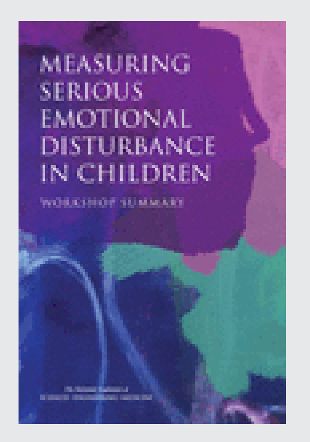


Measuring Serious Emotional Disturbance in Children: Workshop Summary

DETAILS

88 pages | 6 x 9 | PAPERBACK
ISBN 978-0-309-38818-4 | DOI: 10.17226/21865

AUTHORS

Krisztina Marton, Rapporteur; Committee on National Statistics; Board on Behavioral, Cognitive, and Sensory Sciences; Division of Behavioral and Social Sciences and Education; Board on Health Sciences Policy; Institute of Medicine; National Academies of Sciences, Engineering, and Medicine

MEASURING SERIOUS EMOTIONAL DISTURBANCE IN CHILDREN

WORKSHOP SUMMARY

Krisztina Marton, *Rapporteur*

Committee on National Statistics and
Board on Behavioral, Cognitive, and Sensory Sciences,
Division of Behavioral and Social Sciences and Education

and

Board on Health Sciences Policy, Institute of Medicine

The National Academies of
SCIENCES · ENGINEERING · MEDICINE

THE NATIONAL ACADEMIES PRESS
Washington, DC
www.nap.edu

THE NATIONAL ACADEMIES PRESS 500 Fifth Street, NW Washington, DC 20001

This activity was supported by Contract No. HHSP233201400020B/HHSP23337002 between the National Academy of Sciences and the U.S. Department of Health and Human Services. Support of the work of the Committee on National Statistics is provided by a consortium of federal agencies through a grant from the National Science Foundation (No. SES-1024012). Any opinions, findings, conclusions, or recommendations expressed in this publication do not necessarily reflect the views of the organization or agency that provided support for the project.

International Standard Book Number-13: 978-0-309-38818-9
International Standard Book Number-10: 0-309-38818-X

Additional copies of this report are available for sale from the National Academies Press, 500 Fifth Street, NW, Keck 360, Washington, DC 20001; (800) 624-6242 or (202) 334-3313; http://www.nap.edu.

Printed in the United States of America

Suggested citation: National Academies of Sciences, Engineering, and Medicine. (2016). *Measuring Serious Emotional Disturbance in Children: Workshop Summary.* K. Marton, Rapporteur. Committee on National Statistics and Board on Behavioral, Cognitive, and Sensory Sciences, Division of Behavioral and Social Sciences and Education. Board on Health Sciences Policy, Institute of Medicine. Washington, DC: The National Academies Press.

STEERING COMMITTEE FOR THE WORKSHOP ON INTEGRATING NEW MEASURES OF SERIOUS EMOTIONAL DISTURBANCE IN CHILDREN INTO THE SUBSTANCE ABUSE AND MENTAL HEALTH ADMINISTRATION'S DATA COLLECTION PROGRAMS

KATHLEEN RIES MERIKANGAS *(Chair)*, National Institute of Mental Health, U.S. Department of Health and Human Services
GLORISA CANINO, School of Medicine, University of Puerto Rico
MICHAEL DAVERN, NORC at the University of Chicago
GRAHAM KALTON, Westat, Rockville, MD

KRISZTINA MARTON, *Study Director*
JEANNE RIVARD, *Senior Program Officer*
MICHAEL SIRI, *Program Associate*

COMMITTEE ON NATIONAL STATISTICS

BOARD ON BEHAVIORAL, COGNITIVE, AND SENSORY SCIENCE

BOARD ON HEALTH SCIENCES POLICY

Acknowledgment of Reviewers

This workshop summary has been reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise, in accordance with procedures approved by the Report Review Committee of the National Academies of Sciences, Engineering, and Medicine. The purpose of this independent review is to provide candid and critical comments that will assist the institution in making its published report as sound as possible and to ensure that the report meets institutional standards for objectivity, evidence, and responsiveness to the charge. The review comments and draft manuscript remain confidential to protect the integrity of the process.

We thank the following individuals for their review of this report: William Copeland, Center for Developmental Epidemiology, Department of Psychiatry and Behavioral Sciences, Duke University School of Medicine; Susanna Visser, Division of Human Development and Disability, National Center on Birth Defects and Developmental Disabilities, Centers for Disease Control and Prevention; Jerry West, Mathematica Policy Research, Washington, DC; and Alan M. Zaslavsky, Department of Health Care Policy, Harvard Medical School.

Although the reviewers listed above provided many constructive comments and suggestions, they were not asked to endorse the content of the report nor did they see the final draft of the report before its release. The review of this report was overseen by Brian Baird, president, 4Pir2 Communication, Edmonds, Washington. Appointed by the National

Academies of Sciences, Engineering, and Medicine, he was responsible for making certain that an independent examination of this report was carried out in accordance with institutional procedures and that all review comments were carefully considered. Responsibility for the final content of this report rests entirely with the rapporteur and the institution.

Contents

APPENDIXES

1

Introduction

BACKGROUND

This report summarizes the presentations and discussions at the Workshop on Integrating New Measures of Serious Emotional Disturbance in Children into the Substance Abuse and Mental Health Services Administration's (SAMHSA) Data Collection Programs, held in Washington, D.C., in June 2015. The workshop was organized as part of a study sponsored by SAMHSA and the Office of the Assistant Secretary for Planning and Evaluation (ASPE) of the U.S. Department of Health and Human Services (DHHS) to assist SAMHSA in its responsibilities to expand the collection of behavioral health data in several areas. The workshop was structured to bring together experts in the measurement of serious emotional disturbance in children and in health survey methods to facilitate discussion of measures and mechanisms most promising for expanding SAMHSA's data collections in this area.

