potential applicability of these techniques in psychological research settings, we describe our work on Cross Study and present exemplar findings that we believe could be obtained only by using IDA.

A Motivating Example: Cross Study

Cross Study is an ongoing project funded by the National Instititutes of Health. Data are pooled from three existing longitudinal studies of adolescent development, and there is a particular focus on identifying developmental pathways that lead to substance use and disorder. All three studies oversampled offspring who had at least one biological parent with alcoholism (i.e., children of alcoholics) and included matched controls of offspring who had neither biological parent diagnosed with alcoholism. The first study, the Michigan Longitudinal Study (MLS; Zucker et al., 2000), has amassed a broad data archive beginning with a sample of 2- to 5-year-olds who were assessed over four waves (at the time) into early adulthood. The second study is the Adolescent and Family Development Project (AFDP; Chassin, Rogosch, & Barrera, 1991); families were first interviewed when adolescents were age 11–15, with ongoing assessments continuing well into adulthood over five waves. The third study is the Alcohol, Health and Behavior Project (AHBP; Sher, Walitzer, Wood, & Brent, 1991), which began intensive assessments with college freshmen and has continued to survey participants over six waves into their 30s. Together, these three studies span the first four decades of life, in which early risk factors for later substance outcomes first emerge (childhood), substance use initiation typically occurs (adolescence), peak rates of substance use disorders are evident (young adulthood), and deceleration in substance involvement is first apparent (adulthood). Table 1 presents a summary of the pooled sample as a function of study membership and chronological age. Each cell in the table identifies the number of individuals assessed in a given wave of a given study at a given age. The column totals identify the total number of individuals assessed at a given age and pooled across study and wave.

Cross Study presented many methodological challenges within the context of studying early symptom trajectories associated with various forms of parent alcoholism. Notably, the design of the three contributing studies varied substantially in terms of issues such as participant recruitment, assessment strategies, and instrumentation (see Table 2 for a summary of design characteristics). One of our goals in Cross Study was to use IDA to control for such between-study differences as we examined our substantive questions of interest. The pursuit of this goal via IDA permitted us to study a longer developmental period than in any one study, larger subsamples of families with specific forms of alcoholism, and trajectories of symptomatology in analyses with greater statistical power.

Our approach to IDA in this work is best exemplified in Curran et al. (2008), in which trajectories of internalizing symptomatology between ages 10 and 33 were examined by combining data from all three studies. The pooled sample consisted of a total of 1,827 individual participants (512 drawn from the MLS, 830 from the AFDP, and 485 from the AHBP). Each individual provided between one and five repeated measures, resulting in a total of 7,377 person-by-time observations. We operationalized internalizing symptomatology using 27 dichotomous self-reported items; of these, 12 were drawn from the Anxiety and Depression subscales of the Brief Symptom Inventory (BSI; Derogatis & Spencer, 1982) and 15 were drawn from the Anxiety and Depression subscales of the Child Behavior Checklist (CBCL; Achenbach & Edelbrock, 1978). Six of these items were unique to the BSI, nine were unique to the CBCL, and six items were shared by the BSI and the CBCL. Given the six shared items, we used 21 unique items to define internalizing symptomatology (see Table 1 in Curran et al., 2008, for details). All of the 21 items were administered in the MLS, 10 of the 15 CBCL items and none of the BSI items were administered in the AFDP, and all 12 of the BSI items but none of the CBCL items were administered in the AHBP. We applied a series of item response theory (IRT; e.g., Thissen & Wainer, 2001) models to calculate scale scores for each individual at each time point. It should be noted that these scale scores were all anchored to a shared metric, regardless of the set of items to which the individual responded or the study to which the individual belonged. Finally, we fitted a series of multilevel piecewise growth models with which to examine the fixed and random effects

Table 1
Sample Sizes by Study and Chronological Age

		Participant age (years)																																		
Study	N	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38+
MLS1	399	18	143	121	88	27	2																															
MLS2	339				7	99	115	89	28	1																												
MLS3	401							14	128	139	102	18																										
MLS4	418										13	145	133	113	12	1		1																				
MLS5	500													19	144	149	147	31	10																			
MLS6	482																17	143	148	144	30																	
MLS7	482																			17	143	148	144	30														
MLSA1	168									19	147	2																										
MLSA2	158										17	132	9																									
MLSA3	204											8	188	8																								
MLSA4	247												17	214	16																							
MLSA5	219													11	193	15																						
MLSA6	202														11	179	12																					
MLSA7	203															14	182	7																				
AFDP1	454									32	78	85	107	102	46	4																						
AFDP2	449										29	77	84	106	101	48	4																					
AFDP3	447											29	75	86	103	100	50	4																				
AFDP4	749																10	147	116	123	128	71	45	44	38	26	1											
AFDP5	755																					13	34	55	50	84	100	92	143	91	62	20	11					
AFDP6	735																										10	30	50	50	80	105	90	140	90	60	20	10
AHBP1	485																8	396	70	4	4	2	1															
AHBP2	480																	8	394	68	4	4	1	1														
AHBP3	468																		8	362	88	4	4	1	1													
AHBP4	467																			6	340	110	6	3	2													
AHBP5	454																						3	283	154	8	4	0	2									
AHBP6	406																											112	225	59	7	1	2					
AHBP7	383																																31	221	113	12	5	1
Total		18	143	121	95	126	117	103	156	191	386	496	613	657	625	509	429	736	743	724	737	352	238	417	245	118	115	234	420	200	149	126	134	361	203	72	25	11

Note. MLS = Michigan Longitudinal Study; MLSA = Michigan Longitudinal Study annual assessments (see Zucker et al., 2000); AFDP = Adolescent and Family Development Project; AHBP = Alcohol, Health and Behavior Project. Each number appended to the study indicator indicates wave of assessment (e.g., AFDP3 is the third assessment wave in the AFDP project).

Table 2
Exemplar Study Design Differences in Cross Study

	MLS	AFDP	AHBP
Design			
Recruitment	Rolling community-based recruitment with COA families identified through father's court-arrest records and community canvassing.	A community-based sample with alcoholic parents identified through court records, HMO wellness questionnaires, and telephone surveys.	Recruited through a screening of 3,156 first-time freshmen at the University of Missouri who reported on paternal alcoholism using the father SMAST.
Assessment schedule	Mothers and fathers completed up to four assessments when the children were between ages 2 and 5, 6 and 8, 9 and 11, and 11 and 15 at 3-year intervals.	Mothers, fathers, and one child completed the first three annual waves of data on children age 10–17 and two subsequent follow-up waves at 5-year intervals; age-appropriate siblings were included as targets in the follow-up waves.	Children completed four annual assessments (Years 1–4) and two additional post-college follow-ups (at 3- and 4-year intervals, or Years 7 and 11).
Variable			
Parent alcoholism	Lifetime diagnosis was made by a trained clinician based on *DSM–IV* criteria with parent self-report at each wave using three instruments: Diagnostic Interview Schedule, the SMAST, and the Drinking and Drug History Questionnaire.	Lifetime diagnosis was made by interviews based on *DSM–III* criteria with parent self-report at the first wave using the computerized version of the Diagnostic Interview Schedule. In cases when a biological parent was not directly interviewed, the reporting parent was the informant on the FH–RDC.	Lifetime diagnosis was made by survey assessment based on *DSM–III* criteria with target (child) report at baseline using the parent SMAST and FH–RDC.

Note. MLS = Michigan Longitudinal Study; AFDP = Adolescent and Family Development Project; AHBP = Alcohol and Health Behavior Project; COA = child of alcoholic; *DSM–III* = Diagnostic and Statistical Manual of Mental Disorders (3rd ed.); *DSM–IV* = Diagnostic and Statistical Manual of Mental Disorders (4th ed.); FH–RDC = Family History Research Diagnostic Criteria; SMAST = Short Form of the Michigan Alcohol Screening Test.

characterizing the developmental trajectories of internalizing symptomatology between ages 10 and 33.

We have used these IDA methodologies to test many other theoretical questions with data pooled from two or three of the component studies. For example, in Hussong, Cai, et al. (2008) we used IDA to disaggregate the distal, proximal, and time-varying effects of parental alcoholism on children's internalizing symptomatology between ages 2 and 17. In Hussong et al. (2007) we used IDA to examine the relation between the number of alcoholic parents in the family, the specific subtype of parental alcoholism, and the gender of the child in the prediction of developmental trajectories of externalizing symptomatology between ages 2 and 17. And in Hussong, Flora, et al. (2008) we used IDA to examine the unique predictability of trajectories of child internalizing symptoms from parental alcoholism above and beyond the parental comorbid diagnoses of depression and antisocial personality disorder. Over the course of this work, we believe, we have been able to use IDA to empirically test hypotheses in ways that would not otherwise be possible. However, this has not been without a cost. We have addressed a seemingly endless parade of challenges, some foreseen and others not; ultimately, some were surmountable and others were not. We draw upon these experiences to organize a more general discussion about the potential advantages and disadvantages IDA may hold for other applications in the social sciences.

