Integrative Data Analysis in Clinical Psychology Research

Andrea M. Hussong, Patrick J. Curran, and Daniel J. Bauer

Department of Psychology, University of North Carolina at Chapel Hill, North Carolina 27599; email: hussong@unc.edu, curran@unc.edu, dbauer@email.unc.edu

Annu. Rev. Clin. Psychol. 2013. 9:61–89

First published online as a Review in Advance on February 1, 2013

The *Annual Review of Clinical Psychology* is online at http://clinpsy.annualreviews.org

This article's doi:
10.1146/annurev-clinpsy-050212-185522

Keywords

measurement harmonization, pooled data analysis, collaborative science, data sharing, research methods, cumulative science

Abstract

Integrative data analysis (IDA), a novel framework for conducting the simultaneous analysis of raw data pooled from multiple studies, offers many advantages including economy (i.e., reuse of extant data), power (i.e., large combined sample sizes), the potential to address new questions not answerable by a single contributing study (e.g., combining longitudinal studies to cover a broader swath of the lifespan), and the opportunity to build a more cumulative science (i.e., examining the similarity of effects across studies and potential reasons for dissimilarities). There are also methodological challenges associated with IDA, including the need to account for sampling heterogeneity across studies, to develop commensurate measures across studies, and to account for multiple sources of study differences as they impact hypothesis testing. In this review, we outline potential solutions to these challenges and describe future avenues for developing IDA as a framework for studies in clinical psychology.

Contents

INTRODUCTION

With the accrual of high-quality databases, both within our national archives and individual laboratories, and the economic pressures of big science research to do more with less, the scientific community is looking for innovative methods that leverage existing resources to answer novel questions. Responsive to this call, methodologists from different fields are developing multiple approaches for pooled data analysis that combine information collected across multiple studies into a single analytic design. These methods have been used to examine the efficacy of medications versus cognitive behavior therapy for severe depression (DeRubeis et al. 1999), the relation between fat intake and breast cancer (Hunter et al. 1996), the pharmacogenetics of tardive dyskinesia (Lerer et al. 2002), the relation of height, weight, and breast cancer risk (van den Brandt et al. 2000), and the mediators of fluoxetine effects on youth suicidal ideation (Gibbons et al. 2012). Not surprisingly, these methods may share little in their analytics beyond the common goal of data pooling. Together, however, they form a toolkit for researchers interested in analyzing pooled data. We offer to this toolkit an approach that we call integrative data analysis (IDA). IDA is a framework for conducting the simultaneous analysis of raw data pooled from multiple studies. Our goals for this review are to define the IDA framework for pooled data analysis; describe the advantages and limitations of this approach; discuss sampling, measurement, and hypothesis testing within the IDA framework; and provide future directions for IDA applications and methodological development.

Integrative data analysis (IDA): a framework for conducting the simultaneous analysis of raw data pooled from multiple studies

A Cumulative Approach to Science

> It is simply a sad fact that in soft psychology theories rise and decline, come and go, more as a function of baffled boredom than anything else; and the enterprise shows a disturbing absence of that *cumulative*

> character that is so impressive in disciplines like astronomy, molecular biology, and genetics. (Meehl 1978, p. 807; italics as in original)

Although Meehl's reflection references the science of 34 years ago, lament over the lack of rapid progress in psychology and the need for more collaborative, cumulative efforts to advance our field have echoed throughout the interceding decades (e.g., Schmidt 1996). Given the current information explosion, the challenge of our science is clearly not in simply amassing findings but in creating a coherent body of knowledge that considers the strengths and weaknesses of contributing studies to provide answers to the core questions of our era.

Literature reviews have been a fundamental, classic technique for study integration, serving a variety of purposes including generalization of findings, resolving conflicts in the literature, providing a linguistic bridge across studies, critiquing methodology and theory, and identifying key issues and future directions for research (Cooper 2010). However, given a burgeoning number of studies in clinical psychology, literature reviews are often limited in their ability to adequately describe each contributing study and instead rely on prototype studies, box score summaries (contrasting the raw number of supportive versus nonsupportive results), or selective coverage of the literature. These approaches make integrative statements that subsume the broader literature difficult to support and critique. Even when possible, such integrative statements are unable to estimate the strength of a relationship and instead concern whether (and not how much of) a relationship exists in the literature.

In response to these limitations, meta-analysis emerged as an alternative approach and has since become an important part of the toolkit for pooled data analysis (e.g., Cooper et al. 2009). In its simplest form, meta-analysis produces a weighted average effect size for a given association of interest across multiple studies based on summary statistics from each contributing study. Current applications of meta-analysis may also take into account characteristics of contributing studies (e.g., sample size, instrumentation differences) in deriving the weighted average effect size, providing more nuanced interpretations of the synthesized findings. Despite the many advantages of meta-analysis (see Cooper et al. 2009), a key limitation for certain pooled data analysis problems is their reliance on summary statistics derived from completed studies as the unit of analysis, restricting topics of inquiry to those questions that have already been addressed within individual studies.

Although both literature reviews and meta-analysis make significant contributions to the pooled data analysis toolkit, each relies on completed studies to synthesize research findings within a body of work. These approaches are inappropriate for testing novel hypotheses not examined in prior studies. This limitation is noteworthy, as tests of novel hypotheses may play a critical role in synthesizing apparent discrepant findings across studies (e.g., whether discrepancies reflect gender differences that emerge in between-study comparisons). IDA provides an alternate framework for pooled data analysis that retains the rich details of individual studies to simultaneously test novel hypotheses within and across studies. For many questions in clinical psychology, IDA is an important addition to the toolkit for pooled data analysis.

Pooled Data Analysis: Why Now?

Given that the importance of a cumulative approach to science was first heralded long ago, the current push for pooled data analysis may appear woefully belated. However, several current trends in science indicate that now is an important time for the dissemination of techniques for pooled data analysis. We have amassed many matured data sets relevant to the study of clinical psychology and have increased public access to national data archives. At the same time, we have seen rapid advances in statistical methods and related software that facilitate data pooling analyses in the broader research community. Moreover, this amassing of rich data archives and advanced statistical analysis intersects with a scientific zeitgeist emphasizing collaborative (particularly transdisciplinary) efforts, fueling a "big science" initiative.

On the policy side, we have also seen shifts that support greater use of pooled data analysis. For example, a fundamental issue in conducting pooled data analysis is data sharing. In line with ethical guidelines (Am. Psychol. Assoc. 2002), top-tier journals in clinical psychology (e.g., the *Journal of Abnormal Psychology* and the *Journal of Consulting and Clinical Psychology*) and federal funding agencies (Natl. Inst. Health 2003, Natl. Sci. Found. 2011) have long encouraged data sharing to monitor the quality and veracity of published findings. But recent efforts extend the goals of data sharing to generating data structures that encourage pooled data analysis. For example, the National Institutes of Health (NIH) currently provides support for building "metadata structures" through initiatives that "assist in data retrieval and pooled data analysis across sites" (RFA-HD-10-001; Natl. Inst. Child Health Hum. Dev. 2010) and encourages applicants "to collaborate with investigators holding private data sets, use innovative statistical strategies to link methodologically comparable datasets, or utilize public use data readily available" (PAR-10-018; Natl. Inst. Drug Abuse 2010). Other NIH initiatives have established data repositories for genomewide association studies (the Database for Genotypes and Phenotypes; Mailman et al. 2007), autism research (Natl. Database Autism Res. 2011), and clinical trials of substance abuse (Natl. Inst. Drug Abuse Clin. Trials Netw. 2012), seeking to incorporate common measures and data sharing into their research plans from the beginning phases of data collection. Other funded efforts, such as the Collaborative Data Synthesis for Adolescent Depression Trials study funded by the National Institute of Mental Health (Northwest. Univ. Feinberg Sch. Med. 2012) and support of our three-study IDA by the National Institute on Drug Abuse, bolster data pooling efforts as a secondary analysis design. Finally, NIH-supported measurement archives such as the Patient-Reported Outcomes Measurement Information System (PROMIS) and the PhenX toolkit serve as resources for investigators as they initiate new data collection efforts, identifying expert-recommended assessment tools that can be widely used in independent studies to create a potentially broad data base for pooled analysis.

These examples mark a field-wide recognition of the potential benefits of pooled data analysis, though many of the challenges posed by this approach to science remain stumbling blocks for some applications (see sidebar Building an IDA Study Research Team). For researchers able to overcome these challenges, choices about methodological frameworks that will guide their pooled data analysis are still evolving. In the face of such scientific evolution, IDA offers a novel framework for pooled data analysis.

Defining Integrative Data Analysis as a Framework

It would be a mistake to view IDA as a wholly new methodology. Pooled data analysis has a long history, and other techniques for combining raw data have also been introduced in the literature, such as mega-analysis (another approach to the statistical analysis of individual raw-scores from previous studies; McArdle et al. 1998) and individual patient data analysis from the field of medicine (e.g., Ioannidis et al. 1999). Despite these options, research studies in clinical psychology rarely use pooled data analysis (though see Lorenz et al. 1997; McArdle et al. 1998, 2000). Nonetheless, we believe that these techniques, and IDA in particular, offer unique advantages to the field of clinical psychology that may help mitigate endemic problems in studies of abnormal and clinical behavior (Curran & Hussong 2009).

Among data-pooling techniques, relatively unique advantages of IDA primarily result from the level of data pooling; in IDA, pooling is based on raw data (e.g., item responses) from individual participants rather than summary statistics at the level of completed studies (as in meta-analysis).

BUILDING AN IDA STUDY RESEARCH TEAM

Regarding transdisciplinary research teams, Hirsch Hadorn et al. (2008) suggest that building an infrastructure for clear communication and maintaining mutual respect are keys to successful science collaborations. Similarly, IDA studies are potential logistical nightmares without supportive structures that balance the individual needs and demands faced by contributing investigators with those of the IDA team. This is particularly important for academic researchers, for whom the pressures of the academy run counter to the demands of big science collaboration (Mischel 2008, 2009). Issues to consider when building an IDA research team include:

- Determining the relative responsibility and resources available for data preparation given to the original study teams versus the IDA team;
- Distinguishing the aims of research projects within the purview of the original study teams from the unique, value-added aims of the IDA study;
- Coordinating intersecting science teams across manuscripts within the IDA study and the original study teams;
- Acknowledging the contributions of many researchers in a way that balances their diverse needs for publication credit; and
- Communicating about study differences, resolving discrepancies between new IDA study findings and original study publications, and encouraging an open dialogue about replication effects (or the lack thereof) across studies.

This approach yields larger sample sizes than typical single-study designs, a particularly important advantage for examining low-base rate behaviors that are commonly of interest in clinical psychology. Although the pooled data set will have an average base rate that remains within the range of contributing studies (i.e., most simply, each contributing study may have 5% of the sample reporting some form of psychopathology or heavy drug use and thus so will the pooled data set), the absolute number of individuals engaging in the behavior will necessarily be greater in the pooled sample relative to the individual contributing studies. 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.