The overall effort is being overseen by the Standing Committee on Integrating New Behavioral Health Measures into the Substance Abuse and Mental Health Services Administration's Data Collection Programs.[1] In addition to new measures of serious emotional disturbance in children, SAMHSA and ASPE are interested in expanding data collection on specific mental illness diagnoses with functional impairment, on trauma, and

[1] For a description of the overall study, see http://sites.nationalacademies.org/DBASSE/CNSTAT/Behavioral_Health_Measures_Committee/index.htm [October 2015].

on recovery from substance use or mental disorder. Workshops on all four topics are being planned as part of the overall effort.

WORKSHOP FOCUS

In his introductory remarks about SAMHSA's goals for the workshop, Neil Russell (SAMHSA) explained that the agency has a legislative mandate to provide national and state-level estimates of serious emotional disturbance in children. The primary motivation for collecting the data is the block grant for community mental health services, which was established by the Alcohol, Drug Abuse, and Mental Health Administration Reorganization Act of 1992. The block grant is administered by SAMHSA and provides funds to support state-level services for children with serious emotional disturbance and for adults with serious mental illness.

Russell said that for the purposes of SAMHSA's work, the definition of child serious emotional disturbance is the definition published in a 1993 *Federal Register* notice (58 FR 29425, May 20), which included the following elements:

- children from birth up to age 18
- who currently or any time during the past year
- have had a diagnosable mental, behavioral, or emotional disorder of sufficient duration to meet diagnostic criteria specified within DSM-III-R[2]
- which has resulted in functional impairment which substantially interferes with or limits the child's role or functioning in family, school, or community activities.

Two important aspects of the definition are mental disorders and impairment. The intent of the definition is to ensure the availability of targeted grant fund allocations for those children with the most severe functional impairments. The definition also intentionally excludes substance use disorders because of the availability of other resources for substance abuse facilities and substance abuse treatment organizations.

Prior to commissioning this study from the National Academies of Sciences, Engineering, and Medicine, SAMHSA convened two panels of its own to address some fundamental questions associated with the mea-

[2] The DSM-III-R was the 1987 (current at the time) edition of the *Diagnostic and Statistical Manual of Mental Disorders*, a standard classification of mental disorders published by the American Psychiatric Association.

surement of child serious emotional disturbance.[3] One question was how the new edition of the *Diagnostic and Statistical Manual of Mental Disorders* (American Psychiatric Association, 2013) or DSM-5, affects the definition, and in particular, which disorders should be included in the definition. The second question was whether suitable instruments are available to measure these disorders and impairment within the context of the definition. A summary of the two panel meetings was made available to the steering committee for this workshop.

The workshop and the standing committee are expected to assist SAMHSA with exploring these options further, both in terms of deciding what instruments to use and the data collection approach. Some of the measurement considerations include whether the instruments available are suitable for use with the data collection modes that may be feasible and for the population of interest. Another consideration is whether the instrument has been validated for use with Spanish-speaking populations.

In terms of data collection approach, Russell described four options that SAMHSA has been considering. One option would be to use the existing National Survey on Drug Use and Health (NSDUH). The main challenge with this approach is that the average NSDUH interview is 60 minutes, and SAMHSA would prefer not to increase the amount of time, so adding new questions would require identifying existing questions that could be dropped.

A second option would be to reinstate the Mental Health Surveillance Study (MHSS), which was conducted as a follow-up study to the NSDUH between 2008 and 2012. The MHSS was designed to be used in a statistical model that would generate estimates of serious mental illness. Although there were some challenges associated with this approach (described in the afternoon by Heather Ringeisen and Jeremy Aldworth, RTI International), a similar design could be considered.

If the scope of the planned data collection project is expected to be larger than can be accommodated as part of the NSDUH, a third option would be to design a new stand-alone data collection. A fourth option would be to identify an existing data source that already includes the data being sought. For all of these options, both direct and model-based estimation methods could be considered.

[3] Substance Abuse and Mental Health Services Administration, Center for Behavioral Health Statistics and Quality. (In press). *Serious Emotional Disturbance (SED) Expert Panel Meeting: Operationalizing the SED Definition for the Production of National and State Prevalence Estimates, September 8, 2014, Gaithersburg, MD* [meeting notes]. Rockville, MD: Author; Substance Abuse and Mental Health Services Administration, Center for Behavioral Health Statistics and Quality. (In press). *Serious Emotional Disturbance (SED) Expert Panel Meeting: Instrumentation and Measurement Issues When Estimating National and State Prevalence of Childhood SED, November 12, 2014, Gaithersburg, MD* [meeting notes]. Rockville, MD: Author.

BOX 1-1
Statement of Task

A steering committee will organize a public workshop that will feature invited presentations and discussions on options for expanding SAMHSA's behavioral health data collections to include measures of serious emotional disturbance in children. The discussion will explore new measures and efficient mechanisms for collecting the data. Possibilities include adding new measures to existing surveys, initiating new data collections, or implementing model-based estimation procedures that take advantage of existing data sources, in the event that primary data collection methods are cost-prohibitive or not necessary. Survey and questionnaire design tradeoffs, as well as the potential impact of any changes to existing surveys, will also be discussed. An individually authored summary of the presentations and discussions at the workshop will be prepared by a designated rapporteur in accordance with institutional guidelines.

Russell clarified that there is no specific budget amount allocated for producing data on child serious emotional disturbance, but the workshop participants should consider cost implications.

WORKSHOP CHARGE

The specific statement of task for the workshop (shown in Box 1-1) was developed on the basis of the charge for the overall project: to expand behavioral health data collections on several topics. The main goals of the workshop were to discuss options for collecting data and producing estimates on serious emotional disturbance in children, including consideration of the available measures and possible data collection mechanisms.

ORGANIZATION OF THE REPORT

This summary describes the workshop presentations and the discussions that followed each topic: see the workshop agenda in Appendix A. Chapter 2 provides an overview of existing measures and data on child serious emotional disturbance, including discussions of measuring mental disorders and measuring functional impairment, the two components of the definition of serious emotional disturbance in children. The chapter includes an overview of approaches used in previous studies to estimate the prevalence of mental disorders in the United States; a discussion of the role of measuring impairment as part of efforts to measure serious emotional disturbance in children, along with measures that may be particularly suitable to meet SAMHSA's goals; an overview of ongoing federal surveillance systems that collect data on child mental health; and an update on trends in mental health impairment among children, based on data from a national survey.