Potential Advantages of IDA

IDA is not universally appropriate for pooling data from any two independent studies. Nor is IDA intended to replace meta-analysis or any other method of research synthesis. Rather, it is an additional tool that may be used for the purposes of study integration and comparison, given conducive research contexts. Indeed, compared with available alternatives, IDA offers a host of significant advantages under the proper conditions. We believe that there are seven specific advantages of IDA that are particularly salient when the technique is used within many areas of psychological inquiry.

Replication

IDA provides a direct mechanism with which we can test whether a set of findings replicates across independent studies by explicitly quantifying effects that represent tests of our hypotheses both within and between studies.[2] Unlike other approaches to research synthesis that are based on summary statistics, IDA can directly model potential influences on between-study heterogeneity at the level of the observed data. Such modeling permits explicit analysis of study equivalence at multiple levels of design that can often incorporate differences in sampling, geographic region, history, assessment protocol, psychometric measurement, and even hypothesis testing. This advantage also makes IDA well suited to the task of testing novel hypotheses that may not have been considered in the original within-study analysis of the data. Thus, IDA may provide tests of replication of novel hypotheses within a single analysis of independent studies. Moreover, IDA permits an exploration of between-study differences that helps mitigate the need for creating new studies designed to resolve conflicting findings across studies posited to result from between-study design differences.

Increased Statistical Power

IDA has the potential to provide substantial increases in statistical power for testing research hypotheses through the combination of multiple individual data sets. It is well known that most research applications within psychology are often chronically underpowered, such that there is an unacceptably low probability that a given effect will be found if that effect truly exists in the population (Cohen, 1992; Maxwell, 2004). However, when multiple independent samples are combined, there is often a marked increase in power when the same hypotheses are tested on the basis of the aggregated versus independent sample.

Increased Sample Heterogeneity

A closely related advantage is that IDA often allows for a more heterogeneous pooled sample. For a variety of reasons, many studies in psychology use sampling methods that result in the underrepresentation of potentially important subgroups in the population of interest (e.g., groups based on gender, race, socioeconomic status, age). However, a pooled sample that is aggregated across multiple studies, each of which may have been conducted in a different geographic setting or with a different sampling mechanism, allows for simultaneous consideration of more distinct groups or individual characteristics. Moreover, given adequate sample representation within studies, group comparisons may be possible within IDA that are not possible, due to small sample sizes, within the individual studies. This in turn increases the external validity of the IDA findings fitted to the aggregated data.

Increased Frequencies of Low Base-Rate Behaviors

The same logic is evident as an advantage of IDA when pooling studies of low base-rate behaviors. For example, each contributing study may have 5% of the sample reporting heavy drug use. Although such behaviors will retain an overall low base rate in the pooled IDA analyses (e.g., assuming equal sample sizes, the aggregate sample would still reflect 5% heavy drug use), the overall absolute number of individuals engaging in the behavior will necessarily be greater in the pooled sample relative to the individual contributing studies (e.g., there may be 20 of 400 individuals reporting heroin use in a single study but 80 of 1,600 individuals reporting heroin use when four studies are pooled). As a result, the stability of model estimation is improved, the influence of extreme observations is reduced, and more complicated models can be fitted than would otherwise be possible within the individual studies.

Broader Psychometric Assessment of Constructs

IDA often results in a broader and more rigorous psychometric assessment of the key theoretical constructs under study. In any single-study design, theoretical constructs are typically assessed with a discrete set of items shared across all members of the sample (e.g., all subjects respond to the same 10-item scale assessing depression). A common challenge in many areas of psychological research is the need to reconcile the wide array of operationalizations of our constructs across studies. This is a seeming limitation for study-to-study comparison but is in turn a distinct advantage for increased construct validity in IDA. Researchers typically select the psychometric instruments for any given study on the basis of the specific characteristics of their sample (e.g., age, gender, ethnicity), although this in turn limits the generalizability of the subsequent results to the characteristics of the sample under study. Yet when multiple samples are combined, the psychometric assessment of a given construct can often be substantially broadened by incorporating the multiple methods of assessment that were used in each individual study. This approach often results in much stronger psychometric properties of the assessment of the theoretical constructs in the aggregated sample than in any given single sample.

[2] Because of the close relation between *study* and *sample* within our focus on IDA (in which a single study results in a single sample of data), we use these terms interchangeably throughout our article.

Extended Period of Developmental Study

Although researchers may use IDA to pool data drawn from cross-sectional (i.e., single time point) or longitudinal (i.e., multiple time point) studies, the pooling of longitudinal studies presents several distinct advantages. Most important, a single sample is obviously limited to the age range observed within that study. However, when multiple longitudinal studies are combined, a much broader developmental period can be considered, given overlapping age ranges across the set of contributing studies. For example, in Cross Study the MLS assessed individuals between the ages of 2 and 24, the AFDP assessed individuals between the ages of 10 and 34, and the AHBP assessed individuals between the ages of 17 and 40. Under many situations, IDA can allow for inferences across the entire range of ages from 2 through 40, even though each individual study followed participants for a fraction of this time (e.g., see Table 1). This advantage is then amplified by the combined psychometric assessments of the theoretical constructs that were deemed optimal within the particular age range under study.

Support of Data Sharing and Building a Cumulative Science

Finally, IDA is directly supportive of recent practical concerns about efficiency in psychological research, namely, increased calls for data sharing and decreased resources available to support new research endeavors. First, the issue of data sharing has been addressed both at the level of federal funding mechanisms (e.g., both the National Science Foundation and the National Institutes of Health have formal policies regarding data sharing) and at the level of appropriate ethical practices in research (e.g., Section 8.14 of the "Ethical Principles of Psychologists and Code of Conduct"; American Psychological Association, 2003). Not only are individual researchers increasingly called upon to share data but technological advances further support these efforts through the accessible electronic storage and distribution of even the largest data sets. Second, in recent years, non-defense-related federal funding for research and development has stagnated and sometimes decreased (American Association for the Advancement of Science, 2008).[3] As a result, the analysis of existing data is an extremely cost-efficient mechanism for conducting research (Kiecolt & Nathan, 1985). This efficiency is further realized by considering not just one but multiple existing samples of data. Thus, IDA meets several practical needs in terms of data sharing and maximizing limited resources.

Summary

Despite the many potential advantages to adopting an IDA framework for data integration, this remains an uncommon practice within the psychological sciences. One potential reason may be that conducting such analyses can be an extremely complex and challenging task. Key practical issues associated with data acquisition and data management are often eclipsed by a multitude of difficulties that arise from sometimes substantial study-to-study differences. Whereas the initial temptation might be to embark on IDA with the desire to minimize between-study heterogeneity (i.e., to attempt to carefully select contributing studies that are as similar as possible), we believe that certain types of between-study differences can actually help us simultaneously understand both within-study and between-study differences in our findings. We next turn to a closer examination of potential sources of between-study heterogeneity that are likely to arise in many areas of psychological research.

Potential Sources of Between-Study Heterogeneity

When two or more independent data sets are pooled within an IDA framework, one should closely consider the combination of study characteristics that uniquely defines each individual study. For the purposes of study integration, we are typically most interested in controlling for these differences, so that we may obtain findings that are maximally externally valid (e.g., Shadish, Cook, & Campbell, 2002). For the purposes of study comparison, we may be directly interested in between-study heterogeneity as a means of testing the generalizability of our findings. Whether for purposes of control or exploration, identifying important sources of between-study heterogeneity is a critical aspect of IDA. This process is complicated, because some sources of between-study heterogeneity are confounded and thus cannot be disentangled (i.e., geographic influences and ethnicity cannot be distinguished in pooled analyses of a study of Caucasian youths in Indiana and a separate study of Latino youths in Arizona). Another complicating factor is that there are multiple sources of heterogeneity that must be considered simultaneously, but many of these may interact with one another in potentially complex ways. Fortunately, it is not necessary in IDA to independently model and identify all sources of between-study heterogeneity for purposes of study integration. Rather, we can use techniques that control more globally for between-study differences to obtain findings across studies and to determine in which studies these findings hold.

However, for purposes of study comparison, we can use information about between-study heterogeneity to model potential moderating influences on the generalizability of our findings to the extent that we are aware of and able to

[3] This does not include the effects of the 2009 Recovery and Reinvestment Act federal stimulus package that was passed in the same week that we completed this article.

identify and operationalize these sources of variance. Understanding the points of convergence and divergence in findings among a set of studies can often inform our understanding of findings within the individual studies themselves. As we describe later, not only can between-study heterogeneity be directly factored into many IDA applications but some of these study-to-study differences may be of substantive interest in their own right. Indeed, this latter point is what makes IDA such an intriguing endeavor. Although there are many sources of between-study heterogeneity that we might consider, here we focus on five: sampling, geographic region, history, design characteristics, and measurement.

Heterogeneity Due to Sampling

As we explore throughout our article, one of the key benefits of IDA is that it prompts us to think more carefully about important issues that typically receive limited or no attention in single-sample analysis. One prime example is sampling. By sampling we mean the implicit or explicit mechanism by which a sample of observations is drawn from a targeted population with the goal of making inferences back to a population (e.g., Cochran, 1977). On only one occasion in the past several decades has either of us been asked to address sampling issues within any of our single-study papers or grant applications; in contrast, this issue has been raised in some form or another on every single manuscript that has come out of Cross Study. Sampling is clearly an important issue within all areas of psychological research, but this is particularly salient within the IDA framework.