A second advantage of IDA's approach to pooling data at the item level is increased sample heterogeneity. Although an initial temptation when embarking on IDA is to minimize between-study heterogeneity (i.e., to carefully select contributing studies that are as similar as possible), the presence of certain types of between-study differences can facilitate our ability to distinguish within-study and between-study variation in our findings. For example, many studies in clinical 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). By pooling participants across such samples, the representation of these subgroups may be increased, allowing for distinct groups to be simultaneously considered. Similarly, given adequate sample representation within studies, group comparisons may be possible within IDA that are not possible due to small samples sizes within the individual studies. This in turn increases the external validity of the IDA findings relative to those of the individual contributing studies.

IDA also offers several advantages for measurement, often resulting in a broader and more rigorous psychometric assessment of key constructs. In any single-study design, we typically assess such constructs using a discrete set of items that are shared across all members of the sample (e.g., all subjects respond to the same 10-item scale assessing depression) and selected based on the specific characteristics of a given sample (e.g., age, gender, ethnicity). A common challenge in many areas of psychological research is the need to reconcile the wide array of operationalizations of our constructs across studies and to evaluate the generalizability of our measures across populations of interest. This state of affairs is a clear stumbling block for study-to-study comparison but is in turn a distinct advantage for increased construct validity in IDA. Through data pooling and related measurement development, 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 and by examining the performance of these measures across subpopulations within the pooled sample. This in turn results in much stronger psychometric properties of measures in the pooled design versus any single contributing study.

Importantly, IDA also permits tests of hypotheses that have not or sometimes cannot be tested within a single contributing study (e.g., due to the greater age heterogeneity in the pooled sample created by combining longitudinal studies to cover a broader swath of the lifespan) and directly evaluates replication of hypothesis tests across contributing studies. Unlike other approaches to research synthesis that are based on summary statistics, IDA may be used to test even complex associations among variables (e.g., moderation and mediation) that were not previously tested in contributing studies, providing a means for accelerating the pace of novel hypothesis testing. By accounting for potential between-study sources of heterogeneity, IDA also permits tests of whether these associations differ in magnitude or form over study, a direct evaluation of replication across the pooled studies. Depending on the application, this approach may be extended to explicitly model factors that may account for between-study differences in findings, with potential factors spanning multiple levels of design such as study differences in sampling, geographic region, history, and assessment protocol. Thus, IDA permits an exploration of between-study differences that helps mitigate the need for creating new studies designed to resolve conflicting findings among existing studies posited to result from between-study design differences.

These advantages, however, are not realized in all applications of IDA, and as with any tool for conducting pooled data analysis, IDA is not appropriate for every multistudy application. For example, in our experience, limitations in deriving comparable measures across studies and in identifying age overlap for studies pooled through cohort sequential methods may make IDA infeasible or at least indefensible for some questions. One difficulty in stating when IDA may or may not be useful is that we view IDA as a methodological framework for pooled data analysis rather than a set of specific techniques or analyses. In part, this is because the specific techniques and analyses used in IDA depend on the application at hand. The IDA framework is relevant to testing a variety of questions within clinical psychology, including those about measurement, cross-sectional associations, longitudinal prediction, and treatment effects. Just as we may use a variety of statistical techniques and analyses to answer questions about these issues in a single study, the IDA framework provides a set of guidelines that may also be widely applied across a range of techniques and analyses using a pooled data analysis approach. These guidelines have evolved over time through efforts to conduct IDA with the goal of drawing inferences about substantive hypotheses at the level of the pooled analysis. These guidelines thus reflect ways of addressing challenges in accounting for between-study heterogeneity using IDA. Although in previous work we have discussed five sources of between-study heterogeneity that make IDA challenging (i.e., sampling, measurement, geographical region, history, and design characteristics; Curran & Hussong 2009), we now highlight two of these—sampling and measurement—and consider the implications of these sources of between-study heterogeneity for the ultimate goal of IDA, hypothesis testing.

BETWEEN-SAMPLE HETEROGENEITY DUE TO SAMPLING

One particularly intriguing aspect of IDA is that we are prompted to think more closely about issues not commonly considered in single-study designs. One salient example is sampling. Sampling refers to the mechanism by which a finite group of individual observations is selected from a larger population with the purpose of drawing inferences from the sample back to the population (e.g., Cochran 1977). Whereas the importance of sampling in clinical psychology is often undervalued in single-study designs, between-sample heterogeneity due to sampling in multistudy designs is a significant potential threat to the internal validity of an IDA. Most importantly, we do not want to misidentify effects as theoretically meaningful when they are instead artifacts resulting from differences in sampling composition across contributing studies that were not properly modeled. We can consider two well-developed approaches to sampling: model-based and design-based procedures.

Model-Based and Design-Based Approaches

Like many topics of real importance, there has been substantial disagreement about the best way to construct a sample of observations that optimally reflects the population of interest (for an excellent review, see Sterba 2009). Core to this disagreement is the distinction between model-based and design-based approaches. These two approaches permit us to make inferences from samples back to populations, with differences between the two approaches motivating different sampling designs. Briefly, the model-based approach was first proposed by Fisher (1922), who believed that obtaining a truly random sample from a given population was typically not possible. Instead, Fisher proposed building a statistical model that explicitly linked the substantive theory to the sample data by approximating the mechanism by which the dependent variable was generated. However, the statistical model required the imposition of certain distributional assumptions that many researchers believed to be both subjective and fallible. In response to these concerns, Neyman (1934) developed a design-based approach that introduced randomness via a set of known selection probabilities. This allowed for the selection process to be both controlled and known, avoiding untenable distributional assumptions. The selection probabilities were then used when fitting models to weight the sample in accordance with the characteristics of the population (e.g., Pfeffermann 1993).

The degree to which design- versus model-based approaches are used in practice varies by discipline and the nature of the research question under study, and this is further complicated by recent hybrids of the two approaches (e.g., Lenhard 2006). Nonetheless, model-based sampling procedures implicitly underlie the majority of data sets within clinical psychology. Thus from a strictly practical perspective, most applications of IDA in clinical psychology will likely not pool data from samples in which probability weights are even available.[1] For this reason, we focus our attention on the use of model-based approaches in IDA.

Extending Single-Study Model-Based Approaches to IDA

Whereas sampling of individual observations within a single study is characterized by more than a century of research and development, sampling of studies into an IDA application has received virtually no attention. Nonetheless, we can draw on two well-developed methods commonly used in single-study designs to allow us to explicitly incorporate a model-based approach to sampling in IDA. First, from the field of multilevel modeling (MLM; Raudenbush & Bryk 2002) we can use a two-tiered sampling method to define a random-effects IDA. More specifically, within the

[1] This is not always the case; data sets such as Add Health (Harris et al. 2008) and the National Longitudinal Survey of Youth (Bur. Labor Stat. 2002) were constructed using complex probability sampling designs that could be included in an IDA application. However, how to best incorporate such weights into IDA is currently unclear and remains an important topic for future research.

Random-effects IDA: an approach to IDA in which contributing studies are treated as random draws from a defined population of studies, permitting the use of random-effects models (e.g., multilevel modeling) to estimate study effects as a level of analysis in which individual studies are nested to control for the effects of between-study heterogeneity in hypothesis testing

Fixed-effects IDA: an approach to IDA in which contributing studies are treated as the set of available studies to which inferences will be drawn (rather than as a random sample), permitting the use of fixed-effects models in which study effects are directly modeled as predictors to control for between-study heterogeneity in hypothesis testing

MLM it is possible to first randomly sample a finite number of groups from a population of groups (e.g., schools within a district or hospitals within a county) and then randomly sample a finite number of individuals from a population of individuals who are nested within each group (e.g., students within a school or patients within a hospital). This two-tiered sampling framework allows for the estimation of random components at both the level of the group (level 2: contributing study) and of the individual (level 1: individual observation within study). However, a key practical requirement of this approach is that a sufficient number of groups (in our case, contributing studies) have been sampled to obtain a representative sample of the population and to support the numerical estimation of the group-level random components. There are no firm rules on the required minimum number of groups, but in many situations at least 10 or 20 groups must be observed in order to allow for population representation and proper model estimation.

In the absence of a large number of contributing studies, we may instead prefer a model-based approach in which study membership is treated as a fixed (rather than a random) factor. In fixed-effects IDA, the set of available studies is not construed as a random sample of a broader population of studies but instead constitutes the universe of all of the studies of interest. Following Fisher's (1922) guideline to include all measures associated with the sampling framework when fitting models to the data, we include study membership as an explanatory variable in all analyses. For example, in our work we pool data drawn from three studies, each of which used a high-risk sampling design to oversample children of alcoholic parents. We controlled for selection criteria used to create our fixed sample of studies by creating two dummy-coded variables to capture variance associated with the three contributing studies, and these variables are exogenous predictors in all of our fitted models. This accounts for the sampling framework at the level of the pooled analysis. However, we also implement Fisher's recommendation within each individual study. To do so, we control for factors that influence selection into the individual studies, namely parent alcoholism.

The fixed-effects approach, while powerful and practical, does not allow one to generalize results to a broader universe of studies and instead requires one to limit the evaluation of study differences to the particular studies at hand. Thus, although inferences cannot be made back to the population of studies of children of alcoholic parents, any systematic study-to-study differences are estimated and accounted for when examining the effects of the predictors of substantive interest. This in turn offers a powerful line of protection for internal validity in that substantively meaningful findings are less likely to be due to differences among contributing studies.

For the purposes of study integration, this fixed-effects approach to modeling between-sample heterogeneity accomplishes a primary goal, to control for differences among participants (in this case based on study membership) so that we may obtain findings and draw inferences about associations of theoretical interest that are maximally valid (e.g., Shadish et al. 2002). However, in IDA we may have a second goal of identifying sources of 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. As we describe in a later section, 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.

Summary

In sum, any IDA application must carefully consider both the mechanism that resulted in the sample of contributing studies and in the sample of individual observations within each study. Sampling simultaneously presents a danger and an opportunity within IDA. The danger is primarily a threat to internal validity such that effects perceived to be of theoretical importance may instead be attributable to heterogeneity in sampling mechanisms across the contributing studies. The opportunity is that sampling heterogeneity can become a topic of theoretical interest in and of itself such that gaining a better understanding of study-to-study differences can help us develop a better understanding of the substantive processes under study (e.g., Curran & Hussong 2009). Regardless of intent, sampling is a critical issue that must be closely considered in any application of IDA.

Commensurate measures: measures that have the same meaning and metric across studies despite potentially significant between-study heterogeneity in instruments or modalities of assessment

BETWEEN-SAMPLE HETEROGENEITY DUE TO MEASUREMENT

A second fundamental issue in IDA is measurement. Often an initial stumbling block in conducting IDA is the availability of commensurate measures, which have the same meaning and metric across studies despite potentially significant differences in assessment instruments or modalities. Next, we describe various scenarios reflecting between-study heterogeneity in measurement that vary in their complexity and feasibility for constructing commensurate measures for IDA and provide guidelines for creating commensurate measures.