Chapter 3 presents international perspectives on collecting data on child serious emotional disturbance. Researchers in other countries, including Canada, Australia, and the United Kingdom, have been grappling with some of the same challenges as researchers in the United States, and their experiences contributed additional points of view to the discussion.

Chapter 4 includes presentations on study design and estimation options and challenges to inform a discussion about implementation options. It covers the design of the National Survey of Children's Health and lessons learned from that survey; design and estimation considerations in multiphase studies; pilot studies conducted on behalf of SAMHSA to produce prevalence rates of child serious emotional disturbance and of adult serious mental illness, using model-based estimation; and a project that used small-area estimation to produce prevalence rates of serious emotional disturbance in children from school-based samples.

Chapter 5 summarizes the discussions throughout the day, which occurred primarily during two formal discussion sessions, one after the morning presentations and the other one after the afternoon presentations.

This report has been prepared by the workshop rapporteur as a factual summary of what occurred at the workshop. The steering committee's role was limited to planning and convening the workshop. The views contained in the report are those of individual workshop participants and do not necessarily represent the views of all workshop participants, the steering committee, or the National Academies of Sciences, Engineering, and Medicine.

2

Existing Measures and Data

NATIONAL DATA ON THE PREVALANCE OF MENTAL DISORDERS

Kathleen Merikangas (National Institute of Mental Health, NIMH) began her presentation with a reference to *Mental Health: A Report of the Surgeon General*, a document that discusses definitions of mental disorders (albeit definitions that have changed over time), the perceived causal mechanisms of mental disorders, and the social and other contextual factors that influence mental disorders and treatment strategies. When it was first published in 1999,[1] the report brought national attention to the public health relevance of mental disorders and also led to several initiatives by NIMH that focused on collecting data on the prevalence and magnitude of mental disorders, as well as on associated impairments in both adults and children.

One of the NIMH initiatives was to include the Strengths and Difficulties Questionnaire (SDQ) in the National Health Interview Survey (NHIS). The NHIS is a nationally representative annual household survey. The SDQ (described in further detail below) was included in the NHIS between 2001 and 2003. The sample included children aged 4-17, and the interviews were completed by their caretakers, in either English or Spanish. The children included in the NHIS sample are the youngest age group

[1] U.S. Department of Health and Human Services. (1999). *Mental Health: A Report of the Surgeon General*. Washington, DC: Author. Available: http://profiles.nlm.nih.gov/ps/access/NNBBHS.pdf [October 2015].

for which national data are available, using this measure. The SDQ measures severe difficulties in emotional or behavioral functioning during the 6 months preceding the interview. The resulting estimates from the NHIS surveys were around 5 percent, with higher rates for boys than girls.

Another NIMH initiative described by Merikangas was to add selected modules from a structured interview, the Diagnostic Interview Schedule for Children (DISC), to the National Health and Nutrition Examination Survey (NHANES). The NHANES is another nationally representative annual survey. The DISC was included in the NHANES between 2001 and 2004. It was administered to children 8 and older. In the case of children under 16, a parent interview was also conducted. Although the DISC can measure a large number of disorders, only a subset of the disorders were selected for inclusion in the NHANES: generalized anxiety disorder, panic disorder, eating disorder, elimination disorder, major depression, dysthymic disorder, attention deficit hyperactivity disorder (ADHD), and conduct disorder. Merikangas pointed out that one challenge for researchers considering the use of the NHANES data is that in order to protect respondent confidentiality, these data can only be accessed on-site, through the Research Data Center of the National Center for Health Statistics.

NIMH was also involved in the development of the National Comorbidity Survey Replication Adolescent Supplement (NCS-A), a supplement to the National Comorbidity Survey Replication (NCS-R). The NCS-A was a nationally representative survey conducted between 2000 and 2004 that included adolescents aged 13-18. The dual sampling frame covered both households and schools, and the interviews with adolescents were supplemented with a self-administered questionnaire completed by parents. The survey included a modified version of the World Mental Health Organization Composite International Diagnostic Interview 3.0 (CIDI), and it collected data on the full spectrum of disorders. The sample was limited to adolescents 13 and older because there was a lack of agreement about the measures that would be appropriate for use with younger children. In addition to producing prevalence rates, the survey also provided information about correlates of mental disorders.

Merikangas said that it would be valuable to conduct a follow-up study with the adolescents (or a subsample of the adolescents) included in the NCS-A. Such a study could shed light on whether the presence of mental disorder in adolescence and the thresholds used are good predictors of impairment in young adult life.

The data from the surveys described illustrate several challenges encountered when deciding how to measure child serious emotional dis-

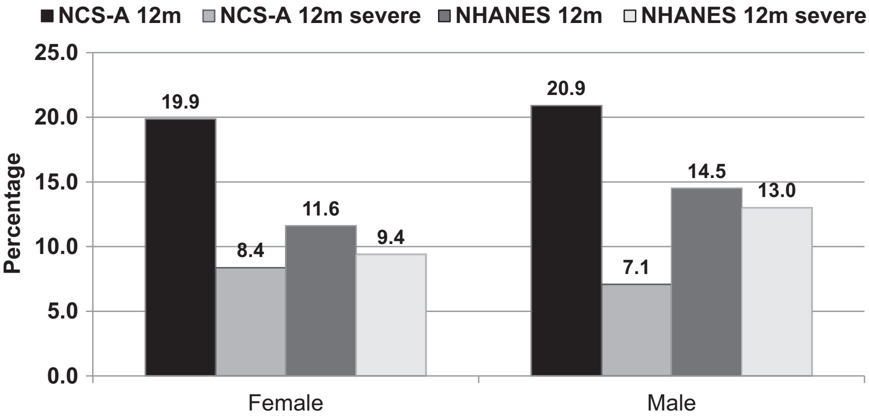


FIGURE 2-1 Prevalence of 12-month mental disorder by sex and severity, NCS-A and NHANES.
NOTES: Any mental disorder includes anxiety (generalized anxiety disorder and panic disorder), mood disorder (major depressive disorder and dysthymia), attention deficit hyperactivity disorder (ADHD), conduct disorder, and eating disorder. NCS-A, National Comorbidity Survey Replication Adolescent Supplement; NHANES, National Health and Nutrition Examination Survey.
SOURCE: Workshop presentation by Kathleen Merikangas, June 2015.