Briefly, there are two general approaches to sampling. The first approach is *probability sampling,* in which all members of a defined population have some known probability of being selected into the sample; examples of this type include simple, systematic, cluster, and stratified sampling. Because the probability of selection is known, one or more sampling weights are typically available for each individual included in the sample (e.g., Pfeffermann, 1993). The second approach is *nonprobability sampling,* in which some (or, more typically, all) members of the population have an unknown probability of being selected into the sample; examples of this type include convenience, snowball, and quota sampling. Because the probability of selection is unknown, no sampling weights are available for individuals included in samples obtained with nonprobability procedures. The majority of research conducted in the psychological sciences is characterized by nonprobability sampling (Sterba, 2009).

Which statistical methods are needed for making valid inferences from the sample back to the population depends upon which sampling mechanism was used to obtain the sample. There are two approaches that characterize nearly all research applications in the social sciences. The first is the *model-based* procedure, proposed by Fisher (1922), which can be used to make population inferences based upon samples that were drawn using a nonprobability framework.[4] Fisher invoked three conditions to allow for this: a structural model must be hypothesized, a parametric distribution must be assumed, and any design characteristics impacting sample selection must be included in the model (e.g., oversampling or cluster sampling). The second approach is the *design-based* procedure, developed by Neyman (1934), which can be used to make population inferences based upon samples that were drawn using a probability sampling framework. Neyman developed these methods to overcome what he viewed as inappropriate subjectivity that was inherent in the model-based approach, particularly with respect to the selection of the hypothesized model and assumed distribution. In contrast to the model-based approach, the design-based approach allows for valid inferences to be made through the direct incorporation of the individual-specific sampling weights into the statistical analysis.[5]

Most important for our discussion here, combining data from two or more studies within the IDA framework provides an exciting opportunity to examine these same sampling issues much more carefully than was previously possible. In particular, IDA offers the potential to conduct direct empirical evaluations of the effects of heterogeneity in sampling mechanism in the pooled sample that cannot be conducted within any single-sample analyses. From a practical standpoint, the first step is to determine what type of sampling mechanism was used within each study that is to be pooled within the IDA. In some situations this mechanism may be unambiguous, such as when data were drawn from probability-design-based studies such as Add Health (Harris et al., 2008) or the National Longitudinal Survey of Youth (Bureau of Labor Statistics, U.S. Department of Labor, 2002); for data such as these, one or more sets of sampling weights, strata indicators, and cluster indicators will likely be available that can potentially be used in the pooled IDA. However, in other situations the sampling mechanism may be less clear, such as when data were drawn from a nonprobability design in which subjects were sampled from an undergraduate psychology subject pool or a volunteer sample was identified via publicly posted fliers. More often than not, the available data sets will likely fall

[4] Model-based procedures can also be used with probability samples if the selection mechanism is explicitly incorporated into the fitted model (Sterba, 2009).

[5] For more detailed discussions about the similarities and differences between model-based and design-based procedures, as well as current hybrid designs that combine the two approaches, see Guo and Hussey (2004); R. J. A. Little (2004); Lenhard (2006); Muthén and Satorra (1995); and Sterba (2009).

between these two extremes. For example, all three contributing samples used in Cross Study were nonprobability samples, yet all three oversampled on certain known demographics (e.g., parental alcoholism) and all three incorporated strict inclusion and exclusion criteria. This information can then be directly incorporated into a model-based framework that fulfills Fisher's (1922) necessary conditions for valid inference.

The critical point to appreciate about the role of sampling in IDA is that our goal is not necessarily to unequivocally establish that all of the component samples are precisely equivalent with respect to sampling prior to proceeding to tests of our substantive questions. This is sometimes a misplaced goal in IDA (e.g., to include only samples that are deemed functionally identical to one another). Instead, we must test for potential differences across the samples, incorporate adjustments for these differences, and, if possible, come to some understanding about why such differences exist. Indeed, meaningful between-study heterogeneity with respect to sampling mechanism not only may be incorporated within the IDA framework but may provide a unique opportunity to empirically evaluate the role of sampling in ways that would not be possible within typical single-study designs.

Heterogeneity Due to Geographic Region

Although in principle two or more studies might have been independently conducted within the same geographic region, this is not likely to be the case. Indeed, even if two independent studies had been conducted in New York City, the samples might not be comparable as a function of the specific borough in which they were conducted; if they had been conducted in the same borough, they might not be comparable as a function of neighborhood (and so on). To further complicate matters, sampling mechanism and geographic region are closely intertwined and often completely confounded in most IDA applications. For example, if data are drawn from three independent studies, it is likely not only that each individual study incorporated a unique sampling mechanism but that each study was located in a distinctly different geographic region. In such situations, it may not be possible to disentangle differences due to sampling from differences due to region. One exception is the use of multiple sites within a single study design. When data are pooled from a multisite study, so-called site differences offer unique insights into heterogeneity in characteristics associated with region while they hold the sampling mechanism constant. However, we must jointly consider sampling and geographic location in the majority of IDA applications.

When evaluating potential between-study heterogeneity associated with geographic location, one should consider which characteristics of the specific location may be responsible for the observed differences across the studies. There may be nothing inherently meaningful in the direct comparison of a sample of individuals drawn from Denver with a comparable sample of individuals drawn from Indianapolis. Instead, the challenge is to identify the specific characteristics that serve to distinguish Denver from Indianapolis that are manifested within the data set. There are likely many such characteristics: ethnic composition, median income, availability of social services, per capita rates of crime, seasonal weather patterns, and proximity to neighboring urban centers, to name just a few. It may be difficult even to identify a discrete region within which a given study was conducted. For example, one study may have recruited subjects from all incoming college students at a single university, whereas another study may have recruited subjects from incoming college students at all the universities within a single state. In this case, there is not a comparable geographic location with which to contrast the two samples; all that might be deduced is that the samples were obtained from two different sources. It is not possible to move beyond this gross level of assessment.

Heterogeneity Due to History

Whereas heterogeneity due to region is associated with potential differences in place, heterogeneity due to history is associated with potential differences in time. We use the term *history* here to capture the essence of this concept as it arises within quasi-experimental designs (e.g., Shadish et al., 2002). Broadly speaking, history as a threat to internal validity refers to any events that might have occurred during a study that could have accounted for an observed effect. Although this concept is implicitly longitudinal (i.e., an event that occurred during a study), history can also play an important role when one conducts an IDA of pooled cross-sectional (i.e., single time point) data sets. For example, if two cross-sectional studies were conducted 10 years apart, there could be historical influences that differentially impacted each of the two studies (e.g., two studies on general anxiety conducted pre- and post-9/11; two studies on minority aspirations conducted pre- and post-President Obama).

A key question that arises within the IDA framework is whether some effect that is observed in the pooled sample can be at least partially attributable to between-study heterogeneity associated with historical period. Although there are many conceivable measures of time (e.g., chronological age, time since diagnosis, age of a particular event), here we primarily focus on calendar time. In other words, are there influences associated with the particular year (or month or day) on which a given subject was assessed? Just as it is likely that two or more independent studies were conducted with different sampling mechanisms within different geographic regions, it is likely that the studies were conducted within different historical periods. It is thus important to consider the potential ways in which the effects of history might be manifested in both cross-sectional and longitudinal IDA applications. Although historical influences can be present when IDA is used to evaluate pooled cross-sectional data, this tends to be a far greater challenge when two or more independent longitudinal data sets are being pooled.

When pooling cross-sectional data, one must evaluate the comparability of two samples of individuals who were assessed at different points in time. For example, in a cross-sectional IDA of adolescent drug use, one might need to consider the impact of historical context as a function of whether the subjects were assessed in 1970, 1980, or 1990. That is, did societal norms, legal sanctions, drug processing, and popular routes of administration differ across these 3 decades in a way that might influence the very meaning of drug use within each of the independent samples? Epidemiological data shows that the annual prevalence of illicit drug use among graduating high school seniors was 54% in 1979, 35% in 1989, and 42% in 1999 (Johnston, Bachman, & O'Malley, 2006); given this, one must take care not to treat the absolute measure of drug use as necessarily equivalent when pooling data collected over this 20-year period. Drug use is just one example, and similar issues can arise across a whole host of psychologically relevant outcomes that might be of interest with a cross-sectional IDA application.

Outcomes are embedded in time when longitudinal data are pooled, and this further complicates potential history effects. In longitudinal IDA, heterogeneity across studies must then be considered in over-time trajectories rather than cross-sectional levels of behavior over time. Extending the cross-sectional example above, we might need to compare developmental trajectories of drug use for individuals who matured through adolescence during the 1970s in one study, the 1980s in a second study, and the 1990s in a third study. This issue is closely related to the classic age–period–cohort distinction first raised nearly 50 years ago in developmental psychology (e.g., Schaie, 1965). Briefly, Schaie argued that to understand the development of an individual over time, one must simultaneously take into account the individual's chronological age, the historical period in which the individual was assessed, and the birth cohort to which the individual belongs. Although initial attempts were made to simultaneously disentangle these three influences, later work suggested that knowledge of any two dimensions of time defined the third (e.g., knowing the individual's age at assessment and the year of assessment in turn defines the cohort in which the individual was born; Palmore, 1978; Schaie, 1994). Thus, we must often choose the historical unit of interest (e.g., age, period, cohort), given our methodological and theoretical application. As such, a common goal for those using an IDA of pooled longitudinal data sets is to evaluate and potentially control for between-study heterogeneity associated with the historical period, however indexed, during which a set of repeated observations was obtained.