IDA Measurement Scenarios

It might seem that the ideal IDA scenario occurs when identical measures are used across contributing studies (i.e., all studies measure the same variable in precisely the same way). In this scenario, between-study heterogeneity in measurement may initially seem an irrelevant concern. However, even when identical measures are used across studies, distinct subpopulations may interpret or respond to the same item in different ways, above and beyond any actual differences in the underlying construct. These differences may reflect systemic influences of local norms in how participants view their research participation (i.e., participants in one location or sociocultural context may respond with less veracity than participants in other locations or contexts), in how items are interpreted within the context of the larger assessment battery of each study (e.g., the content of surrounding items; Rivers et al. 2009, Tourangeau et al. 2000), or in how items are administered across study (e.g., by interviewer, paper and pencil, or computer; Meade et al. 2007, Richman et al. 1999). In such cases, despite the fact that the item is identical, the values obtained from the different study samples would not necessarily have an identical scale or meaning.[2]

This underlying issue is often clearer within more complex IDA measurement scenarios. As shown in **Table 1**, an example from our own work involves three studies that each used slightly different ways of measuring the frequency of alcohol consumption: Study 1 assessed a six-month time frame and responses were open-ended, whereas Studies 2 and 3 assessed a 12-month time frame with binned, ordinal response options. Studies 2 and 3, however, each used a different set of frequency bins for the responses. Clearly we cannot simply pool the responses from these three studies given these measurement differences. We can, however, harmonize these items by transforming the original items to have logically equivalent response scales. In this case, some of the response options for assessing the frequency of alcohol use are the same in Studies 2 and 3, and we can collapse other response options to create comparable frequency intervals across studies, thereby obtaining a common set of frequency intervals across the response options for these two studies. For Study 1, we can convert the responses to annualized estimates by multiplying the monthly average by 12 and then bin the responses into the same common set of intervals as

[2] This problem is not unique to IDA. Even within a single study, commensurate measurement cannot always be assumed to hold across subpopulations within the sample (e.g., male and female participants or young and old participants). Like sampling, this issue is simply more salient within the IDA setting.

Table 1 Variation in the measurement of two constructs across three studies

	Study 1	Study 2	Study 3	Harmonized item
	Consumption of alcohol			
Prompt	Over the past 6 months, on the average, how many days a month have you had a drink?	How often did you drink wine or beer or wine coolers in the past year?	Think of all the times in the past year when you had something to drink – how often have you had some kind of beverage containing alcohol?	Past-year frequency of alcohol use
Response scale	Days per month	0. Never 1. 1–2 times 2. 3–5 times 3. More than 5 times but less than once a month 4. 1–3 times a month 5. 1–2 times a week 6. 3–5 times a week 7. Every day	0. Twice a day or more 1. Once a day 2. Nearly every day 3. 3 to 4 times a week 4. Once or twice a week 5. 2 to 3 times a month 6. About once a month 7. 6–11 times a year 8. 1–5 times a year 9. Didn't drink this past year	0. Never 1. 1–5 times 2. 6–11 times 3. 1–3 times a month 4. 1–2 times a week 5. 3+ times a week
	Positive expectancies about alcohol: relaxation			
Prompt	Drinking alcohol makes me relaxed	Drinking alcohol relaxes me	Drinking helps me to relax	Expectation that alcohol helps to relax
Response scale	0. Never 1. Very rarely 2. Rarely 3. Occasionally 4. Frequently 5. Very frequently 6. Always	1. Strongly agree 2. Agree 3. Neither agree nor disagree 4. Disagree 5. Strongly disagree	1. Not at all 2. A little bit 3. Somewhat 4. Quite a bit 5. A lot	?

Studies 2 and 3. Recoding the data this way results in the harmonized item shown in the right column of **Table 1**. The harmonized item is designed to be equivalent across studies in time frame and response options; however, it still may not be truly commensurate because different responses to the item may continue to reflect factors other than actual individual differences in alcohol use. In fact, the assumption that all individuals interpret and respond to the item in the same way is more tenuous in this measurement scenario since the item was not in fact administered in an identical format across studies, enhancing the potential for context effects.

Harmonized item: an item that has been altered within a study to make it comparable to similar items assessed in other studies that will be combined for pooled data analysis (e.g., changes in response scale, or combined items)

A third and still more difficult measurement scenario, again drawing from our own work, involves three studies in which participants were asked whether they believe or expect alcohol to relax them (see bottom half of **Table 1**). The response formats for this item varied greatly between studies, and it is not immediately obvious how to construct a harmonized item. Without additional information to bridge the different item structures (see the later section on bridging studies), there is little justification for pooling these responses to enable IDA.

Commensurate Measures in the IDA Framework

All of the previous scenarios involve single-item measures; often, contributing studies use multi-item scales. In discussing multi-item scales within the IDA framework, we distinguish between

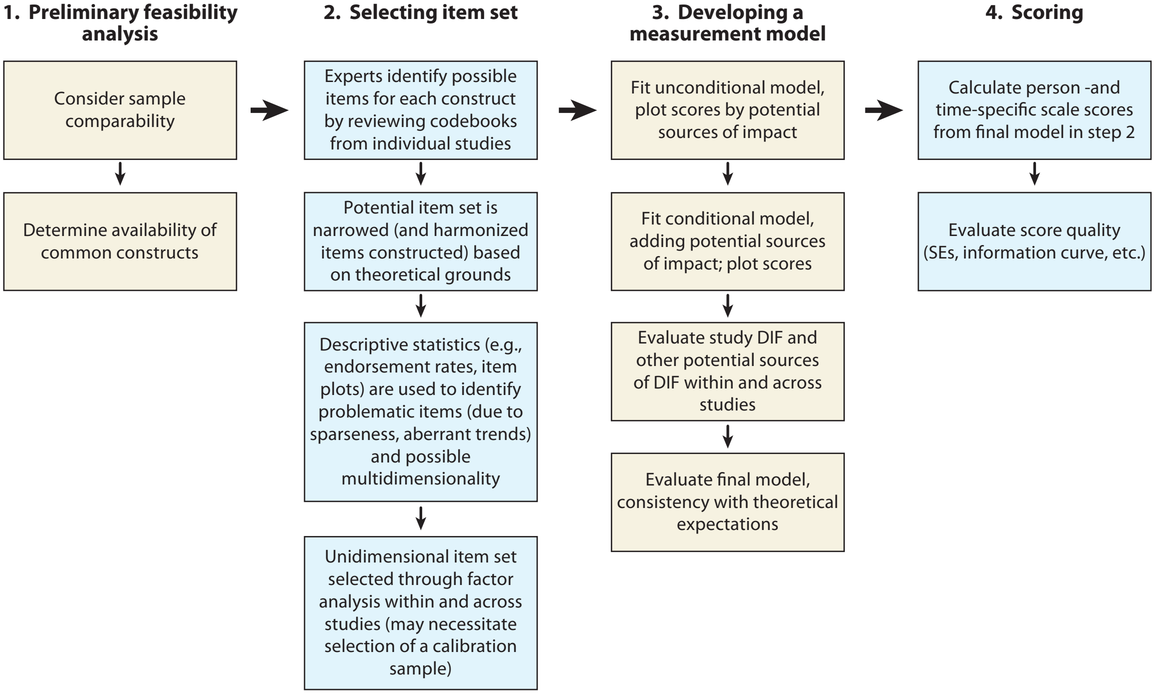
1. Preliminary feasibility analysis
2. Selecting item set
3. Developing a measurement model
4. Scoring
Consider sample comparability
Determine availability of common constructs
Experts identify possible items for each construct by reviewing codebooks from individual studies
Potential item set is narrowed (and harmonized items constructed) based on theoretical grounds
Descriptive statistics (e.g., endorsement rates, item plots) are used to identify problematic items (due to sparseness, aberrant trends) and possible multidimensionality
Unidimensional item set selected through factor analysis within and across studies (may necessitate selection of a calibration sample)
Fit unconditional model, plot scores by potential sources of impact
Fit conditional model, adding potential sources of impact; plot scores
Evaluate study DIF and other potential sources of DIF within and across studies
Evaluate final model, consistency with theoretical expectations
Calculate person -and time-specific scale scores from final model in step 2
Evaluate score quality (SEs, information curve, etc.)

Figure 1

Building commensurate measures for integrative data analysis. DIF, differential item functioning; SEs, score estimates.

creating harmonized items, a term that we use to refer to creating comparable items across studies (by transforming the original item/response scales), and creating commensurate measures, a term that we use to refer to creating scales that have the same meaning and metric across studies (by using measurement models). The process of creating commensurate measures is thus more complex than simple harmonization.

Multi-item scales offer two key advantages for creating commensurate measures. First, for identical or harmonized items, we can test whether participants from different studies (and/or subpopulations within studies) respond to the items in the same way. Even if a subset of items fails this test, we may still be able to construct a commensurate measure while adjusting for between-study heterogeneity in the item responses. Second, we can often retain items from individual studies that cannot be harmonized across studies but that nevertheless provide information to improve measurement within a given study. Both of these advantages are realized by using psychometric models. Accordingly, our strategy for creating commensurate measures borrows heavily from work on measurement invariance testing in factor analysis (Meredith 1993, Millsap & Meredith 2007, Vandenberg & Lance 2000, Widaman & Reise 1997) and on linking and equating test scores in the educational assessment literature (Holland 2007, Holland & Dorans 2006). As shown in **Figure 1**, our analytic guidelines include four key steps: preliminary feasibility analysis, selecting an item set, developing a measurement model, and scoring.

We illustrate these steps through an example from our own work, namely, the scoring of dimensions of internalizing behavior across three independent studies (also see P.J. Curran, J. McGinley, D.J. Bauer, A.M. Hussong, A. Burns, L. Chassin, K. Sher, R.A. Zucker, manuscript in preparation). This project, funded by the National Institute on Drug Abuse, pools data from three existing longitudinal studies that oversampled offspring who had at least one biological alcoholic parent and included matched controls of nonalcoholic parents. These three studies vary in the ages when assessments began, the number of waves and spacing of assessments, and methods of assessment, among other dimensions. They include the Michigan Longitudinal Study (MLS; Zucker et al. 2000), the Adult/Adolescent and Family Developmental Project (AFDP; Chassin et al. 1991), and the Alcohol, Health, and Behavior Project (AHBP; Sher et al. 1991). **Table 2** presents a summary of the pooled sample as a function of study membership, wave of assessment, and chronological age of participants. 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 pooling across study and wave.

Common item: either identical or harmonized items that link measures across studies; common items may be present in all studies, or only pairs of studies, so long as sufficient pairs exist with which to link measurement across studies

Preliminary feasibility analysis. A critical first step in the IDA framework for developing commensurate measures is to conduct a feasibility analysis. This involves assessing the extent of between-study heterogeneity in measurement by defining the measurement scenario present in the pooled data design and evaluating whether and how the construct(s) of interest are measured in each study. IDA requires that at least some common items are present across studies, where common items may be either identical or harmonized items. Common items allow us to link measures across studies. Common items may be present in all studies or only pairs of studies, so long as sufficient pairs exist with which to link measurement across studies (e.g., Studies 1 and 2 share items and Studies 2 and 3 share items such that measurement in Studies 1 and 3 can be linked through Study 2). Unique items, or items present in only one study, do not help in linking measures but still provide useful information for estimating participants' scale scores and more fully assessing constructs of interest. In our example, all three studies included at least one measure of internalizing behavior, with some overlap in item content.

Selecting an item set. The second step in developing commensurate measures is item selection. This includes a review of all individual items assessing a target construct, with the goal of identifying a core set of items that define a unidimensional factor. This core set of items will include both common and unique items. This step involves the use of content experts to identify and narrow the potential item pool, descriptive statistics to identify potentially problematic items, and exploratory factor analyses to evaluate dimensionality. One goal of this step is to identify an item pool that maximizes the number of common items that are core to the construct of interest and provide opportunities to link measurement across studies.