turbance in large-scale surveys.[2] Figure 2-1 shows the NCS-A and the NHANES data on the rate of mental disorders experienced in the 12 months preceding the survey. The gender differences in the data are well known. The more interesting differences are in the rates obtained from the two surveys, particularly the large differences that are a result of applying a severity threshold. When a severity criterion is applied, there is a larger drop in the case of the NCS-A (which measured a broader range of disorders) than in the NHANES. These differences illustrate the importance of obtaining input from experts with clinical experience and carefully considering what definitions to use and how the thresholds map onto what is intended by the diagnostic criteria.

Merikangas noted that another issue that deserves careful consideration is the role of informants. There has been a lot of discussion about

[2] Merikangas, K.R., He, J.P., Brody, D., Fisher, P.W., Bourdon, K., and Koretz, D.S. (2012). Prevalence and treatment of mental disorders among U.S. children in the 2001-2004 NHANES. *Pediatrics, 125*(1), 75-81; Merikangas, K.R., He, J.P., Burstein. M., Swanson, S.A., Avenevoli, S., Cui, L., Benjet, C., Georgiades, K., and Swendsen, J. (2010). Lifetime prevalence of mental disorders in U.S. adolescents: Results from the National Comorbidity Survey Replication-Adolescent Supplement (NCS-A). *Journal of the American Academy of Child and Adolescent Psychiatry, 49*(10), 980-989.

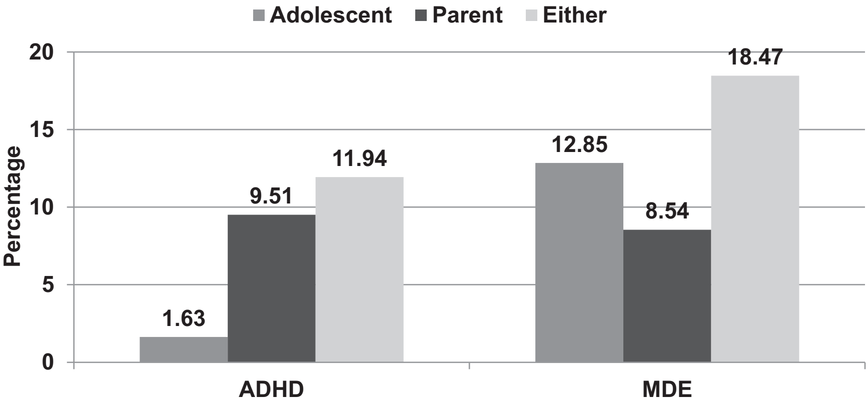


FIGURE 2-2 Lifetime prevalence of ADHD and MDE by different informants, NCS-A
NOTE: The "either" bars reflect adolescent and parent reports combined at symptom level using an "or" rule.
SOURCE: Workshop presentation by Kathleen Merikangas, June 2015.

the relative advantages of children's reports and parents' reports over the years: it seems that who is most knowledgeable depends on the specific question and the disorder. In addition, some research indicates that for children under the age of 12, teacher reports may also be necessary in order to increase reliability. Figure 2-2 shows differences in the NCS-A data in reporting between adolescents and parents for ADHD and major depressive episode (MDE). In the case of ADHD, the parents' estimates are more than five times higher than those of the adolescents. In the case of MDE, a different pattern emerges: the adolescents were more likely to report criteria for major depression than the parents. In the case of disorders such as depression and anxiety, it appears that the older the child, the less likely parents are to be able to provide adequate information, Merikangas said. Teachers are also likely to underestimate mood disorders or anxiety disorders, particularly social anxiety.

Another consideration, Merikangas noted, is the large proportion of children with more than one disorder. In the NCS-A study, approximately 40 percent of the adolescents with a disorder had multiple disorders. In other words, reports of the percentage of children with various disorders tend to count the same children under several disorders. In addition, data from the recent Neurodevelopmental Genomics Study show that there is an association between the severity of mental disorders and physical disorders. This association suggests that in order to understand impairment, it is important to learn as much as possible about a child, including information about the full spectrum of physical and mental conditions, and the home environment.

Merikangas said that the studies that have been conducted over the years have yielded a wealth of information about the magnitude and correlates of mental disorders in children, in addition to simple prevalence estimates. However, there are several limitations. The cross-sectional design of the studies means that the data collection is limited to one point in time. In some cases, such as the NHANES, the number of disorders captured is very limited. Measures of psychosis are missing from all the nationally representative surveys conducted in the United States, although they are usually included in similar surveys in other countries. Developmental disorders are also missing and would be important to include if there is an interest in impairment. Finally, there are inconsistencies in the reporting depending on the informant, and the lack of data from teachers is also a limitation. Merikangas pointed out that although the nationally representative surveys sometimes lack depth of information, there are several regional studies that have collected rich, comprehensive data, including in the Northwest, the Smoky Mountain region, New York state, and Puerto Rico.

Thinking about SAMHSA's mandate to produce prevalence estimates of children with serious emotional disturbance, Merikangas said that there are several challenges that are particularly important to consider. One challenge is that the instruments available to measure child serious emotional disturbance are tied to the DSM system, which has limited predictive or biologic validity. In addition, Merikangas noted, there is increased dissatisfaction with the DSM-5, in particular, for certain applications. The DSM is also not typically used outside of the United States. Other countries use the International Classification of Diseases (ICD), currently the ICD-10, and the two systems are not comparable. The field as a whole needs to begin a discussion about why the United States has its own version of diagnostic criteria and how those can be mapped to the ICD-10.