Heterogeneity Due to Other Design Characteristics

Another likely source of between-study heterogeneity is study-to-study differences in characteristics of the study design. Such design characteristics may include methods of data collection and sample retention, which in turn exert an important impact on the sampled data. As a sobering example,[6] Harford (1994) found that discrepancies in the order and style of two items assessing alcohol use at two different periods of time resulted in a large and observable change in the reported levels of heavy alcohol use. This finding demonstrates that something as simple as altering the order of presentation of two items at two points in time can lead to substantial differences in how subjects respond to the items, even within the very same study. Given this, care must clearly be taken in pooling across multiple studies that may differ in structural design characteristics.

Yet here lies another challenge. It is typically possible to generate an extraordinarily long list of potentially meaningful differences between all of the contributing studies, and this list grows exponentially as the number of individual studies increases. Examples are abundant. Were subjects assessed face-to-face or via telephone or Internet? If the assessment was face-to-face, was it conducted at the home of a subject or was the subject brought into a controlled facility? Was the assessment conducted as a personal interview, or was it computer-based? Was the same assessment battery used at each time period, or were items deleted and others added over time? And on and on and on. This list-generation exercise is not dissimilar to "medical school hypochondriasis," in which first-year medical students diagnose themselves as having each new condition they learn about in class.

Clearly, there are an unmanageable number of differences across studies in design characteristics. In most typical IDA applications these differences are confounded with one another. Thus, it is not only unrealistic but also not useful to exhaustively identify, track, and code the entire set of differences in design characteristics across the set of contributing samples. A more useful goal is to identify those specific characteristics that are thought to be most salient for the given application at hand. The selection of these characteristics can be guided by the same factors related to drawing valid inferences from between-group single-study designs (Shadish et al., 2002). For example, for studies on a sensitive topic, such as illegal or high-risk behaviors, were individual assessments conducted with face-to-face interviews or with confidential, computer-assisted personal in-

[6] Pardon the pun.

terview methods? For studies on diagnostic criteria, were all items presented to all subjects or could individual subjects "skip out" of a set of items that were deemed irrelevant as a function of the individual's response to some qualifying item? For studies on in-group/out-group behavior, were subjects assessed individually or in the presence of other subjects? Our goal here is not to develop an all-inclusive catalog of potential design-related differences. Instead, the goal is to identify those study-to-study characteristics that are most likely to be related to the specific constructs that are under study. These measures can then be included in the analytic strategies we describe later.

Heterogeneity Due to Measurement

We have saved what is often the most important and yet most challenging source of between-study heterogeneity for last: heterogeneity due to measurement. Here we use the classic definition offered by Stevens (1946) that broadly establishes measurement as "the assignment of numerals to objects or events according to rules" (p. 677). Of course, measurement plays a critical role in nearly all areas within the psychological sciences, and the field of psychometrics is characterized by more than a century of work in this domain (e.g., Cudeck & MacCallum, 2007). In most IDA applications, a key goal is to optimally capture the measurement of specific theoretical constructs both within each sample individually and, more important, within the aggregated sample as a whole. Indeed, measurement might be considered the most fundamental source of between sample heterogeneity in IDA, because the reliability and validity of the analytic results rely directly on the reliability and validity of the contributing measures drawn from each individual study.

As with other factors that arise in IDA, issues of measurement within the aggregated sample often prompt us to more closely consider these issues within each contributing study. Our motivating goal here is to develop an analytic framework that allows us to create a valid and reliable aggregate measure that is sensitive to potential study differences on dimensions such as design characteristics, specific items administered, subject age, and calendar year. As we will see, in many IDA applications we will be able to implement such a framework, although naturally this comes at a cost. Simultaneously balancing the need to maximize the quality of measurement and the need to minimize the associated cost is one of the most salient challenges of IDA. Here we briefly explore two closely related measurement issues that are often of greatest importance in IDA: measurement invariance and measurement comparability.[7] (Further discussions of these important issues are presented in Bauer & Hussong, 2009, and McArdle, Grimm, Hamagami, Bowles, & Meredith, 2009, and in other articles from our project; see Curran, Edwards, Wirth, Hussong, & Chassin, 2007; Curran et al., 2008, Flora, Curran, Hussong, & Edwards, 2008.)

To think more specifically about these measurement issues, let us again consider the Cross Study project in which we estimated developmental trajectories of internalizing symptomatology using data pooled from three separate studies (Curran et al., 2008). For our measurement models we used 21 items drawn from the BSI and CBCL. In the first study subjects responded to all 21 items from the BSI and CBCL; in the second study subjects responded to 10 of the 15 CBCL items and none of the BSI items; and in the third study subjects responded only to the BSI items. On the basis of our theoretical perspective, we believe that there is some underlying individual-specific latent propensity for a child to experience depressive symptomatology and that this latent propensity is manifested in the child's response to a particular set of items. In our study we had a set of internalizing symptomatology items shared across all three studies as well as items that were unique to a given study or studies (e.g., 3 items were administered in just one study, 14 items in two studies, and 4 items in all three studies). Despite these differences, all three studies attempted to assess precisely the same underlying latent construct. The concept of measurement invariance applies to the studies that share an item set; the concept of measurement comparability applies to the studies that use a unique item set.[8]

Measurement invariance. Generally speaking, measurement invariance addresses the extent to which a set of items reliably and validly assesses an underlying construct in a similar (if not identical) fashion across groups or over time. Classic examples of measurement invariance include examining gender differences or racial differences in the psychometric properties of a particular measure (e.g., Rusticus, Hubley, & Zumbo, 2008) and examining developmental differences in the expression of a set of behaviors over time (e.g., Pentz & Chou, 1994). Nearly all prior research on measurement invariance has focused on differences across groups or over time within a single sample of observations (but see Nesselroade, Gerstorf, Hardy, & Ram, 2007, for a recent discussion of idiographic approaches to measurement invariance). However, we can draw on this literature to apply concepts of measurement invariance across groups and over time not only within a particular

[7] The term *measurement comparability* is not widely used in the traditional invariance literature, but the concept becomes an important issue when one considers the psychometric equivalence of the assessment of a theoretical construct across multiple independent studies.

[8] This distinction between invariance and comparability is an oversimplification, in that both concepts will ultimately apply to both situations. We believe, however, that this distinction offers a helpful starting point from which to understand the core issues at hand.

study but, more important, between two or more independent studies. Indeed, given sufficient empirical data we can examine the interaction between measurement invariance across group or over time with study group membership. For example, on Cross Study we were able to explicitly compare differences in the magnitude of noninvariance in the assessment of depressive symptomatology as a function of age and gender across our three contributing studies (Curran et al., 2008; Flora et al., 2008).

The topic of measurement invariance itself is extremely broad, and seminal contributions include Thurstone (1947); Horn and McArdle (1992); Millsap (1995, 1997); and Meredith (1993). There are actually many different types of measurement invariance that have been proposed over the years, and different terminologies have been used to describe similar types (Vandenberg & Lance, 2000). Approaching measurement using concepts drawn from factor analysis can help us understand these types of invariance. Within a factor analytic framework, a set of items (or indicators) is used to define an underlying latent variable, the existence of which is believed to have given rise to the pattern of observed correlations among the items (e.g., Bollen, 2002). The items are linked to the underlying factor via the factor loadings; these serve as partial regression coefficients that index the extent of change in the individual items resulting from a one-unit change in the latent factor. Each item is also defined by an item-specific intercept and residual variance, and each factor (or latent variable) is defined by a mean and variance. The different types of invariance conditions (e.g., strong, weak, configural) are then defined with respect to the equality (i.e., invariance) or inequality (i.e., noninvariance) of these model parameters.

At the extreme, strict invariance (Meredith, 1993; also called complete invariance; Millsap, 1995) means that all the parameters that define the measurement model are equal across group or over time. In our earlier example of depression, a measurement structure that is strictly invariant indicates that boys and girls express depression identically, and thus a single set of parameters validly defines the assessment of this construct for both genders. However, if one or more parameters are found to differ across group or over time, the items are characterized by weaker forms of invariance (e.g., configural, pattern, scalar). Given a partially noninvariant measurement structure, we would conclude that boys and girls express depression in a functionally different way and that, most important, different parameters are needed to validly define the assessment of depression within each gender. If noninvariance is ignored (whether within a single sample or within the pooled sample), we cannot unambiguously establish whether an observed relation between depression and some other construct is valid or is instead an artifact of imposing an improper measurement model; see Meredith (1993); Millsap (1995); and Vandenberg and Lance (2000) for further details.