In our example, we selected items assessing internalizing symptoms administered as part of the Child Behavior Checklist (CBCL; Achenbach & Edelbrock 1981), completed by MLS and AFDP participants, and the Brief Symptom Inventory (BSI; Derogatis & Spencer 1982), completed by MLS and AHBP participants. Because MLS participants completed both instruments, the MLS data provide a bridge between item sets. We also harmonized BSI and CBCL items with similar content to create common items across AFDP and AHBP and, given sparseness concerns, we recoded all items to be binary indicators of symptom presence or absence. After harmonization, we identified 33 items that we believed to be theoretically relevant to the construct of internalizing symptoms.

In the next step, we examined descriptive statistics and performed graphical analysis for these items to evaluate endorsement rates by age and study (see **Figure 2**). If all items represent a single factor, plots of endorsement rates should show similar developmental trends, each reflecting the developmental trend of the underlying factor. Similarly, study differences in endorsement rates should be consistent across item plots, reflecting study differences at the level of the factor. Dissimilar study trends in these item plots generally signal either that the item set is not unidimensional or that differential item functioning (DIF) is present. DIF (sometimes called factorial invariance) indicates that a given item does not reflect the underlying factor in the same way across individuals (e.g., at all ages or across all studies); in other words, DIF indicates that the endorsement of an item does not mean the same thing for all individuals, even if the item is worded the same way or has been harmonized to have an equivalent metric. Study DIF could arise due to differences in the harmonized items or simply due to differences in how the sampled populations interpreted and responded to items.[3] **Figure 2** highlights that one item, "overtired," shows an atypical developmental trend relative to other items and exclusively in the AFDP study. Given this aberrant pattern, we removed the item from further consideration.

Remaining differences in item trends were less stark but still may suggest possible multidimensionality or DIF, which we explored further using a nonlinear exploratory factor analysis model to account for the binary nature of the items. Because the factor analysis procedures we used assume that observations are independent of one another, we performed our analyses on a cross-sectional calibration sample drawn by randomly selecting one repeated observation per participant from the longitudinal data.[4] Initial exploratory factor analyses estimated within and across studies suggested that the item set was indeed multidimensional. Factors generally consistent with anxiety and depression emerged; item plots grouped by these two factors also displayed greater homogeneity (i.e., anxiety items showed more similar patterns to one another than to depression items), corroborating a multidimensional structure. Although we could choose to develop separate commensurate measures for both anxiety and depression, here we focus our efforts on developing a commensurate measure of depression from the 17 items that most consistently loaded on the depression factor.

Developing a measurement model. The third step in our strategy was to develop a formal measurement model from which we could generate commensurate scale scores for depression. Although there are several options, we used an extension of traditional factor analytic and item response theory methodology referred to as moderated nonlinear factor analysis (MNLFA; Bauer & Hussong 2009). MNLFA assumes that observations are independent, so these models are also fit to the calibration sample. Given that model complexity in IDA applications of MNLFA can increase rapidly, we prefer a model-building approach beginning with a simple model and adding complexity through iterative model tests.

The first model we fit is an unconditional model, which is a unidimensional factor model without predictors. Unlike traditional factor analysis, the relationships between the factor and the items are not necessarily linear and depend on the scales of the indicators. For instance, for the binary symptom indicators in our example we specified a logistic function, so that the probability of endorsing any given symptom is bounded between zero and one. The latent variable is viewed as a common cause of all of the symptoms (i.e., someone high in latent depression is more likely to endorse multiple symptoms compared to someone low in depression) and is assumed to account for all associations among the symptom indicators. This is a standard assumption of many latent variable models and is often referred to as local independence. Additionally, because depression is a latent variable, we must set its scale, and we do so by setting the mean to zero and variance

Differential item functioning (DIF): when a given item does not reflect the underlying factor in the same way for all people (e.g., at all ages, or across all studies), indicating that the endorsement of an item does not mean the same thing for everyone, even if the item is worded the same way or has been harmonized to have a logically equivalent metric

Calibration sample: a subsample of available observations drawn to meet the assumption of independence (i.e., a single observation per person in the study) for analyses to develop a measurement model prior to scoring

Moderated nonlinear factor analysis (MNLFA): a factor analytic model that allows for nonlinear relationships between the latent factor and the items used to measure it (e.g., a logistic relationship for a binary item) and that also allows the model parameters (e.g., item intercepts, factor loadings, factor mean, and/or factor variance) to vary as a function of one or more observed moderator variables (e.g., study, gender, and/or age)

[3] Although not the focus of between-study heterogeneity in measurement, these plots may also reflect sources of DIF other than study. For example, age DIF may also be observed if opportunities for expressing a behavior change with age independently of changes in the factor itself (e.g., as abilities or norms shift), reflecting a kind of heterotypic continuity.

[4] We selected observations randomly to retain diversity in the calibration sample in age to permit us to test for DIF in items across the full age range of our observed sample. However, there is no need for a calibration sample drawn from the full, scoring sample if the observations are already independent.

Table 2 The Cross Study design

<table>
<tr><th>Age</th><th>2</th><th>3</th><th>4</th><th>5</th><th>6</th><th>7</th><th>8</th><th>9</th><th>10</th><th>11</th><th>12</th><th>13</th><th>14</th><th>15</th><th>16</th><th>17</th><th>18</th><th>19</th><th>20</th><th>21</th><th>22</th><th>23</th><th>24</th><th>25</th><th>26</th><th>27</th><th>28</th><th>29</th><th>30</th><th>31</th><th>32</th><th>33</th><th>34</th><th>35</th><th>36</th><th>37</th><th>38</th><th>39</th><th>40</th></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td colspan="4">MLS-battery 9, $n = 17$</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td colspan="6">MLS-battery 8, $n = 68$</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td colspan="10">MLS-battery 7, $n = 136$</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td colspan="5">MLS-battery 6, $n = 68$</td><td></td><td></td><td></td><td></td><td></td><td colspan="14">AFDP-battery 6, cohort 2; $n = 349$</td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td colspan="5">MLS-battery 5, $n = 510$</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td colspan="9">AFDP-battery 6, cohort 1; $n = 408$</td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td colspan="5">MLS-battery 4, $n = 511$</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td colspan="12">AFDP-battery 5, cohort 2; $n = 345$</td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td colspan="5">MLS- battery 3, $n = 432$</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td colspan="10">AFDP-battery 5, cohort 1; $n = 411$</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td colspan="6">MLS-battery 2, $n = 349$</td><td></td><td></td><td></td><td></td><td></td><td></td><td colspan="11">AFDP-battery 4, cohort 2; $n = 327$</td><td></td><td></td><td></td><td></td><td></td><td colspan="5">AHBP-battery 7, $n = 380$</td><td></td><td></td><td></td></tr>
<tr><td></td><td colspan="6">MLS-battery 1, $n = 371$</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td colspan="7">AFDP-battery 4, cohort 1; $n = 422$</td><td></td><td></td><td></td><td></td><td colspan="6">AHBP- battery 6, $n = 406$</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td colspan="7">AFDP-battery 3, cohort 1; $n = 447$</td><td></td><td></td><td></td><td></td><td colspan="7">AHBP- battery 5, $n = 454$</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td colspan="7">AFDP-battery 2, cohort 1; $n = 449$</td><td></td><td></td><td colspan="6">AHBP-battery 4, $n = 467$</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td colspan="7">AFDP-battery 1, cohort 1; $n = 454$</td><td></td><td></td><td colspan="7">AHBP- battery 3, $n = 468$</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td colspan="7">AHBP- battery 2, $n = 480$</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td colspan="7">AHBP- battery 1, $n = 485$</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
</table>

Age	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**
MLS	11	137	111	91	127	122	107	160	150	132	179	165	164	167	166	152	58	18	22	51
AFDP	---	---	---	---	---	---	---	---	32	107	191	266	294	250	152	64	151	116	123	129
AHBP	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	8	404	472	440	436
Total	11	137	111	91	127	122	107	160	182	239	370	431	458	417	318	224	613	606	585	616

Age	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**	**37**	**38**	**39**	**40**	**Total**
MLS	42	34	24	25	20	11	8	7	1	---	---	---	---	---	---	---	---	---	---	671
AFDP	88	101	164	179	129	89	96	98	147	144	144	108	84	54	38	33	27	11	3	850
AHBP	120	15	288	157	8	4	112	227	59	7	1	33	220	112	12	5	---	---	---	485
Total	250	150	476	361	157	104	216	332	207	151	145	141	304	166	50	38	27	11	3	2006

Note: 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 pooling across study and wave. Abbreviations: AFDP, Adult/Adolescent and Family Developmental Project; AHBP, Alcohol, Health, and Behavior Project; MLS, Michigan Longitudinal Study.