Another challenge, Merikangas noted, is related to defining serious emotional disturbance for children younger than 6. The average 6- or 7-year-old with ADHD is likely to have had other difficulties by that age, so finding a way to expand the data collection to include children under 6 will be important. The NCS-A is an example of a project that involved expanding an existing survey to include adolescents, but something similar could not easily be done for children younger than 6 because the issues characteristic of that age group are very different. In addition, as mentioned, comorbidity, both among mental disorders and with physical conditions, is pervasive, and it will be important to find a way to integrate such information into defining impairment in children.

THE ROLE OF MEASURING FUNCTIONAL IMPAIRMENT

Glorisa Canino (University of Puerto Rico) provided an in-depth overview of the concept of functional impairment and its role as part of measuring serious emotional disturbance in children. She proposed defining impairment on the basis of the International Classification of Functioning, Disability and Health for Children and Youth (ICF-CY), which describes disabilities as negative functional outcomes resulting from health conditions, involving significant deviation from or loss of "normal" or "expected" function. Negative functional outcomes can occur at the individual level, for example, if a child has difficulties with some activities. They can also occur in a social context, for example, as difficulties with participation in contexts involving family, school, peers, or the community at large.

The ICF-CY definition implies that impairment should be viewed as separate from the disorder or health condition. However, Canino noted, in mental health, this separation is not possible because of the absence of biological markers or clinically useful measurements of severity: without such markers or measurements, it is not possible to separate normal and pathological symptom expressions. This is particularly true in the case of disorders in children, because it is very difficult to assess what is normal and what is pathological. As a result, a diagnostic criterion requiring distress and disability has been used to establish disorder thresholds in the DSM. Thus, the definition of the health condition is dependent on the presence of functional impairment.

However, as discussed during Russell's presentation (see Chapter 1), the legislation governing the provision of mental health services separates diagnosis and impairment, and the presence of impairment is a necessary condition for the purposes of allocating funds. Insurance companies also use similar definitions.

Indeed, impairment is the best predictor of need for services, Canino said. Declines in functioning and unexpected deviations in behavior are the most common reasons that parents first seek mental health services for children. Impairment in functioning is more likely to lead parents to seek treatment for their children than a psychiatric disorder. Furthermore, most epidemiologic studies find that the perception of disability is more significant than a diagnosis in predicting the use of mental health services.

Impairment is also the pivotal information basis for decisions about interventions, Canino said. While diagnosis is important for prognosis, impairment tends to be the determining factor in planning and developing interventions. Furthermore, improvement in functioning is the main outcome used for determining the effectiveness of an intervention.

Another reason that impairment is important is its prevalence. As discussed by Merikangas, approximately 5 percent of children had a definite or severe impact score based on the SDQ that was part of the 2001-2003 NHIS. Even higher prevalence rates were found in the 2010-2012 Medical Expenditure Panel Survey data, discussed by Benjamin Druss below, which estimated that 11 percent of children have severe mental health impairment, as measured by the Columbia Impairment Score (CIS). As Merikangas pointed out earlier, different definitions and instruments can lead to different prevalence estimates, which underscores the need for operationalizing substantial impairment.

Canino said that the criteria for establishing the symptom according to DSM IV or DSM-5 are well established. However, what constitutes substantial impairment in social, work, or other areas of functioning is variable and somewhat arbitrary, depending on the clinician, measurement instrument, and population-based statistical scores.

Because of the lack of conceptual clarity, some researchers determine impairment severity empirically, by establishing population norms (for example, one or two standard deviations from the mean) or by determining cutoffs based on cost or research purpose. Canino offered an illustration: for the 12-item adult World Health Organization Disability Assessment Schedule (WHO-DAS), RTI International uses a scoring system from 0 to 58. A WHO-DAS score of 17 will correspond to the 90th percentile of the population, and around 10 percent of the population will have a score higher than 17. A WHO-DAS score of 31 would translate into 5 percent of the population. Using the WHO-DAS, it is not straightforward to determine whether the 10 percent or the 5 percent figure represents the population with substantial impairment, and this is a decision that has to be made by SAMHSA.

There are other impairment scales, such as the Brief Impairment Scale (BIS) and the CIS. These instruments do not provide a severity score: rather, they simply determine whether impairment is present or not, based on the specificity and sensitivity of the instrument, using either a population-based or clinical sample.

Canino summarized the characteristics of an impairment measure that are best suited in her view to measure child serious emotional disturbance as follows:

- has scoring that is suitable to determine severity or significant impairment;
- able to assess functioning in a variety of contexts, such as school, family, friends (multidimensional);
- suitable for use with a wide range of ages;
- available for both parent and child reports;
- has good psychometric properties for the U.S. population in both English and Spanish; and

- does not require that the person administering the interview have prior knowledge of the child to make an assessment.

None of the existing measures meets all of these requirements, Canino said. The child WHO-DAS has potential because it is based on the ICF, which she considers to be the best definition of impairment or disability in children. However, use of the ICF for youth is not widely known or accepted by clinicians in the United States. There has been progress in the conceptual definition of the construct since the publication of the child ICF, but the problems for measurement persist because of the imprecise operationalization and validity of the construct for children.

Canino said she recognizes that there are both advantages and disadvantages to the ICF-CY. One of the advantages is that it is based on an international classification. Another advantage is that training manuals are available and offer criteria for clinicians for assessing different types of disability and the contextual factors (e.g., school, family, community) that can contribute to the presentation, occurrence, and outcome of disabilities. Finally, it can be useful for the treatment and prevention of both mental and physical impairments.

One of the disadvantages of the ICF-CY is that its applicability to children with serious emotional disturbance is limited, due to the operationalization challenges discussed. The WHO-DAS for adults was used successfully with adolescents in the NCS-A, but further research would be needed on the psychometric properties of the instrument, based on that study. A DSM-5 workgroup also developed a child version of the WHO-DAS, and it has been used in DSM-5 clinical trials, but psychometric data have not been published for either the English or Spanish version. Canino also noted that it is not clear who owns the copyright for the instrument, the American Psychological Association or the World Health Organization.