The importance of this situation within the IDA framework is highlighted in an extreme hypothetical example in which there is strict measurement invariance across all observed groups and over all time points within each independent contributing study. Although there is unequivocal evidence for strict invariance within each sample, this is not sufficient to imply that there is measurement invariance across the pooled set of samples. This may occur, for example, if measurement invariance holds within a given range of age (e.g., invariance holds within each of three studies that assessed subjects from ages 5 to 10, ages 10 to 15, and ages 15 to 20, respectively) yet invariance does not hold across the full pooled range of ages (e.g., from 5 to 20). Thus, for any number of reasons, children might express depression in a functionally different way across studies even though they responded to precisely the same set of items. However, this situation becomes even further complicated when subjects in one study respond to a different set of items than do subjects in another study. This brings us to the topic of measurement comparability.

Measurement comparability. Measurement comparability is less well studied relative to measurement invariance, yet it is no less important. The reason that it has received much less attention is that this issue does not often arise within a single study. In typical applications a single sample of subjects responds to a given set of items, and questions of invariance arise with respect to how the shared set of items might operate differently across group or over time. But when IDA is applied to an aggregated sample comprising two or more independent samples, it is common to encounter the use of partially or wholly different scales to assess a shared underlying construct. In this situation, the core issues that constitute classic measurement invariance simply do not apply. In the classic invariance scenario, the hypothesis being tested is that a given set of items relates to the underlying construct in equivalent ways across groups and over time. Thus, it is illogical to raise the question of whether Item 1 is related to the underlying construct in the same way as is Item 2 across two separate studies; in many cases this is analogous to comparing apples with oranges, and our classic methods of assessing invariance are not applicable.

Much more is known about this issue in the field of education using IRT, particularly as applied to standardized testing of academic skills. For example, it is often theorized that there is some underlying individual-specific latent math ability, but this ability is assessed with fundamentally different items across grade (e.g., first by addition, then fractions, then algebra, then calculus). A vast literature exists that focuses on methods in IRT that allow for test equating, scaling, and linking to deal with these complex issues in practice (see, e.g., Kolen & Brennan, 2004; Thissen & Wainer, 2001). The very same issues arise in many (if not most) IDA applications, yet several factors arise that make dealing with them particularly challenging. Issues such as small sample sizes, small numbers of shared items, and multidimensional factor structures combine to limit the use of standard IRT equating and linking procedures (Curran et al., 2007, 2008). Given these current limitations, we must think extremely carefully about issues related to both measurement invariance and measurement comparability when combining multiple data sets into one. We cannot simply compute the standardized mean of 10 items in Study A and compute the standardized mean of 15 items in Study B and assume that these are equivalent measures of the same underlying construct within the pooled sample. We must instead consider all theoretical and empirical evidence that can strengthen our confidence in whether we are assessing the same construct within each individual sample as well as within the pooled sample in a psychometrically equivalent way.

General Analytic Strategies

We have reviewed five potential sources of between-study heterogeneity that can arise when applying IDA to pooled data: sampling, geography, history, other design characteristics, and measurement. There are many other sources that we have not addressed here. However, in our own research, we have sought to address each of these sources of between-study heterogeneity and, in doing so, have come to see them as among the most likely to be encountered in IDA applications. To reiterate an important earlier point, it is not paramount that all of the contributing samples be precisely comparable on all possible dimensions to allow for a valid analysis of the aggregated data. Indeed, this was certainly not the case with Cross Study, in which our three studies differed along all five dimensions. As we will see, there are several analytic strategies that allow us to directly test and incorporate potential between-study differences into the analysis of pooled data; this will be done in a way that is consistent with Fisher's model-based procedures, which we described earlier. Further, better understanding potential sources of between-study heterogeneity may offer us unique insights into each sample individually while we endeavor to generalize our findings in the aggregated sample as a whole. Strategies for addressing these sources of heterogeneity in IDA are newly emerging, and we next summarize several general approaches to IDA that address one or more sources of between-study heterogeneity.

Evaluating Heterogeneity Due to Study-Specific Characteristics

We can think of our available set of independent samples that are to be combined within the IDA in two distinct ways. The first is to conceptualize the collection of data sets as randomly drawn realizations from a homogeneous population of data sets; we refer to this approach as random-effects IDA. The second is to treat the collection of data as fixed as known; we refer to this approach as fixed-effects IDA. There are advantages and disadvantages of each approach, the relative utility of which ultimately depends on the specifics of the IDA application at hand. We briefly explore each in turn.

Random-effects IDA. In any given single study, we typically assume, the sample of observations is randomly selected from some larger (if not infinitely large) population. Random sampling is one of the cornerstones of inferential statistics that permit probabilistic inferences to be drawn from samples to populations. We can extend this concept within the IDA framework, in which we consider our multiple samples to *themselves* be randomly drawn from some larger population. In this way we literally have a random sample of random samples. This concept is similar to the standard multilevel model, in which a sample of independent sampling units (ISUs; e.g., schools) is randomly selected and individual observations are then randomly selected within ISU (e.g., students within schools). Here, independent studies are randomly sampled from a population of studies, and individual subjects are sampled within each study. This two-stage sampling introduces two potential sources of variability into our observed data: variability due to the sampling of studies and variability due to the sampling of individual observations within study (Raudenbush & Bryk, 2002). Therefore, we can potentially approach IDA from a multilevel perspective.

There are two issues that we must consider here. First, we must somehow establish, even if in theory only, that the multiple data sets can be meaningfully considered as representing random draws from a homogeneous population of data sets. If the independent samples are deemed uniquely and distinctly different from one another in theory or design, they should likely not be treated as random draws from a single population; the fixed-effects IDA approach should considered instead. Second, from a more practical perspective, there must be a sufficient number of independent samples to allow for the reliable estimation of the random variability both within and between the samples. There are no infallible rules that dictate how many independent samples are sufficient to allow for proper estimation of the random effects, but in the general multilevel framework 20 to 30 are often viewed as a minimum (e.g., Kreft & de Leeuw, 1998). Although there are certainly potential applications of IDA that would have access to this many independent data sets, most applications within the social sciences would typically be based on substantially fewer samples than this (e.g., in Cross Study we combined just 3 data sets, yet this still provided nearly 2,000 individual observations followed over nearly a 40-year period). As before, if an insufficient number of studies is available to allow for the estimation of a random-effects IDA, we would instead need to consider a fixed-effects approach.

There are, however, a variety of significant advantages to conducting a random-effects IDA if the necessary data are available. Most important, if the multiple samples can be treated as randomly drawn from a homogeneous population, we can consider incorporating study-level predictors to model between-study variability on the outcome measure of interest. This is precisely analogous to having multiple students nested within one of multiple schools and being able to disaggregate student-level effects, school-level effects, and the Student × School cross-level interactions (e.g., Bauer & Curran, 2005). In IDA, we have multiple subjects nested within one of multiple samples, and we are able to disaggregate subject-level effects, sample-level effects, and Subject × Sample cross-level interactions. Examples of sample-specific measures might include the type of sampling mechanism that was used, the geographic location of the study, and whether data were collected via personal interview or over the Internet. A multilevel model could be estimated that simultaneously evaluates the main effects of the within-sample predictors on the outcome, the main effects of the between-sample predictors on the outcome, and the interaction between the within-sample and between-sample predictors on the outcome. Not only is sample-to-sample heterogeneity directly incorporated into the overall model but this becomes an important empirical and theoretical question in its own right.

As an important final point, although an anonymous reviewer noted that our above arguments were compelling, the reviewer also felt that the conceptualization of a truly random-effects IDA was "something of a stretch." The reviewer's primary concern was that because individual studies are so complexly determined on so many different dimensions, it may be inherently impossible to consider these as truly random realizations from a population of potential studies. We agree. However, we also agree with the reviewer's subsequent suggestion that, under the right circumstances, a random-effects IDA allows for direct estimation of the between-study variability and that such estimation could be quite useful if for nothing other than a description of the relative heterogeneity among a set of studies. That is, although we might not want to make direct inferences back to some hypothetical population of studies, the random-effects IDA nonetheless provides intriguing insights into the similarities or dissimilarities among a set of studies that would otherwise not be possible. That said, most applications of IDA in the social sciences may simply lack the necessary number of individual data sets to support the estimation of the random-effects model. We thus must turn to the fixed-effects counterpart.

Fixed-effects IDA. Within random-effects IDA, we treat each data set as an independent random draw from a homogeneous population of data sets. Within fixed-effects IDA, we instead treat study membership as a fixed and known characteristic of each individual observation nested within that study. To accomplish this, we simply incorporate one of several available coding schemes (e.g., dummy codes, effect codes, weighted effect codes) to denote study membership as a fixed characteristic of each individual observation. This is precisely the same strategy we might use to incorporate gender or ethnicity as a fixed characteristic of a given individual; however, here the fixed characteristic of the individual is the study to which he or she belongs. For example, for any given observation, our design matrix might denote a particular individual as male, African American, and belonging to Study 4. These dummy- or effect-coded variables are then entered as predictors in our fitted models in a way consistent with Fisher's (1922) model-based inferential methods, as described earlier. A key advantage of this strategy is that we can also estimate multiplicative interactions between individual characteristics (e.g., gender, ethnicity) and study group membership. This in turn allows for the potential of differential impact of individual characteristics on the outcome across the set of studies. We have used precisely this strategy extensively on Cross Study (e.g., Hussong, Cai, et al., 2008; Hussong, Flora, et al., 2008; Hussong et al., 2007).