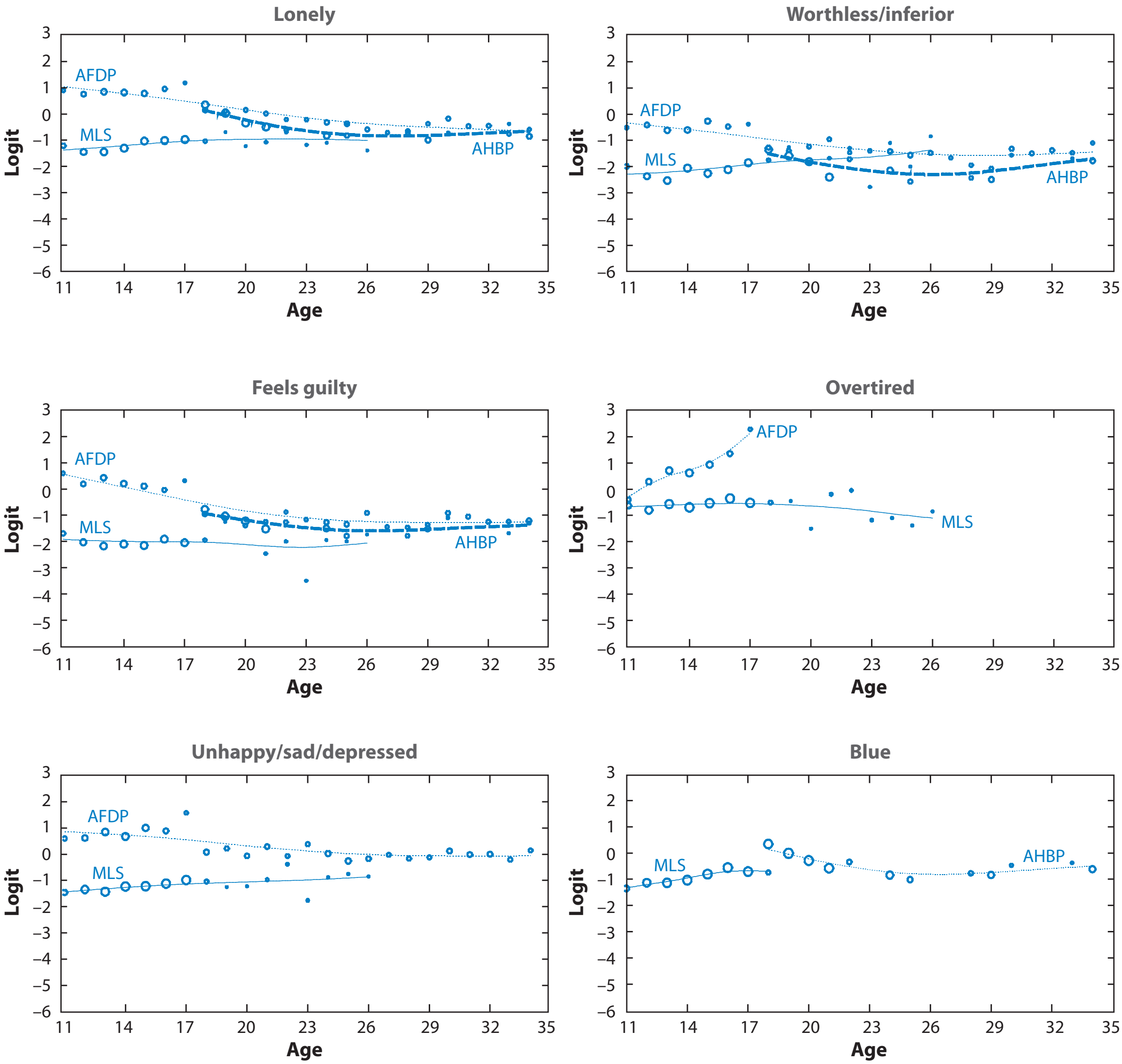

Figure 2

Age and study trends in select items. Note: Endorsement is plotted in the logit scale, or the log of the ratio of the proportion endorsing over the proportion not endorsing the item, since this scale is also used in the models ultimately fit to the data. A logit of one corresponds to a proportion of 0.5, lower logits correspond to lower rates of endorsement, and higher logits correspond to higher rates of endorsement. AFDP, Adult/Adolescent and Family Developmental Project; AHBP, Alcohol, Health, and Behavior Project; MLS, Michigan Longitudinal Study.

to one. Thus the factor scores we generate from this model are on a standard normal metric (see sidebar What Makes a Factor Score?). The resulting model is equivalent to a binary factor analysis with a logit link and also to a two-parameter logistic item response theory model (Takane & de Leeuw 1987).

The unconditional model is a useful starting point, providing preliminary information on the quality of the indicators via inspection of factor loadings, tracelines, and information

WHAT MAKES A FACTOR SCORE?

Because latent variables are unobserved, factor scores for the individual participants are by definition unknown. We can, however, determine the distribution of possible factor scores for each individual. Typically, the expected value or modal value of this distribution is then taken as the best prediction of the factor score, or factor score estimate. Scores based on the expected and modal values are, respectively, known as EAPs (expected a posteriori) or MAPs (modal a posteriori). Both EAPs and MAPs are based on two pieces of information: the item responses of the individual (i.e., the specific symptoms endorsed or not endorsed) and the distribution of the factor in the total population. Each score is pulled (or "shrunken") toward the mean of the population to improve the estimate, especially when there is little information on the individual (e.g., few items or incomplete data). Conceptually, we borrow strength from the whole to improve our estimate for the one. By conditioning the factor mean and variance on covariates, we further refine and improve these estimates by shrinking the scores not toward the grand mean but rather toward the conditional mean, that is, the average for persons who are similar to the target with respect to background characteristics.

functions (P.J. Curran, J. McGinley, D.J. Bauer, A.M. Hussong, A. Burns, L. Chassin, K. Sher, R.A. Zucker, manuscript in preparation). Yet the unconditional model also makes a number of unrealistic assumptions, including that the distribution of the latent variable is homogeneous across subpopulations (i.e., studies). It also assumes that the relationships between the latent variable and the indicators are equal across subpopulations. We test each of these assumptions in turn.

We initially focused on how the distribution of the latent variable may differ across subpopulations. For our example, we expect that depression levels may vary across participants as a function of study, age, history of parental alcoholism, and gender. Such differences are sometimes referred to as impact (Holland & Wainer 1993). To help us better understand potential impact effects, we generated factor scores from the unconditional model and plotted them as a function of the covariates. For instance, to get an initial idea of potential age-related changes in depression, we plotted factor scores as a function of age and study, determining that the age trends in the scores differed across studies but could be well approximated in each study by a cubic function.

Next, we respecified the MNLFA to include effects of predictors on the mean and variance of depression. We specified a linear model for the factor mean and a log-linear model for the factor variance (to prevent negative predicted variances). Not all predictors must appear in the model for the factor mean and variance; rather, in our experience, the variance model is often simpler than the mean model. For our depression example, we detected a cubic age trend on the factor mean, differing by study and gender, and a main effect of parental history of alcoholism. This indicates that changes in depression scores across age follow a cubic trend that differs across study and gender, and that having an alcoholic parent increases mean levels of depression. Regarding the factor variance, we detected an interaction between age and study membership, such that depression levels increased in variability with age in AHBP but not in MLS or AFDP. Bringing these predictors into the model not only makes the specification of the model more realistic (by explicitly incorporating known sources of heterogeneity), but it also provides additional information with which to improve our score estimates for the participants. Indeed, given the incorporation of this extra information, when we generated and plotted factor scores from this model, the plots provided an even stronger visual indication of the covariate effects. **Figure 3**, for instance, depicts study differences in developmental trends, and one can observe both the mean differences across studies and the increasing heterogeneity within AHBP over time.

Impact: when the distribution of a latent variable in an item response theory or factor analysis model differs across subpopulations, typically manifested as covariate effects on the latent variable mean and/or variance (e.g., the distribution of depression may have a higher mean and greater variance for girls than boys)

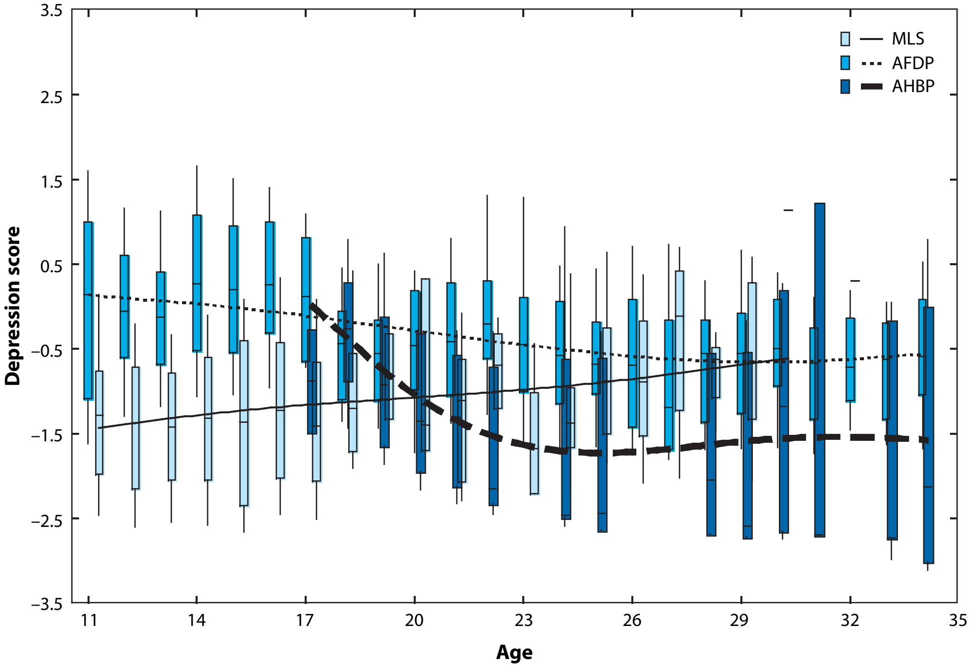

Figure 3

Boxplots of factor score estimates by age and study with mean trends depicted by smoothed lines. Note study differences in mean trends with age, relatively stable score variability in MLS and AFDP, and increasing score variability in AHBP consistent with moderated nonlinear factor analysis results. AFDP, Adult/Adolescent and Family Developmental Project; AHBP, Alcohol, Health, and Behavior Project; MLS, Michigan Longitudinal Study.

These initial effects of covariates focus on study, age, and gender differences in the underlying factor (mean and variance) of depression. Now we return to the items themselves. To this point we have assumed, but have not yet demonstrated, that the items reflect the factor equivalently in all subgroups and across all studies. We identified potential sources of DIF based on our item plots due to participants' age and study, previous literature on gender-related DIF for some items (e.g., Schaeffer 1988, Steinberg & Thissen 2006), and our sampling framework (our oversampled children of alcoholic parents might interpret or respond to some items differently). A clear advantage of the MNLFA framework is that it allows DIF testing across all of these covariates simultaneously, including continuous covariates like age. DIF testing is accomplished by allowing variables indexing these potential sources of DIF to moderate the intercept and factor loading for an item. Using a conservative alpha level of 0.01 to account for multiple testing, we detected DIF by study for six items, DIF by age for eight items, DIF by gender for three items, and DIF by parental history of alcoholism for one item. In total, 11 of the 17 items displayed DIF (P.J. Curran, J. McGinley, D.J. Bauer, A.M. Hussong, A. Burns, L. Chassin, K. Sher, R.A. Zucker, manuscript in preparation). We can now use this final measurement model to obtain scores for subsequent analysis.

Scoring. By incorporating DIF into the scoring model we can correct for bias in the scores that would result from differential measurement across study and as a function of covariates. For instance, suppose that a harmonized item was constructed from two items measuring similar content but with slightly different item stems. If endorsement rates for these items differed in part due to wording differences in the item stems, and not just due to underlying differences in depression, then we would expect to detect study DIF for this item. Failing to account for the study DIF would lead to artificially elevated depression scores for participants receiving the easier-to-endorse prompt. Generating depression scores from a model that incorporates DIF removes such potential sources of bias, ensuring that the scores are commensurate across studies (or other subpopulations).

There is, however, a question of how much DIF is too much DIF. If all items displayed DIF, this would imply no commonality of measurement between subpopulations and a lack of comparability of scores. DIF among some items is often expected, and just how much DIF is tolerable is a matter of debate (Byrne et al. 1989, Cheung & Rensvold 1998, Reise et al. 1993, Steenkamp & Baumgartner 1998). Strictly speaking, only one invariant (non-DIF) item is required to put the measures on an equivalent scale across subpopulations, but the odds of correctly detecting which items are invariant versus not are reduced when many items display DIF (Yoon & Millsap 2007). The less DIF, the more confidence one can have that the scale of measurement is truly invariant across persons. Although a majority of our depression items displayed some DIF, less than half of the items displayed DIF due to any single covariate, suggesting that we could interpret our measure as being commensurate in scale across subpopulations defined by study, age, gender, and parental history of alcoholism.

Bridging study: a new primary data collection for the express purpose of linking together the measures used in the original set of studies in an IDA

In the final step of our measurement strategy, we generate and evaluate the quality of scores for the full sample, including all repeated measures. This quality check could take a number of forms, including cross-validation with estimates obtained from a second calibration sample and convergent/divergent validity analysis with other criterion variables. For our example, we chose to repeat our measurement-building process with a second calibration sample drawn from our pooled sample and found that scores were highly correlated across the two calibration samples. After conducting such sensitivity analysis, we generated scores for the longitudinal data by refitting the final model to the full, pooled sample of repeated measures, holding constant all the parameter estimates at the values previously obtained from the calibration sample, including all necessary impact and DIF parameters. Once we generated scores for all individuals and all repeated measures, we evaluated their reliability. For our depression measure, the largest standard errors for the scores were obtained at levels below the mean, a common feature of most measures of psychopathology (Reise & Waller 2009).