When thinking about other options, it is important to remember that a panel of experts convened by SAMHSA suggested that an instrument that is independent of psychiatric disorders and symptomatology should be used. This approach would exclude the impairment scales based on psychiatric symptoms, such as the DISC and the Child and Adolescent Psychiatric Assessment (CAPA).

Canino briefly covered several global and multidimensional measures that could also be considered. Table 2-1 summarizes the global measures, and Table 2-2 presents the multidimensional measures. In addition to the child WHO-DAS, there is one scale that does come close to meeting the SAMHSA criteria, and that is the Child and Adolescent Functional Assessment Scale (CAFAS). The CAFAS can be used for a wide range of ages, and it has very good psychometric properties in both English and Spanish. The disadvantage of the CAFAS is that it is long, and it does require

TABLE 2-1 Global Impairment Measures

Impairment Measure[a]	Severity Score, Cutoff	Time to Administer (Minutes)	Age Range Covered	Respondent: Parent, Child, or Teacher	Advantages	Disadvantages
CIS	Cutoff No severity scores	3 (13 items)	7-17	P, C	No training needed for administration; good psychometric properties in English and Spanish; short	No severity score, not applicable for children younger than 7; three items confounded with symptoms
C-GAS	Yes	5	4-16	P, C	Good psychometric properties in English and Spanish; short; severity cutoffs	Not applicable for children younger than 4; dependent on prior knowledge of the child and what interviewer/parent thinks is normal functioning; confounded with symptoms
SDQ-Impact	Yes[b]	2 (5 items)	2-17	P, C, T	Good psychometrics and predictive validity in several languages, including Spanish; used in large national survey by the Centers for Disease Control and Prevention; wide age range; short	Not tested in children younger than 5; of five items, only three refer to impact; difficulty disentangling impact from symptoms; all items refer to difficulties with emotions, concentration, behavior, or getting along with others; measures outcome, not impairment

[a] CIS, Columbia Impairment Scale; C-GAS, Child Global Assessment Scale; SDQ, Strengths and Difficulties Questionnaire.

[b] Scores are 0-10: 0, no problem; 1, minor problem; 2-10, definite and severe problem.

SOURCE: Workshop presentation by Glorisa Canino, June 2015.

TABLE 2-2 Multidimensional Impairment Measures

Impairment Measure[a]	Severity Score, Cutoff	Time to Administer (Minutes)	Age Range Covered	Respondent: Parent, Child, or Teacher	Advantages	Disadvantages
CAFAS	Yes	30 (97 items)	5-19	P, C	Good psychometrics, Spanish version, severity cutoffs, wide age range, PCFAS version	Dependent on prior knowledge of the child's symptoms and level of functioning for 10-minute administration, intertwines symptoms with impairment
PECFAS	Yes	30	0-6	P	Same as above	Same as above
BIS	Global cutoff, no severity[b]	10 (23 items)	4-17	P	English/Spanish versions, good psychometrics, brief	No tested child version, not applicable for children younger than 4
BERS	No	15 (52 items)	0-5, 6-18	P, C, T	Wide age range; focuses on strengths; used mostly for placing children and treatment goals	Requires knowledge of child; has not been tested in populations other than whites; no Spanish version

[a] CAFAS, Child and Adolescent Functional Assessment Scale; PECFAS, Preschool and Early Childhood Functional Assessment Scale; BIS, Brief Impairment Scale; BERS, Behavioral and Emotional Rating Scale.

[b] Global cutoff refers to results of receiver operating characteristic (ROC) curve analysis in which a cutoff is determined using external criteria (i.e., mental health service use, another instrument with known psychometric properties) to establish whether the participant is impaired or not. It does not determine severity of impairment.

SOURCE: Workshop presentation by Glorisa Canino, June 2015.

prior knowledge of the child, which is usually obtained by administering another questionnaire.

Canino concluded her talk by noting several areas where further research would be useful to narrow down the options. They include additional analysis of the psychometric properties of the child WHO-DAS or other data, such as the CIS from the Medical Expenditure Panel Survey (see below), as well as work on shortening the CAFAS. She added that forming a workgroup to assist SAMHSA with adapting or developing an impairment measure that meets the agency's goals would also be useful.

ONGOING FEDERAL CHILD MENTAL HEALTH SURVEILLANCE SYSTEMS

Susanna Visser (Centers for Disease Control and Prevention, CDC) discussed the data available from ongoing federal child mental health surveillance systems. Her presentation was based on a report that was the culmination of a multiagency project to produce an inventory of ongoing and recurring federal efforts to monitor mental disorders in children.[3] The goals of such efforts are to (1) document the impact of children's mental health and mental disorders; (2) document the mental health needs of children; (3) build effective programs and services for children and families; (4) inform research on factors that increase risk and promote prevention; and (5) inform policy and resource allocation.

The report compiled information on the prevalence of specific mental disorders and other indicators of mental health among children, which is a first step toward better understanding of the disorders and their impact. With this initiative, the CDC wanted to contribute its expertise in the area of surveillance to an issue that is important to many agencies.

The surveillance systems reviewed are as follows:

- Autism and Developmental Disabilities Monitoring Network
- National Health and Nutrition Examination Survey
- National Health Interview Survey
- National Survey of Children's Health
- National Survey on Drug Use and Health
- National Violent Death Reporting System
- National Vital Statistics System
- National Youth Risk Behavior Survey
- School-Associated Violent Death Surveillance Study
- National Comorbidity Survey Replication Adolescent Supplement

[3] Centers for Disease Control and Prevention. (2013). Mental health surveillance among children—United States, 2005-2011. *Morbidity and Mortality Weekly Report, 62* (Supplement 2).

Visser noted that the list does not include the NCS-R because it is not a recurring survey, and therefore it cannot be described as a surveillance system, but it does provide a wealth of information and it would be useful to repeat.

The report summarizes key information about the data collections, including who is included (e.g., age, sample size, any oversamples); the geographic coverage (e.g., national estimates, state estimates, limited number of sites); the topics covered (e.g., general health, conditions, diagnoses, symptoms); who provided the information (e.g., parents, self-reports); when the data were collected and what time periods are covered by the questions (e.g., current diagnosis, past month, past year, ever); and data collection mode (e.g., telephone, in-person, administrative records).