There are several critical distinctions between the random-effects and fixed-effects approaches. Most important, in the random-effects framework we treat the set of independent samples as random draws from a population and can thus (in principle) make inferences back to an infinite population of samples. In contrast, in the fixed-effects framework we treat the set of independent samples as fixed and known and are thus able to make inferences back only to the specific samples under study. In our view, this is likely a more realistic goal in many IDA applications in psychology. Further, within the random-effects framework we can explicitly disaggregate within-sample effects, between-sample effects, and cross-level interactions. In contrast, with the fixed-effects approach we treat each sample as fixed and known, such that the inclusion of the set of study membership variables removes all between-sample sources of variability from the model (although there are some situations in which more complex contrast coding schemes could be used to partially disaggregate these effects; see, e.g., Maxwell & Delaney, 2004).

Thus, whereas in the random-effects model we can estimate the effects of one or more sample-specific measures (e.g., sampling mechanism or geographic region), in the fixed-effects model it is typically not possible to include sample-specific measures once the dummy- or effect-coded variables have been entered into the model. This last issue can be simultaneously viewed as a limitation and an advantage of the fixed-effects approach. The obvious limitation is that finer distinctions cannot be made among specific characteristics that define each unique sample, because all between-sample variability has been removed from the model. But the associated advantage is that, because the entry of the effect codes eliminates all between-sample sources of variability, any between-sample differences are controlled even if specific measures regarding these differences are not available. Given the plethora of potential sources of between-sample heterogeneity that exist, controlling for all of these simultaneously can be both beneficial and efficient.

Evaluating Heterogeneity Due to Historical Time

Our discussion thus far has focused on fixed- and random-effects models for evaluating between-study heterogeneity associated with study-specific design characteristics (e.g., differences due to sampling mechanism, geographic location). These techniques can equivalently be applied to pooled samples that consist of either cross-sectional or longitudinal data. However, several intriguing opportunities arise when considering the role of time when pooling longitudinal data, particularly when there is an interest in disaggregating within-person from between-person differences over time (e.g., Baltes, Reese, & Nesselroade, 1988). Indeed, we can use a fixed-effects strategy for incorporating information about historical time, as we did earlier for study-specific heterogeneity. To accomplish this, we draw on methods currently available for testing and combining multiple cohorts within a single study (e.g., Miyazaki & Raudenbush, 2000). However, here we must deal with the added complexity that the multiple cohorts are drawn from multiple studies, and this in turn introduces the possibility of a Cohort × Study interaction.

To begin, consider a single longitudinal study that consists of three repeated measures on a sample of children who were 11 to 15 years of age at first assessment (e.g., Chassin et al., 1991). In this situation a cohort-sequential design would typically be used, so that developmental trajectories could be estimated between 11 and 17 years of age. This allows for the estimation of trajectories over a 7-year age span, despite the fact that any given child was assessed only three times (e.g., Mehta & West, 2000). In this situation there are five cohorts (ages 11 to 15 at first assessment), three periods (the first, second, and third assessments), and seven ages (11 to 17 years). Miyazaki and Raudenbush (2000) provided an excellent general discussion of this type of design and proposed an analytic method for testing age, cohort, and Age × Cohort interactions.

Within the IDA framework, we can extend these existing methods in two ways. First, we can adopt the approach of Miyazaki and Raudenbush (2000), in which each specific cohort is dummy-coded and entered as a predictor into the model. A series of nested models is then estimated to evaluate the extent to which a single trajectory underlies the set of repeated observations or if cohort-specific trajectories are needed. The straightforward expansion we can introduce is the estimation of the interactions between cohort and study membership. This provides a direct test of the extent to which cohort may be differentially related to the outcome within each specific study. These differences could each be formally tested and retained to statistically control for Cohort × Study interactions prior to evaluating key theoretical predictors of interest.

A second extension we can consider capitalizes on the potential for large numbers of distinct cohorts that may be available as a result of pooling multiple longitudinal data sets. Because Miyazaki and Raudenbush (2000) considered a single data set with seven distinct cohorts, they logically chose to treat these as discrete groups and thus used a nominal coding scheme of dummy variables and orthogonal contrasts. However, if a sufficient number of cohorts are available, these could instead be included as a quantitative variable that would then allow for additional ways to evaluate age-related and cohort-related change.[9] For example, within the final pooled sample on Cross Study there were 29 unique birth cohorts spanning from 1964 to 1992. Because of the large number of unique birth years, it is possible to include birth year as a quantitatively scaled between-subjects predictor in subsequent models. This in turn introduces the exciting possibility of simultaneously incorporating two measures of time in a single model: between-subject functions of historical time (as measured by birth year) and within-subject functions of developmental time (as measured by chronological age). Given sufficient data, these two dimensions of time could be allowed to interact with one another as well as with study group membership. This would allow for the testing of a variety of important hypotheses in ways not previously possible.

Evaluating Heterogeneity Due to Measurement

There are two general situations in which issues of measurement arise in most IDA applications. The first situation is when studying the measurement properties of a set of items or tests, such that measurement itself is the primary question of interest. For example, multiple independent studies might be pooled to allow for the examination of the factor structure underlying some set of items that assess constructs such as personality traits, intelligence, or psychopathology. The second situation is when measurement is used more as a means to some other end, such that the goal is to produce a reliable and valid scale score that could then be used in other types of analyses. For example, a confirmatory factor analysis (CFA) model might be fitted to a set of items to create a scale score that would then be used as a criterion measure in a separate multilevel model (e.g., Curran et al., 2008). These two applications of measurement are highly related, in that one must examine the measure-

[9] We thank Dan Bauer for the original development of this idea.

ment properties prior to computing an associated scale score. Despite the obvious potential applications of the first situation, here we focus on the second, given the likely broader use of this approach in many IDA applications in the social sciences.

One of the reasons that measurement is so critically important within the IDA framework relates to our goal of making valid inferences back to theory based upon the empirical characteristics of the pooled data set. We must establish that we are measuring the same theoretical construct in the same manner for all individuals across all data sets. How to best accomplish this in large part depends upon the specific characteristics of the data sets to be pooled. For example, say that the theoretical construct of interest is childhood depression. In an ideal situation, all studies have used precisely the same items, response scales (e.g., 1 to 10 Likert scale), and time frame (e.g., past 30 days) to measure depression. A less ideal situation is one in which all studies have used precisely the same items but some studies have used different response scales (1 to 5 Likert scale vs. 1 to 10 Likert scale) and some studies have used different time frames (e.g., past 30 days vs. past 60 days). Finally, the most typical situation probably is one in which each study has used some unique combination of items, response scales, and time frames. Different analytic strategies are more or less well suited for handling these different combinations of item characteristics across studies (see Bauer & Hussong, 2009, and McArdle et al., 2009, for further discussions of these issues).

To establish the optimal measurement strategy, one first identifies the set of available items that is believed to assess the underlying construct pooling across the set of contributing studies. It is extremely helpful if at least some portion of this pooled item set contains items that are shared across all contributing studies. These shared items form a set of "anchor" items that can be of great use in later stages of measurement development (Kolen & Brennan, 2004). For example, there may be an item that is worded in precisely the same way across all studies (e.g., "My child cries easily"); this could unambiguously serve as a potential anchor item. However, there may be items that are not worded in precisely the same manner but that might nonetheless be operating in a psychometrically similar fashion. For example, several studies might have phrased the item in slightly different ways (e.g., "My child cries easily," "My child cries often," "My child cries more frequently than other children"). Although the wording may appear to be closely related across the three items, these differences may or may not be sufficient for these items to be treated as separate indicators. Finally, there will be items that are found in one or a small number of contributing studies and that were clearly not assessed in other studies. For example, a single contributing study may have assessed the item "My child feels sad even in the presence of others"; however, from a theoretical standpoint this item is a valid indicator of childhood depression that should be retained if possible.

Selection of the proper statistical measurement model depends in large part on the form of the response scales that were used for each individual item. The standard linear CFA model is well suited for items that are (at least approximately) interval scaled, although there is some controversy regarding what constitutes "approximately" (e.g., Muthén & Kaplan, 1985, 1992). If individual items do not sufficiently approximate an interval scale, two or more individual items can be combined to create item parcels that better correspond to the assumed continuous distribution (e.g., T. D. Little, Cunningham, Shahar, & Widaman, 2002; MacCallum, Widaman, Zhang, & Hong, 1999). Alternatively, if the items are discretely scaled (e.g., binary or a small number of ordinal responses) and the creation of item parcels is not a viable option, the assumption of linearity underlying the standard CFA model is not well met and a nonlinear measurement model is necessary (Flora & Curran, 2004). There are two key options in this situation: nonlinear factor analysis (NLFA; e.g., Skrondal & Rabe-Hesketh, 2004) and item response theory (IRT; e.g., Thissen & Wainer, 2001). There is a vast literature dedicated to the NLFA and IRT models; a thorough review of more general issues in nonlinear measurement models is given in Wirth and Edwards (2007), and we have discussed these issues as they relate to applications within IDA in Bauer and Hussong (2009) and Curran et al. (2007, 2008).