In sum, completing the sequence of steps illustrated in **Figure 1**, we constructed a measure of depression that is commensurate across our three studies (and across subgroups within studies). The pooled data can now be used in subsequent longitudinal analyses to evaluate, for instance, whether there is support for a negative affect pathway into substance use disorders.

Bridging Studies: What To Do If There Are No Common Items or Measures

The strategy that we described above assumes that there are at least some common items across studies and that a reasonable subset of the common items is invariant (i.e., do not show DIF by study or other covariates). But in some desired IDA applications, there may be no common items (i.e., no items can be harmonized) or too few common items (or common items without DIF) to confidently link measures together. One option to facilitate IDA in such circumstances is to conduct what we refer to as a bridging study. The basic idea is to embark on a new primary data collection for the express purpose of linking together the measures used in the original set of studies. A bridging study would involve recruiting new participants, ideally from a similar population as that sampled in the contributing studies intended for the IDA, and administering items from all of the original studies to these participants. Not all of the original items would have to be administered; a subset of items is sufficient (reducing participant fatigue and eliminating redundancies between items). Pooling the data from the original studies and bridging study, one would then have the opportunity to construct a commensurate measure in the same way described above.

HYPOTHESIS TESTING IN INTEGRATIVE DATA ANALYSIS

As noted previously, IDA is not so much a set of statistical techniques as it is a framework for delineating, controlling for, and exploring sources of between-study heterogeneity in order to create commensurate measures for and test hypotheses in a pooled data analysis. For this reason, the statistical techniques used for hypothesis testing in IDA may draw from many traditions, but they should share the capacity to account for study differences at all plausible points in the modeling sequence. As a result, hypothesis testing in IDA may be challenging due to necessary model complexity, particularly in the context of longitudinal study designs. However, accounting for such between-sample heterogeneity is essential to valid inference testing.

Guiding Principles

Taking into account sources of between-study heterogeneity is most simply done by modeling the effects of study membership for each participant directly in the model. As described previously, fixed-effects IDA treats the study membership of participants as a fixed and known characteristic of each individual observation nested within a given study. Analytic techniques associated with this approach are straightforward; we 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 (as we would gender or ethnicity) and enter these dummy- or effect-coded variables as predictors in our fitted models in a way consistent with Fisher's model-based inferential approach described previously. A key advantage of this strategy is that we can also estimate multiplicative interactions between individual characteristics (e.g., gender, ethnicity) and study membership. This in turn allows us to test the differential impact of individual characteristics on outcomes across the set of studies.

These effect codes eliminate all between-sample sources of variability, and 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 sources simultaneously can be both beneficial and efficient. Importantly, this method also provides a direct test of replication of findings across studies. Significant interactions between study and individual characteristics indicate a failure to replicate an effect across studies. This is a strong test of replication, testing not simply whether an effect is present in all studies but whether the magnitude of the effect is constant across studies.

In this review, we have focused on two key sources of between-study heterogeneity, sampling and measurement, and elsewhere we consider the sources of geographical region, history, and design characteristics (Curran & Hussong 2009). Many other sources of between-study heterogeneity may also be present, and identifying the plausible primary sources is a preliminary step in hypothesis testing. To the extent that study membership is not completely confounded with these sources of between-study heterogeneity (e.g., differences in rates, though not in the presence, of maternal depression across studies), we can isolate the factors that underlie differences in our findings across studies (i.e., factors that explain failures to replicate). Such questions of methodological interest may allow insight into those of theoretical interest, with the ultimate goal of providing some synthesis regarding theoretical associations or even mechanisms. For those sources of between-study heterogeneity that are completely confounded with study, we obviously cannot identify what factors underlie observed study differences, though we can control for them.

Given that final models are often complex in fixed-effects IDA because of the need to control for the main effects of covariates and their interactions with study membership, we recommend using a model-building strategy that includes the practice of model trimming. As is common in single-study analysis, we begin with models that test study differences in the associations between control variables and outcomes to establish a baseline model. We then add to the baseline model theoretical predictors of our outcomes, typically in a sequence of models defined by our substantive theory (e.g., adding unique effect predictors or interactions among theoretical predictors later in the sequence). In each step of the model-building sequence, we test interactions between all variables (including higher-order interaction terms) with study membership.

To address the resulting model complexity, we trim nonsignificant interactions between predictor variables and study membership at each stage of the model-building process. This practice of model trimming is maximally conservative and designed to maintain parsimony, to provide ease of interpretation, and to support greater stability in model estimation. The main effect codes for study membership, however, we do not trim. In our pooled analyses, even our initial models include study membership as a main effect. This may be a particularly important point for testing growth models and using other longitudinal approaches in which the initial model is typically an unconditional model estimated with the goal of identifying the functional form of change over time in a given construct (i.e., in the absence of covariates). However, in the context of IDA, failure to include study membership as a covariate (and predictor of change over time) may occlude the functional form of these trajectories, with differences in the studies contributing observations to the pooled data set over time or age leading to seeming changes in trajectories that are driven not by time trends in the underlying factor but instead by the pooled data design.

Benefits of IDA in Hypothesis Testing

These recommendations for fixed-effects IDA are drawn once again from our experience with the Cross Study (see **Table 1** for sample description). Although we do not provide analytic details regarding the steps in hypothesis testing here given space constraints, we do consider the value added of an IDA approach for hypothesis testing based on applications from this work. For example, IDA allowed us to examine trajectories of internalizing and externalizing symptoms with a larger sample size and over longer periods of development than captured in any one of our contributing studies (Hussong et al. 2007, 2008b). Moreover, IDA allowed us to compare subgroups of children with small representation in the contributing studies to determine, for example, the relative risk for symptomatology among children of antisocial alcoholic parents versus children of depressive alcoholic parents. Given that these subpopulations are rare, important distinctions in the risk profiles of these groups of children of alcoholic parents could not be tested in contributing studies and were only evident in the larger and more highly powered pooled data analysis.

The pooled data set also permitted us to look at the impact of a low base-rate behavior, occurrence of parents' alcohol-related symptoms within any given year of the child's life from ages 2 to 18, on children's functioning (Hussong, et al. 2008a, 2010, 2012). This approach led to an important theoretical distinction involving between-person and within-person effects of parents' alcohol-related symptoms on children's functioning (e.g., Curran & Bauer 2011), differentiating the role of parents' symptoms on identifying which children are at risk for poor functioning (between-person effects) versus identifying the timing of a given child's functional impairment (within-person effects). In this series of studies, we were able to examine these between-person and within-person effects of parents' alcohol-related symptoms on more common indicators of children's functional impairment, such as internalizing symptoms, externalizing symptoms, and alcohol use, as well as on such low base-rate indicators of children's functional impairment as marijuana use and other illicit drug use.

Finally, these applications of IDA collectively demonstrate how we can identify replication of effects across studies as well as the failure for replication. Core findings were largely replicated across studies, despite significant differences in studies due to geography, cohort, assessment, and sampling. However, this was not always the case. In attempting to reconcile differences in our findings across studies, we were able to model potential study differences (e.g., in rates of parental depression and antisociality across studies) that may account for significant interactions between our predictors of interest and study membership in these models. In most cases, these covariates did not account for significant across-study differences in our findings. However, the test of whether these factors accounted for replication failures addressed the plausibility of such comorbid parental disorders as the source of between-study differences, obviating the need for a new study in which this factor is systematically controlled in order to reconcile study differences.

Summary

In sum, our approach to hypothesis testing in the IDA framework uses guidelines for statistical analysis common in single-study analysis, with the goals of controlling for sources of between-study heterogeneity and, when possible, trying to understand how such sources impact study differences. The core principles are to identify and model such sources of between-study heterogeneity at all plausible points in the analysis while using strategies to reduce model complexity and increase interpretability (i.e., model building and trimming strategies). Because these guidelines may be applied using a variety of statistical techniques, IDA offers an incredibly flexible approach to testing hypotheses regarding a wide range of problems of interest in clinical psychology.

RECOMMENDATIONS FOR INTEGRATIVE DATA ANALYSIS IN PRIMARY DATA ANALYSIS

Although our examples of IDA employ a secondary analysis framework, IDA may also be a useful methodological framework for primary data collection; that is, novel data collection with the explicit intention of using IDA. Because data may be used in ways not originally envisioned in the contributing studies, the initial data management task faced by researchers using IDA within a secondary data analysis is often daunting, and adequate resources are needed to support the creation of a reliable integrative database for subsequent analysis. At least a trimmed-down version of this database is often required to determine feasibility of an IDA application, tantamount in effort to conducting a pilot study to establish feasibility for an original data collection. Planning for IDA at the stage of primary data analysis has the potential for a more efficient use of resources and greater likelihood that commensurate measures may be derived, given a preplanned core item set across studies.

Some of the initial challenges in planning for an IDA study with primary data are related to resources (see sidebar Resources for Pooled Data Analysis), but here we focus on some of the scientific challenges. First, a hypothesis-driven approach is likely to avert many problems. We suspect that the types of questions best answered in a primary IDA are those that require larger sample sizes, such as distinguishing among predictors of different kinds of low-base rate behavior (i.e., obsessions versus compulsions), determining whether unusual subgroups of individuals vary in their risk for a clinical outcome (i.e., alcoholic adults with and without comorbid depression),

RESOURCES FOR POOLED DATA ANALYSIS

Pooled data efforts are increasingly supported by funding agencies as an efficient means for conducting research on a larger scale. Whether these efforts are initiated before or after data collection in the contributing studies, an emerging theme is the need for resources to support the infrastructure of the data pooling effort. The creation of a pooled data set is as intensive as the creation of original, single-study data sets. Pooled data sets often require "unpacking" data from the original studies by rescoring, combining, or reanalyzing items in ways not envisioned at the point of data collection. The need to resolve across-study discrepancies places demands on original study teams to provide details about study procedures, requiring an ongoing collaboration between data managers in the original study teams and the IDA team. These discrepancies may become apparent throughout the analysis and interpretation phase, requiring input from original study investigators to identify potential reasons for study differences or replication. For this reason, resources are needed to support effective IDA efforts that target the need for collaboration of both the data management teams and the investigators from the original studies with the IDA study team.

and examining developmental variation in outcomes or predicted associations over time. Each of these types of questions may be difficult to address within any single contributing study, given the relatively lower prevalence of a given behavior, subgroup or age range and more directly tested within IDA given more sample heterogeneity, larger sample sizes, and greater statistical power.