The report concluded that recurring information that meets surveillance criteria is available for seven disorders: ADHD, disruptive behavioral disorders, autism spectrum disorders, anxiety, depression, substance use disorders, and Tourette syndrome. In addition, there are also indicators available for suicide and mentally unhealthy days. Looking at all disorders across the different surveillance systems, Visser noted that it is clear that all demographic groups are affected by mental disorders. Another report finding is that the proportion of children with a mental disorder increases with age, with the exception of autism spectrum disorders. Among children, boys are more likely than girls to have most of the conditions.

The data from these various sources can also be used to focus in on a demographic group of interest. For example, for adolescents between the ages of 12 and 17, the 2010-2011 National Survey of Drug Use and Health (NSDUH) provides data on major depressive episode in the past year (8.1 percent), illicit drug use disorder in the past year (4.7 percent), alcohol use disorder in the past year (4.2 percent), and cigarette dependence in the past month (2.8 percent). The 2010 National Vital Statistics System shows that suicide was the second leading cause of death among adolescents aged 12-17 and that the suicide rate is 4.5/100,000 children. The 2005-2010 NHANES data show that 8.3 percent of 12- to 17-year-olds reported 14 or more mentally unhealthy days in the past month.

An advantage of having national surveys that include these types of indicators is that they provide population estimates that help in understanding the relative sizes of the groups with various disorders, which helps with resource planning: see Figure 2-3. However, Visser noted, there are also some data gaps and challenges. One limitation is that many surveillance systems exclude undiagnosed cases, instead relying heavily on parents' receiving a diagnosis from a health care provider and then reporting that accurately. Another limitation is that data are not readily available on some conditions, especially specific anxiety disorders and

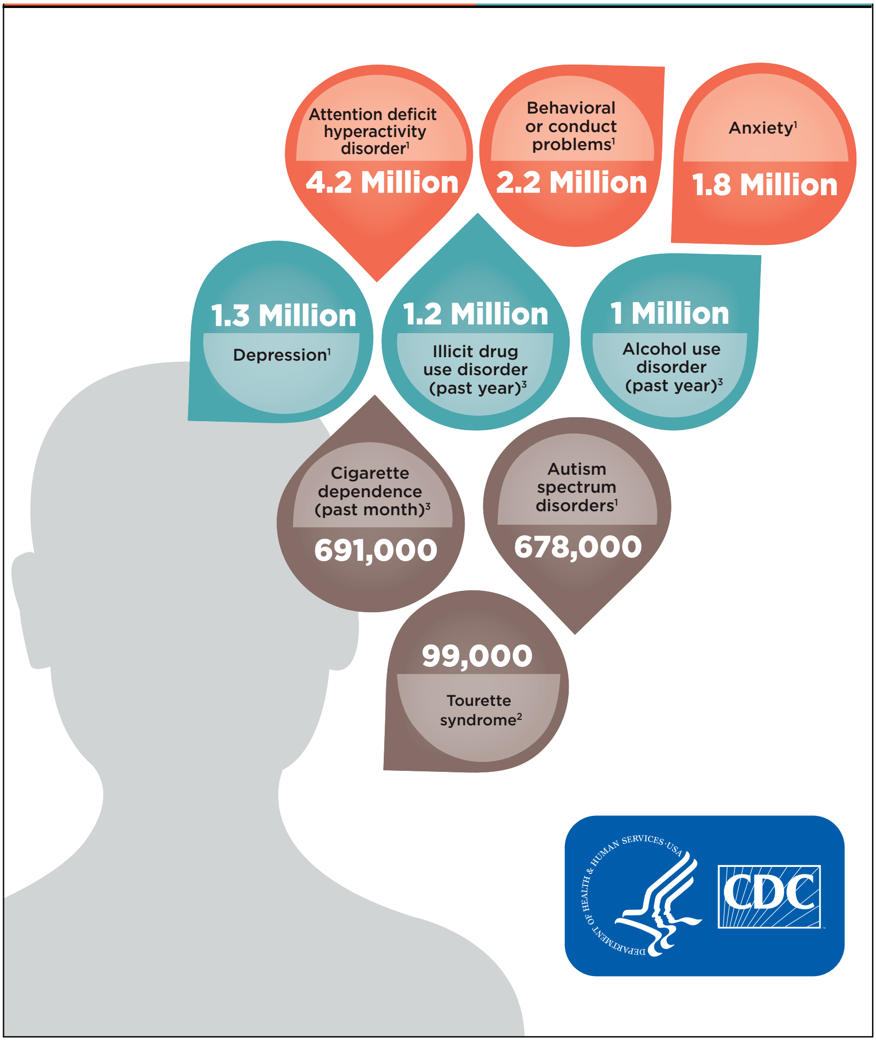


FIGURE 2-3 Estimates of U.S. children with mental disorders.
NOTE: The data cover children aged 3-17 except as follows: for Tourette syndrome, they cover children aged 6-17, and for illicit drug use disorder, alcohol use disorder, and cigarette dependence, they cover children aged 12-17.
SOURCE: Workshop presentation by Susanna Visser, June 2015, based on Centers for Disease Control and Prevention, available: http://www.cdc.gov/media/dpk/2013/docs/Child_menatal_health/Child_menatal_health_infographic.pdf [October 2015].

bipolar disorder. Due to the lack of consistency in the definitions among the surveys, it is not possible to calculate an overall estimate of the prevalence of all mental disorders in children. As others mentioned, the criteria for mental disorders can be subjective, especially in the case of ratings of impairment. Parents, children, and teachers frequently disagree on impairment. In addition, there is very little validation of case ascertainment methods for surveillance.

Visser said that sometimes it is necessary to rely on more than one type of data source to estimate prevalence rates. For example, for research on ADHD, the national survey data, community-based studies, and administrative records are complementary sources, each with its own strengths and limitations.