Once the potential item pool has been identified and the optimal statistical model has been chosen, actual model fitting can begin. As with many other points in our discussion, the specific steps to be taken will depend on the goals and unique characteristics of the given IDA. However, in most applications there are four general steps in the measurement portion of the analysis (see Curran et al., 2008, for a more detailed discussion of these steps). First, some assessment must be made of the dimensionality underlying the set of items. The standard IRT model assumes unidimensionality, although recently developed techniques allow for the estimation of multidimensional IRT models under certain circumstances (e.g., Fox & Glas, 2001; Rabe-Hesketh, Skrondal, & Pickles, 2004). In contrast, the CFA and NLFA models typically allow for either unidimensionality or multidimensionality. Second, measurement models are first fitted within each study separately and then across all studies simultaneously to establish an initial understanding of the psychometric properties of the scales. These properties take the form of factor loadings and item intercepts in the factor analysis models and of discrimination and severity parameters in the IRT model. Third, some type of assessment is needed of measurement invariance across study, across demographic group, or over time. Multiple group analysis provides such tests in the factor analysis framework, and this is accomplished by using differential item functioning in the IRT model. Bauer and Hussong (2009) and McArdle et al. (2009) described other options for evaluating invariance in IDA.

Finally, once a comprehensive measurement model has been established, scale scores are calculated that are jointly based on the observed pattern of responses to the items and the parameter estimates from the final measurement model. These scale scores can be calculated by using one of several available factor score estimates in the factor model (e.g., Grice, 2001) and by using posterior modal estimate or posterior mean estimate scoring in the IRT model (e.g., Thissen & Wainer, 2001). Regardless of approach, the motivating goal is to create a person-specific scale score that incorporates information about study group membership and, potentially, demographic group membership; these scale scores can be used in subsequent statistical analyses.

Conclusion

We view IDA as a product of our times. These methods respond to an increased demand for collaborative efforts that make efficient use of limited resources in the pursuit of a cumulative science. Moreover, these methods call for the integration of cutting-edge techniques concerning longitudinal modeling and measurement evaluation that in tandem have the power to address many of the vexing problems of IDA surrounding study integration and study comparison. We have attempted to outline the advantages of IDA and the types of applications for which IDA may be particularly useful. As we note at the outset, we recognize that IDA will not be possible in all applications. In our own work, we have succeeded in conducting IDA across our three contributing studies in some instances (e.g., Curran et al., 2008), have needed to drop to using only two studies in others (e.g., Hussong, Cai, et al., 2008; Hussong, Flora, et al., 2008), and have relied on parallel, single study analysis on another occasion (Hussong, Bauer, et al., 2008). Thus, in our experience, the feasibility of IDA rests on the characteristics of the contributing studies as they bear on specific questions of theoretical interest. Understanding the boundaries of IDA within those applications where raw data are available for studies of homogenous populations is an important area of further development for these methods.

The focus here on analysis of existing data is merely a starting point for IDA. Collaborative models for IDA that use primary data collection efforts in psychology come from existing multisite designs (e.g., Hofer & Picinnin, 2009). However, we believe that IDA has much more to offer in study design that is yet unrealized. Greater attention to assessing study characteristics, and not simply participant characteristics, would greatly aid efforts to understand between-study sources of heterogeneity. Measurement approaches that are tailored to the target sample of a given study but also include linkage items across studies create important opportunities for IDA. Discipline and field-level efforts to coordinate research may offer enormous potential for IDA applications, such as the suggested core items for assessing alcohol use put forth by the National Institute on Alcohol Abuse and Alcoholism (2003). Such efforts to provide a core but not necessarily exclusive set of items for assessment permit greater study comparison and still retain the need for study innovation. As such, overlapping but nonredundant item sets are often ideal for IDA. Regardless of specifics, efforts to reach beyond single study design planning are clearly needed to facilitate the future integration of findings pooling across multiple studies, and IDA can be an important part of those efforts.

References

Achenbach, T., & Edelbrock, C. (1978). The classification of child psychopathology: A review and analysis of empirical efforts. *Psychological Bulletin, 85,* 1275–1301.

American Association for the Advancement of Science, Intersociety Working Group. (2008). *AAAS Report XXXIII: Research and Development FY 2009* (AAAS Publication No. 08–1A). Washington, DC: American Association for the Advancement of Science.

American Heritage dictionary of the English language, fourth edition. (2006). Retrieved January 29, 2009, from http://dictionary.reference.com/browse/integrate

American Psychological Association. (2003). Ethical principles of psychologists and code of conduct. *American Psychologist, 57,* 1060–1073.

Baltes, P. B., Reese, H. W., & Nesselroade, J. R. (1988). *Life-span developmental psychology: Introduction to research methods.* Hillsdale, NJ: Erlbaum.

Bauer, D. J., & Curran, P. J. (2005). Probing interactions in fixed and multilevel regression: Inferential and graphical techniques. *Multivariate Behavioral Research, 40,* 373–400.

Bauer, D. J., & Hussong, A. M. (2009). Psychometric approaches for developing commensurate measures across independent studies: Traditional and new models. *Psychological Methods, 14,* 101–125.

Berlin, J. A., Santanna, J., Schmid, C. H., Szczech, L. A., & Feldman, H. I. (2002). Individual patient- versus group-level data meta-regressions for the investigation of treatment effect modifiers: Ecological bias rears its ugly head. *Statistics in Medicine, 21,* 371–387.

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology, 53,* 605–634.

Bureau of Labor Statistics, U.S. Department of Labor. (2002). *National Longitudinal Survey of Youth 1979 cohort, 1979–2002 (rounds 1–20)* [Data file]. Columbus: Center for Human Resource Research, Ohio State University.

Chassin, L., Rogosch, F., & Barrera, M. (1991). Substance use and symptomatology among adolescent children of alcoholics. *Journal of Abnormal Psychology, 100,* 449–463.

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: Wiley.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155–159.

Cooper, H. M., Hedges, L. V., & Valentine, J. C. (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). New York: Russell Sage Foundation.

Cooper, H. M., & Patall, E. A. (2009). The relative benefits of meta-analysis conducted with individual participant data versus aggregated data. *Psychological Methods, 14,* 165–176.

Cudeck, R., & MacCallum, R. C. (2007). *Factor analysis at 100: Historical developments and future directions.* Mahwah, NJ: Erlbaum.

Curran, P. J., Edwards, M. C., Wirth, R. J., Hussong, A. M., & Chassin, L. (2007). The incorporation of categorical measurement models in the analysis of individual growth. In T. Little, J. Bovaird, & N. Card (Eds.), *Modeling ecological and contextual effects in longitudinal studies of human development* (p. 89–120). Mahwah, NJ: Erlbaum.

Curran, P. J., Hussong, A. M., Cai, L., Huang, W., Chassin, L., Sher, K. J., & Zucker, R. A. (2008). Pooling data from multiple prospective studies: The role of item response theory in integrative analysis. *Developmental Psychology, 44,* 365–380.

Derogatis, L. R., & Spencer, P. (1982). *Brief Symptom Inventory (BSI*). Baltimore: Clinical Psychometric Research.

DeRubeis, R. J., Gelfand, L. A., Tang, T. Z., & Simons, A. D. (1999). Medications versus cognitive behavior therapy for severely depressed outpatients: Mega-analysis of four randomized comparisons. *American Journal of Psychiatry, 156,* 1007–1013.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, Series A, 222,* 309–368.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9,* 466–491.

Flora, D. B., Curran, P. J., Hussong, A. M., & Edwards, M. C. (2008). Incorporating measurement non-equivalence in a cross-study latent growth curve analysis. *Structural Equation Modeling, 15,* 676–704.

Fox, J.-P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika, 66,* 271–288.

Gans, H. J. (1992). Sociological amnesia: The noncumulation of normal social science. *Sociological Forum, 7,* 701–710.

Glass, G. V. (1976). Primary, secondary, and meta-analysis. *Educational Researcher, 5,* 3–8.

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods, 6,* 430–450.

Guo, S., & Hussey, D. L. (2004). Non-probability sampling in social work research: Dilemmas, consequences, strategies. *Journal of Social Service Research, 30,* 1–18.

Harford, T. C. (1994). The effects of order of questions on reported alcohol consumption. *Addiction, 89,* 421–424.

Harris, K. M., Halpern, C. T., Entzel, P., Tabor, J., Bearman, P. S., & Udry, J. R. (2008). *The National Longitudinal Study of Adolescent Health: Study design.* Retrieved February 12, 2009, from http://www.cpc.unc.edu/projects/addhealth/design

Higgins, J. P. T., Whitehead, A., Turner, R., Omar, R. Z., & Thompson, S. G. (2001). Meta-analysis of continuous outcome data from individual patients. *Statistics in Medicine, 20,* 2219–2241.

Hofer, S. M., & Picinnin, A. M. (2009). Integrative data analysis through coordination of measurement and analysis protocol across independent longitudinal studies. *Psychological Methods, 14,* 150–164.

Horn, J. L., & McArdle, J. (1992). A practical guide to measurement invariance in research on aging. *Experimental Aging Research, 18,* 117–144.