A second practical challenge is creating a common core battery. As described previously, the common items within the selected item set need not be identical across studies but must include a minimum number of items that relate to the underlying construct similarly across studies. This provides flexibility for primary data collection. Batteries for individual studies may include a core set of items for a given construct that are then augmented for individual studies to capture differences in the aims of individual investigations or samples of interest. For example, a core set of 8 items may assess antisocial behavior over time in all studies, but a larger set of 10 to 15 items may be used in three individual studies contributing to this IDA that are tailored to differences in the age ranges of the samples (e.g., ages 2 to 10, 8 to 16, and 14 to 22). This strategy would allow us to create a pooled set of items that include all core items as well as items unique to each study, permitting the collective measure to better capture potential heterotypic continuity (i.e., changes in the indicators of a given construct over time, despite continuity in the construct itself) over development.

A third consideration in planning IDA within primary data collection is sampling. To the extent that individual studies are completely confounded with sampling designs (e.g., all males in one study and all females in another), researchers will be unable to use IDA to test the unique influence of confounded factors in hypothesis testing apart from the influence of study membership. If a priori hypotheses indicate that particular group differences are of interest, the inclusion of important subgroups within each of the contributing studies can provide leverage for distinguishing the effects of theoretically meaningful factors associated with sampling design (e.g., gender, age, ethnicity) and study membership.

A final consideration in planning IDA within primary data collection is in assessing control variables for hypothesis testing. To the extent that variables can be directly modeled and controlled in hypothesis testing, we can differentiate their impact on study outcomes and predicted associations from the influence of study membership. For this reason, some initial planning regarding important ways in which individuals within and across the contributing samples may differ from one another can identify variables that should be included in the common battery across studies.

For example, individual studies may overlap in the range of socioeconomic status (SES) sampled for their participants, but some studies may cluster on the high end whereas others cluster on the low end. To the extent that SES is correlated, though not wholly confounded, with study membership, we can unpack the influence of study membership and SES on outcomes and predicted associations if we have a commensurate measure of SES across studies to include in our models that test our hypotheses.

This is a powerful aspect of IDA. By including these correlated influences, we can attempt to explain why study differences are present within IDA findings. For example, we may find that once we control for SES, study differences on our depression outcome are significantly diminished. Without this information, we are in the same situation as other methods of study integration, such as literature reviews, in which we can only speculate as to the reason for study differences in a pattern of findings. Such speculations often form the basis for a new line of inquiry and require additional data collection. However, if we include these variables in our IDA, we can directly test these speculations at the point of initial study integration, bypassing the need for a new data collection to resolve anticipated study differences.

In sum, we view IDA as a potentially powerful technique for use in both secondary and primary data analysis. In each case, important design features involving measurement, sampling, and hypothesis testing should be considered, though the options for addressing challenges in each of these areas differ for secondary and primary data analysis approaches. Indeed, the greater flexibility offered by primary data analysis to address these challenges is an exciting reason to consider IDA at the point of study planning. Although a number of investigators are currently attempting this approach to IDA, the fruit of their labors and lessons from their collaborations have yet to be realized.

CONCLUSIONS AND FUTURE DIRECTIONS IN INTEGRATIVE DATA ANALYSIS

We consider IDA to a be an important addition to the toolkit for pooled data analysis, offering such advantages as economy (i.e., reuse of extant data), power (i.e., large combined sample sizes), the potential to address new questions not answerable by a single contributing study (e.g., combining longitudinal studies to cover a broader swath of the lifespan), and the opportunity to build a more cumulative science (i.e., examining the similarity of effects across studies and potential reasons for dissimilarities). We also recognize that IDA may not be appropriate for some questions and may be untenable for pooling studies that have no common items or overlap in variables that would support making important distinctions regarding the effect of between-study differences. Two significant sources of between-study heterogeneity to consider in embarking on an IDA are sampling and measurement, each of which may be addressed in many circumstances through the application of traditional analytic techniques to the unique problem of pooling raw data in IDA.

In many ways, measurement is the core challenge to IDA. One must be able to identify or construct commensurate measures of predictors and outcomes for a pooled data analysis to be sensible. We have described a variety of measurement scenarios that might be encountered in IDA, including situations in which identical, highly similar, or truly distinct items are present in the original studies. Ideally, multiple items will be available from each study for a given scale, and a subset of these should be common items (i.e., either identical or harmonized items). Under these conditions, we have outlined and illustrated a series of steps and procedures that we have found useful in constructing commensurate measures. We have also described the use of bridging studies for enabling IDA when no common items are initially available.

We make no pretense that the procedures we have described are exhaustive: There may be other creative ways to link measures together across studies, and additional research on this topic is needed. Nor do we assume these procedures will be universally applicable. For instance, in the models illustrated here we assumed that the items reflect a continuous underlying latent variable and that, conditional on this latent variable, the item responses were locally independent. These assumptions would not be appropriate if, for instance, one wished to score a categorical latent variable (e.g., diagnosis) or if item responses were nested within reporters (e.g., parent, teacher, and self reports). Other psychometric models would be needed in such circumstances. Nevertheless, we believe that the general set of steps illustrated here for constructing commensurate measures across studies should generalize to a wide variety of measurement scenarios.

The same caveats equally apply to the hypothesis testing guidelines we have offered. Because we believe most applications of IDA in clinical psychology will involve limited numbers of studies not representing a random sampling of studies from a defined population of studies, we focused our discussion on fixed-effects IDA. However, random-effects IDA presents a potentially powerful alternative that is largely unexplored as a pooled data analysis technique. Thoughtful consideration of this approach deserves greater attention in future research.

We also believe that the IDA framework may have broader implications for other types of research designs. For example, what constitutes an "independent" study contributing to a pooled data analysis can sometimes be unambiguously determined and other times cannot. In some cases, minor design differences between samples may be present. For example, independent samples may be collected within a multisite or rolling recruiting 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 the country) or across different time periods (e.g., as recruitment rolls across different school years or birth cohorts). These independent samples are then pooled for analysis with some control for site or cohort differences (e.g., Kaplow et al. 2002, Stark et al. 2005). In other cases, many design differences between samples may be present. For example, multiple independent samples may each be collected as part of different independent studies that were conducted at different historical times using different sampling mechanisms, experimental procedures, and psychometric instruments. Thus what constitutes a "separate" sample ultimately resides on a continuum, and some common research designs within clinical psychology that are typically not considered pooled data analysis (e.g., multisite trials) may also benefit from consideration of the issues raised in the IDA framework.

To conclude, we view IDA as a general framework for pooled data analysis that draws on a strong line of previous methodological advancements. We also view the fruitful directions that IDA may best further develop as closely linked to the substantive questions and data sets from our field. Our pressing goal is to evaluate theoretically derived hypotheses to advance, cumulatively, our science. Thus, we view the advancement of substantive understanding and methodological techniques as inextricably linked, both equally necessary to the task of creating a coherent body of knowledge that provides answers to the core questions of our era.

SUMMARY POINTS

1. Given the recent accrual of several matured data sets, increased public access to these data, rapid advances in statistical methods, and a national push toward collaborative research efforts, pooling data for analysis via IDA, or the simultaneous analysis of data obtained from two or more independent studies, energetically responds to each of these motivations and goals.

2. IDA offers several advantages including economy (i.e., reuse of extant data), power (i.e., large combined sample sizes), the potential to address new questions not answerable by a single study (e.g., combining longitudinal studies to cover a broader swath of the lifespan), and the opportunity to build a more cumulative science (i.e., examining the similarity of effects across studies and potential reasons for dissimilarities).
3. In addition to its many advantages, several challenges are also encountered when conducting IDA, including the need to account for sampling heterogeneity across studies, to develop commensurate measures across studies for both predictors and outcomes, and to account for multiple sources of study differences as they may impact hypothesis testing.
4. Depending on the mechanism that resulted in the sample of contributing studies and in the sample of individual observations within each study, IDA approaches to sampling may follow a model-based or designed-based method, with implications for the appropriate use of statistical techniques for subsequent hypothesis testing.
5. Psychometric modeling techniques offer many advantages for developing scale scores that are commensurate in meaning and metric across studies and across distinct within-study subpopulations.
6. Models used to test substantive hypotheses must account for between-study differences that might otherwise be confounded with the effects of interest and that may even be of interest in their own right.
7. Fixed-effects IDA is likely to be most commonly used in current applications in clinical psychology, though future research may provide powerful techniques for integrative analyses using random-effects approaches to IDA.
8. IDA may be a powerful tool in primary data collection, with implications for individual study design to support later data pooling efforts.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The project described was supported by award number R01DA015398 from the National Institute on Drug Abuse. The content is solely the responsibility of the authors and does not represent the official views of the National Institute on Drug Abuse or the National Institutes of Health. This work owes its strengths to the team efforts of the Cross Study research group, including its current research assistants and postdoctoral fellows who commented on this manuscript (Sierra Bainter, Alison Burns, Andrea Howard, Jim McGinely, and Jess Solis), collaborating investigators (Laurie Chassin, Ken Sher, and Bob Zucker), and numerous research assistants and postdoctoral fellows over the last ten years of productivity.

LITERATURE CITED

Achenbach TM, Edelbrock CS. 1981. Behavioral problems and competencies reported by parents of normal and disturbed children aged four through sixteen. *Monogr. Soc. Res. Child Dev.* 46:1–82

Am. Psychol. Assoc. 2002. Ethical principles of psychologists and code of conduct. *Am. Psychol.* 57:1060–73

Bauer DJ, Hussong AM. 2009. Psychometric approaches for developing commensurate measures across independent studies: traditional and new models. *Psychol. Methods* 14:101–25

Bur. Labor Stat., U.S. Dep. Labor. 2002. *National Longitudinal Survey of Youth 1979 Cohort, 1979–2002 (Rounds 1–20)* [data file]. Columbus, OH: Cent. Hum. Resour. Res., Ohio State Univ.

Byrne BM, Shavelson RJ, Muthén B. 1989. Testing for the equivalence of factor covariance and mean structures: the issue of partial measurement invariance. *Psychol. Bull.* 105:456–66

Chassin L, Rogosch F, Barrera M. 1991. Substance use and symptomatology among adolescent children of alcoholics. *J. Abnorm. Psychol.* 100:449–63

Cheung GW, Rensvold RB. 1998. Cross-cultural comparisons using non-invariant measurement items. *Appl. Behav. Sci. Rev.* 6:93–110

Cochran WG. 1977. *Sampling Techniques*. New York: Wiley. 3rd ed.

Cooper HM. 2010. *Research Synthesis and Meta-Analysis: A Step-by-Step Approach*. Thousand Oaks, CA: Sage

Cooper H, Hedges LV, Valentine JC, eds. 2009. *The Handbook of Research Synthesis and Meta-Analysis*. New York: Sage Found.

Curran PJ, Bauer DM. 2011. The disaggregation of within-person and between-person effects in longitudinal models of change. *Annu. Rev. Psychol.* 62:583–619

Curran PJ, Hussong AM. 2009. Integrative data analysis: The simultaneous analysis of multiple data sets. *Psychol. Methods.* 14:81–100

Derogatis LR, Spencer PM. 1982. *The Brief Symptom Inventory (BSI): Administration, and Procedures Manual-I*. Baltimore, MD: Clin. Psychometr. Res.