Hunter, D. J., Spiegelman, D., Adami, H. O., Beeson, L., van den Brandt, P. A., Folsom, A. R., et al. (1996). Cohort studies of fat intake and the risk of breast cancer: A pooled analysis. *New England Journal of Medicine, 334,* 356–361.

Hunter, J. E., & Schmidt, F. L. (1996). Cumulative research knowledge and social policy formulation: The critical role of meta-analysis. *Psychology, Public Policy, and Law, 2,* 324–347.

Hussong, A. M., Bauer, D. J., Huang, W., Chassin, L., Sher, K. J., & Zucker, R. A. (2008). Characterizing the life stressors of children of alcoholic parents. *Journal of Family Psychology, 22,* 819–832.

Hussong, A. M., Cai, L., Curran, P. J., Flora, D. B., Chassin, L. A., & Zucker, R. A. (2008). Disaggregating the distal, proximal, and time-varying effects of parent alcoholism on children's internalizing symptoms. *Journal of Abnormal Child Psychology, 36,* 335–346.

Hussong, A. M., Flora, D. B., Curran, P. J., Chassin, L. A., & Zucker, R. A. (2008). Defining risk heterogeneity for internalizing symptoms among children of alcoholic parents: A prospective cross-study analysis. *Development and Psychopathology, 20,* 165–193.

Hussong, A. M., Wirth, R. J., Edwards, M. C., Curran, P. J., Chassin, L. A., & Zucker, R. A. (2007). Externalizing symptoms among children of alcoholic parents: Entry points for an antisocial pathway to alcoholism. *Journal of Abnormal Psychology, 116,* 529–542.

Johnston, L. D., Bachman, J. G., & O'Malley, P. M. (2006). *Monitoring the Future: Questionnaire responses from the nation's high school seniors, 2005.* Ann Arbor, MI: Institute for Social Research.

Kaplow, J. B., Curran, P. J., Dodge, K., & the Conduct Problems Prevention Research Group. (2002). Child, parent, and peer predictors of early-onset substance use: A multi-site longitudinal study. *Journal of Abnormal Child Psychology, 30,* 199–216.

Kiecolt, K. J., & Nathan, L. (1985). *Secondary analysis of survey data.* Thousand Oaks, CA: Sage.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer.

Kreft, I., & de Leeuw, J. (1998). *Introducing multilevel modeling.* London: Sage.

Lambert, P. C., Sutton, A. J., Abrams, K. R., & Jones, D. R. (2002). A comparison of summary patient-level covariates in mega-regression with individual patient data meta-analysis. *Journal of Clinical Epidemiology, 55,* 86–94.

Lenhard, J. (2006). Models and statistical inference: The controversy between Fisher and Neyman-Pearson. *British Journal of Philosophy of Science, 57,* 69–91.

Lerer, B., Segman, R. H., Fangerau, H., Daly, A. K., Basile, V. S., & Cavallaro, R. (2002). Pharmacogenetics of tardive dyskinesia: Combined analysis of 780 patients supports association with dopamine D3 receptor Gene Ser9Gly polymorphism. *Neuropsychopharmacology, 27,* 105–119.

Little, R. J. A. (2004). To model or not to model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association, 88,* 546–556.

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling, 9,* 151–173.

Lorenz, F. O., Simons, R. L., Conger, R. D., Elder, G. H., Johnson, C., & Chao, W. (1997). Married and recently divorced mothers' stressful events and distress: Tracing change across time. *Journal of Marriage and the Family, 59,* 219–232.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4,* 84–99.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9,* 147–163.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Erlbaum.

McArdle, J. J., Grimm, K. J., Hamagami, F., Bowles, R. P., & Meredith, W. (2009). Modeling life span growth curves of cognition using longitudinal data with multiple samples and changing scales of measurement. *Psychological Methods, 14,* 126–149.

McArdle, J. J., Hamagami, F., Meredith, W., & Bradway, K. P. (2000). Modeling the dynamic hypotheses of Gf-Gc theory using longitudinal life-span data. *Learning and Individual Differences, 12,* 53–79.

McArdle, J. J., Prescott, C. A., Hamagami, F., & Horn, J. L. (1998). A contemporary method for developmental–genetic analyses of age changes in intellectual abilities. *Developmental Neuropsychology, 14,* 69–114.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46,* 806–834.

Mehta, P. D., & West, S. G. (2000). Putting the individual back into individual growth curves. *Psychological Methods, 5,* 23–43.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58,* 525–543.

Millsap, R. E. (1995). Measurement invariance, predictive invariance, and the duality paradox. *Multivariate Behavioral Research, 30,* 577–605.

Millsap, R. E. (1997). Invariance in measurement and prediction: Their relationship in the single factor case. *Psychological Methods, 2,* 248–260.

Miyazaki, Y., & Raudenbush, S. W. (2000). A test for linkage of multiple cohorts from an accelerated longitudinal design. *Psychological Methods, 5,* 44–63.

Muthén, B. O., & Kaplan, D. (1985). A comparison of some methodologies for the factor-analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology, 38,* 171–180.

Muthén, B. O., & Kaplan, D. (1992). A comparison of some methodologies for the factor analysis of non-normal Likert variables: A note on the size of the model. *British Journal of Mathematical and Statistical Psychology, 45,* 19–30.

Muthén, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology, 25,* 267–316.

National Institute on Alcohol Abuse and Alcoholism. (2003). *Task Force on Recommended Alcohol Questions, National Council on Alcohol Abuse and Alcoholism recommended sets of alcohol consumption questions, October 15–16, 2003.* Retrieved February 9, 2009, from http://www.niaaa.nih.gov/Resources/Research Resources/TaskForce.htm

Nesselroade, J. R., Gerstorf, D., Hardy, S. A., & Ram, N. (2007). Idiographic filters for psychological constructs. *Measurement, 5,* 217–235.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society, 109,* 558–606.

Palmore, E. (1978). When can age, period, and cohort be separated? *Social Forces, 57,* 282–295.

Pentz, M. A., & Chou, C. P. (1994). Measurement invariance in longitudinal clinical research assuming change from development and intervention. *Journal of Consulting and Clinical Psychology, 62,* 450–462.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review, 61,* 317–337.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika, 69,* 167–190.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). *Publication bias in meta-analysis: Prevention, assessment and adjustments.* West Sussex, England: Wiley.

Rusticus, S. A., Hubley, A. M., & Zumbo, B. D. (2008). Measurement invariance of the Appearance Schemas Inventory—Revised and the Body Image Quality of Life Inventory across age and gender. *Assessment, 15,* 60–71.

Schaie, K. (1965). A general model for the study of developmental problems. *Psychological Bulletin, 64,* 92–107.

Schaie, K. (1994). Developmental designs revisited. In S. H. Cohen & H. W. Reese (Eds.), *Life-span developmental psychology: Theoretical issues revisited.* Hillsdale, NJ: Erlbaum.

Schmidt, F. L. (1984, August). *Meta-analysis: Implications for cumulative knowledge in the behavioral and social sciences.* Paper presented at the annual convention of the American Psychological Association, Toronto, Ontario, Canada.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1,* 115–129.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal* inference. Boston: Houghton Mifflin.

Sher, K. J., Walitzer, K. S., Wood, P. K., & Brent, E. E. (1991). Characteristics of children of alcoholics: Putative risk factors, substance use and abuse, and psychopathology. *Journal of Abnormal Psychology, 100,* 427–448.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal and structural equation models.* Boca Raton, FL: Chapman & Hall/CRC.

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist, 32,* 752–760.

Stark, L. J., Janicke, D. M., McGrath, A. M., Mackner, L. M., Hommel, K. A., & Lovell, D. (2005). Prevention of osteoporosis: A randomized clinical trial to increase calcium intake in children with juvenile rheumatoid arthritis. *Journal of Pediatric Psychology, 30,* 377–386.

Sterba, S. K. (2009). *Alternative inferential frameworks for nonprobability sampling in psychology and probability sampling in allied fields: Polarization to reconciliation.* Manuscript submitted for publication.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103,* 677–680.

Thissen, D.,& Wainer, H. (2001). *Test scoring.* Mahwah, NJ: Erlbaum.

Thurstone, L. L. (1947). *Multiple-factor analysis.* Chicago: University of Chicago Press.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement in-variance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3,* 4–69.

van den Brandt, P. A., Spiegelman, D., Yaun, S. S., Adami, H. O., Beeson, L., Folsom, A. R., et al. (2000). Pooled analysis of prospective cohort studies on height, weight, and breast cancer risk. *American Journal of Epidemiology, 152,* 514–527.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods, 12,* 58–79.

Zucker, R. A., Fitzgerald, H. E., Refior, S. K., Puttler, L. I., Pallas, D. M., & Ellis, D. A. (2000). The clinical and social ecology of childhood for children of alcoholics: Description of a study and implications for a differentiated social policy. In H. E. Fitzgerald, B. M. Lester, & B. S. Zuckerman (Eds.), *Children of addiction: Research, health and policy issues* (pp. 174–222). New York: Garland Press.

Received June 26, 2008
Revision received March 12, 2009
Accepted March 12, 2009 ■