DeRubeis RJ, Gelfand LA, Tang TZ, Simons AD. 1999. Medications versus cognitive behavior therapy for severely depressed outpatients: mega-analysis of four randomized comparisons. *Am. J. Psychiatry* 156:1007–13

Fisher RA. 1922. On the mathematical foundations of theoretical statistics. *Philos. Trans. R. Soc. Lond. A* 222:309–68

Gibbons RD, Brown CH, Hur K, Davis JM, Mann J. 2012. Suicidal thoughts and behavior with antidepressant treatment: reanalysis of the randomized placebo-controlled studies of fluoxetine and venlafaxine. *Arch. Gen. Psychiatry* 69:580–87

Harris KM, Halpern CT, Entzel P, Tabor J, Bearman PS, Udry JR. 2008. *The National Longitudinal Study of Adolescent Health: Study Design*. Chapel Hill, NC: UNC Carol. Popul. Cent. **http://www.cpc.unc.edu/projects/addhealth/design**

Hirsch Hadorn G, Hoffmann-Riem H, Biber-Klemm S, Grossenbacher-Mansuy W, Joye D, et al. 2008. The emergence of transdisciplinarity as a form of research. In *Handbook of Transdisciplinary Research*, ed. G Hirsh Hadorn, H Hoffman-Riem, S Biber-Klemm, W Grossenbacher-Mansuy, D Joye, C Pohl, U Weismann, E Zemp, pp. 19–39. Dordrecht, Netherlands: Springer

Holland PW. 2007. A framework and history for score linking. In *Linking and Aligning Scores and Scales*, ed. NJ Dorans, M Pommerich, PW Holland, pp. 5–30. New York: Springer

Holland PW, Dorans NJ. 2006. Linking and equating. In *Educational Measurement*, ed. RL Brennan, pp. 187–220. Westport, CT: Am. Counc. Educ./Praeger. 4th ed.

Holland PW, Wainer H. 1993. *Differential Item Functioning*. Hillsdale, NJ: Erlbaum

Hunter DJ, Spiegelman D, Adami H, Beeson L, van den Brandt PA, et al. 1996. Cohort studies of fat intake and the risk of breast cancer—a pooled analysis. *N. Engl. J. Med.* 334:356–61

Hussong AM, Cai L, Curran PJ, Flora DB, Chassin LA, Zucker RA. 2008a. Disaggregating the distal, proximal, and time-varying effects of parent alcoholism on children's internalizing symptoms. *J. Abnorm. Child. Psychol.* 36:335–46

Hussong AM, Flora DB, Curran PJ, Chassin LA, Zucker RA. 2008b. Defining risk heterogeneity for internalizing symptoms among children of alcoholic parents: a prospective cross-study analysis. *Dev. Psychopathol.* 20:165–93

Hussong AM, Huang WJ, Curran PJ, Chassin L, Zucker RA. 2010. Parent alcoholism impacts the severity and timing of children's externalizing symptoms. *J. Abnorm. Child Psychol.* 38:367–80

Hussong AM, Huang W, Serrano D, Curran PJ, Chassin L. 2012. Testing whether and when parent alcoholism uniquely affects various forms of adolescent substance use. *J. Abnorm. Child Psychol.* 40(8):1265–76

Hussong AM, Wirth RJ, Edwards MC, Curran PJ, Chassin LA, Zucker RA. 2007. Externalizing symptoms among children of alcoholic parents: entry points for an antisocial pathway to alcoholism. *J. Abnorm. Psychol.* 116:529–42

Ioannidis JPA, Contopoulos-Ioannidis DG, Lau J. 1999. Recursive cumulative meta-analysis: a diagnostic for the evolution of total randomized evidence from group and individual patient data. *J. Clin. Epidemiol.* 52:281–91

Kaplow JB, Curran PJ, Dodge KA, Conduct Probl. Prev. Res. Group. 2002. Child, parent, and peer predictors of early-onset substance use: a multisite longitudinal study. *J. Abnorm. Child. Psychol.* 30:199–216

Lenhard J. 2006. Models and statistical inference: the controversy between Fisher and Neyman-Pearson. *Br. J. Philos. Sci.* 57:69–91

Lerer B, Segman RH, Fangerau H, Daly AK, Basile VS, et al. 2002. Pharmacogenetics of tardive dyskinesia: combined analysis of 780 patients supports association with dopamine D3 receptor gene Ser9Gly polymorphism. *Neuropsychopharmacology* 27:105–20

Lorenz FO, Simons RL, Conger RD, Elder GR, Johnson C, Chao W. 1997. Married and recently divorced mothers' stressful events and distress: tracing change across time. *J. Marriage Fam.* 59:219–32

Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, et al. 2007. The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.* 39:1181–86

McArdle JJ, Hamagami F, Meredith W, Bradway KP. 2000. Modeling the dynamic hypotheses of Gf-Gc theory using longitudinal life-span data. *Learn. Individ. Differ.* 12:53–79

McArdle JJ, Prescott CA, Hamagami F, Horn JL. 1998. A contemporary method for developmental-genetic analyses of age changes in intellectual abilities. *Dev. Neuropsychol.* 14:69–114

Meade AW, Michels LC, Lautenschlager GL. 2007. Are internet and paper-and-pencil personality tests truly comparable? *Organ. Res. Methods* 10:322–45

Meehl PE. 1978. Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *J. Consult. Clin. Psychol.* 46:806–34

Meredith W. 1993. Measurement invariance, factor analysis and factorial invariance. *Psychometrika* 58:525–43

Millsap RE, Meredith W. 2007. Factorial invariance: historical perspectives and new problems. In *Factor Analysis at 100: Historical Developments and Future Directions*, ed. R Cudeck, RC MacCallum, pp. 131–52. Mahwah, NJ: Erlbaum

Mischel W. 2008. Connecting clinical practice to scientific progress. *Psychol. Sci. Public Interest* 9:i–ii

Mischel W. 2009. From personality and assessment 1968 to personality science, 2009. *J. Res. Personal.* 43:282–90

Natl. Database Autism Res. 2011. *National Database for Autism Research*. Bethesda, MD: Natl. Inst. Mental Health, Natl. Inst. Health. **http://ndar.nih.gov**

Natl. Inst. Drug Abuse. 2010. *Accelerating the Pace of Drug Abuse Research Using Existing Epidemiology, Prevention, and Treatment Research Data (R01)*. Bethesda, MD: Natl. Inst. Health. **http://grants.nih.gov/grants/guide/pa-files/PAR-10-018.html**

Natl. Inst. Drug Abuse. 2012. *Clinical Trials Network (CTN)*. Bethesda, MD: Natl. Inst. Drug Abuse. **http://www.drugabuse.gov/about-nida/organization/cctn/ctn**

Natl. Inst. Child Health Human Dev. 2010. *Systems-Oriented Pediatric Obesity Research and Training (SPORT) Center of Excellence* (U54) (RFA-HD-10–001). Bethesda, MD: Natl. Inst. Health. **http://grants.nih.gov/grants/guide/rfa-files/RFA-HD-10-001.html**

Natl. Inst. Health. 2003. *Final NIH Statement on Sharing Research Data*. Bethesda, MD: Natl. Inst. Health. **http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html**

Natl. Sci. Found. 2011. *Dissemination and Sharing of Research Results*. Arlington, VA: Natl. Sci. Found. **http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/aag_6.jsp#VID4**

Neyman J. 1934. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *J. R. Stat. Soc.* 97:558–606

Northwest. Univ. Feinberg Sch. Med. 2012. *Collaborative Data Synthesis for Adolescent Depression Trials*. Chicago, IL: Dep. Prev. Med., Northwest. Univ. Feinberg Sch. Med. **http://www.preventivemedicine.northwestern.edu/research/collab-data-adolescent-depression.html?t=vD1**

Pfeffermann D. 1993. The role of sampling weights when modeling survey data. *Int. Stat. Rev.* 61:317–37

Raudenbush SW, Bryk AS. 2002. *Hierarchical Linear Models*. Thousand Oaks, CA: Sage. 2nd ed.

Reise SP, Waller N. 2009. Item response theory and clinical measurement. *Annu. Rev. Clin. Psychol.* 5:27–48

Reise SP, Widaman KF, Pugh RH. 1993. Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. *Psychol. Bull.* 114:552–66

Richman WL, Kiesler S, Weisband S, Drasgow F. 1999. A meta-analytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires, and interviews. *J. Appl. Psychol.* 84:754–75

Rivers DC, Meade AW, Fuller WL. 2009. Examining question and context effects in organization survey data using item response theory. *Organ. Res. Methods* 12:529–33

Schaeffer NC. 1988. An application of item response theory to the measurement of depression. In *Sociological Methodology*, vol. 18, ed. CC Clogg, pp. 271–307. Washington, DC: Am. Sociol. Assoc.

Schmidt FL. 1996. Statistical significance testing and cumulative knowledge in psychology: implications for training of researchers. *Psychol. Methods* 1:115–29

Shadish WR, Cook TD, Campbell DT. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton-Mifflin

Sher KJ, Walitzer KS, Wood PK, Brent EE. 1991. Characteristics of children of alcoholics: putative risk factors, substance use and abuse, and psychopathology. *J. Abnorm. Psychol.* 100:427–48

Stark LJ, Janicke DM, McGrath AM, Mackner LM, Hommel KA, Lovell D. 2005. Prevention of osteoporosis: a randomized clinical trial to increase calcium intake in children with juvenile rheumatoid arthritis. *J. Pediatr. Psychol.* 30:377–86

Steenkamp JEM, Baumgartner H. 1998. Assessing measurement invariance in cross national consumer research. *J. Consum. Res.* 25:78–90

Steinberg L, Thissen D. 2006. Using effect sizes for research reporting: examples using item response theory to analyze differential item functioning. *Psychol. Methods* 11:402–15

Sterba SK. 2009. Alternative model-based and design-based frameworks for inference from samples to populations: from polarization to integration. *Multivar. Behav. Res.* 44:711–40

Takane Y, de Leeuw J. 1987. On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika* 52:393–408

Tourangeau R, Rips LJ, Rasinski K. 2000. *The Psychology of Survey Response*. New York: Cambridge Univ. Press

Vandenberg RJ, Lance CE. 2000. A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. *Organ. Res. Methods* 3:4–69

van den Brandt PA, Spiegelman D, Yaun S, Adami H, Beeson L, et al. 2000. Pooled analysis of prospective cohort studies on height, weight, and breast cancer risk. *Am. J. Epidemiol.* 152:514–27

Widaman KF, Reise SP. 1997. Exploring the measurement invariance of psychological instruments: applications in the substance use domain. In *The Science of Prevention: Methodological Advances from Alcohol and Substance Abuse Research*, ed. KJ Bryant, M Windle, SG West, pp. 281–324. Washington, DC: Am. Psychol. Assoc.

Yoon M, Millsap RE. 2007. Detecting violations of factorial invariance using data-based specification searches: a Monte Carlo study. *Struct. Equ. Modeling* 14:435–63

Zucker RA, Fitzgerald H, Refior S, Puttler L, Pallas D, Ellis D. 2000. The clinical and social ecology of childhood for children of alcoholics: description of a study and implications for a differentiated social policy. In *Children of Addiction: Research, Health, and Policy Issues*, ed. H Fitzgerald, B Lester, B Zuckerman, pp. 109–41. New York: Routledge Falmer

Annual Review of
Clinical Psychology

Volume 9, 2013

Contents

Indexes

Errata

An online log of corrections to *Annual Review of Clinical Psychology* articles may be found at http://clinpsy.annualreviews.org