The national surveys, which typically include parents, are a very efficient method for collecting these data. With this approach, very large sample sizes are possible, and they produce estimates that are generalizable and sometimes available even at the state level. The repeated surveys also allow for monitoring trends over time. The main limitations of these surveys are declining response rates, coverage bias, recall bias, and reporting bias.

Data from community-based studies that involve either direct assessments of children or some other type of active surveillance case assessment tend to have more depth and breadth than data from the national surveys. These studies also allow for hypothesis generation and testing. However, Visser noted, coverage bias and lack of generalizability are among the main limitations of these data sources. Response rates are also declining in community-based studies, with families becoming increasingly more reluctant to engage in public health research in general.

In terms of administrative records, Medicaid data are available in every state and for a large number of cases. Using these data also allows for monitoring trends over time. However, the data are submitted for the purposes of payments and contain limited clinical information. Coverage bias is also a limitation for Medicaid records because only a subset of the population is included.

Visser used ADHD as an example of how different data sources can be combined to better understand a disorder. Table 2-3 shows that from 2007-2008 to 2011 there has been an increase in the prevalence of children ever diagnosed with ADHD or current ADHD in the NHIS, but the highest estimate comes from the 2007 National Survey of Children's Health (NSCH). The two surveys ask about diagnosed ADHD in the same way, but the NSCH is a telephone interview, so the differences seem to indicate a mode effect. The rate of current ADHD in the 2007 NSCH was 6.8 percent.

The large sample sizes of the national surveys, such as the NSCH, allow for detailed analyses of ADHD by age group, and the annual administration allows for the close monitoring of cohort effects. The NSCH also produces state-level estimates, and indeed there is great geographic variability in parent-reports of diagnosed ADHD among the states, Visser noted. Nevada has the lowest rate (4.2 percent), and Arkansas and Kentucky have the highest rates (14.6 and 14.8 percent, respectively). Nevada is the only state that requires preauthorization for psychotropic medications for children of all ages, and information about these policy differences among the states can be combined with the data for analytic purposes.

Visser mentioned the concerns associated with declining telephone survey response rates in the NSCH, which were also discussed by Kogan (see below). Despite those concerns, however, the survey estimates and data obtained from administrative records can be remarkably similar. For example, there is high convergent validity between the NSCH estimates of ADHD for insured children between the ages of 5 and 11 in the state of California (4.2 percent) and physician-diagnosed ADHD data from Kaiser Permanente Southern California (4.5 percent).

An analysis of survey data and administrative data in Georgia found that the percentage of children with administrative claims for ADHD, especially medicated ADHD, was comparable to state-based survey estimates of ADHD. In addition, use of administrative data, such as the Georgia Medicaid files, also allowed for further in-depth analyses. For example, the researchers found that the rate of ADHD medication increases with age, but the rate of psychological services does not increase, despite the fact that best practices call for a combination of both for children 6 and older.

The CDC currently operates four community-based sentinel sites, where it has been conducting direct assessments of children since 2003, using the DISC and other instruments. These sites also allowed CDC to examine how the changes in the DSM might affect the estimates of ADHD prevalence. Based on data from the Project to Learn about ADHD, the researchers found that approximately 30 percent of the children will meet the criterion when six or more symptoms are required.[4] The estimate does not change when the criterion of onset before age 12 is added. When the criterion of two or more moderate impairments or one severe impairment is added, the estimate drops to 22 percent. The biggest drop, to 11 percent, occurs with the addition of teacher reports, specifically, at

[4] McKeown, R.E., Holbrook, J.R., Danielson, M.L., Cuffe, S.P., Wolraich, M.L., and Visser, S.N. (2015). The impact of case definition on attention deficit hyperactivity disorder prevalence estimates in community-based samples of school-aged children. *Journal of the American Academy of Child & Adolescent Psychiatry, 54*(1), 53-61.

TABLE 2-3 Prevalence of Children Aged 3–17 Years Who Ever Received a Diagnosis of ADHD or Who Currently Have ADHD, by Sociodemographic Characteristics and Year

	Ever Received a Diagnosis of ADHD (Parent Report)			
	NHIS 2007–2008 (N = 14,970)		NHIS 2009–2010 (N = 18,411)	
Characteristic	%	(95% CI)	%	(95% CI)
Sex				
Male	10.6	(9.7–11.4)	11.5	(10.7–12.3)
Female	4.6	(4.0–5.3)	5.4	(4.8–6.0)
Age Group				
3–5	2.1	(1.5–2.8)	2.5	(1.9–3.4)
6–11	7.4	(6.6–8.3)	8.2	(7.4–9.0)
12–17	10.8	(9.8–11.8)	11.9	(11.0–12.9)
Race/Ethnicity[a]				
Hispanic	4.1	(3.4–4.9)	4.6	(3.9–5.4)
Black, non-Hispanic	8.1	(6.9–9.5)	10.3	(9.1–11.7)
White, non-Hispanic	9.1	(8.3–10.0)	10.0	(9.2–10.8)
Multirace, non-Hispanic	10.2	(7.3–14.0)	11.5	(8.6–15.2)
Other, non-Hispanic	3.1	(1.8–5.3)	2.0	(1.4–2.9)
Highest Education in Household[b]				
Less than high school	6.5	(4.9–8.5)	7.9	(6.5–9.5)
High school graduate	8.9	(7.8–10.1)	10.5	(9.3–11.7)
More than high school	7.4	(6.8–8.2)	8.0	(7.4–8.7)
Health Insurance				
Yes	7.9	(7.3–8.5)	8.7	(8.2–9.3)
No	5.4	(4.2–6.9)	5.9	(4.6–7.5)
Region				
Northeast	6.8	(5.8–8.0)	8.6	(7.4–9.9)
Midwest	8.8	(7.5–10.2)	9.4	(8.3–10.7)
South	8.9	(8.0–10.0)	10.1	(9.2–11.0)
West	5.1	(4.3–6.0)	5.2	(4.4–6.1)
Poverty-Income Ratio[c]				
≤100% FPL	8.9	(7.5–10.4)	11.4	(10.1–12.7)
>100% to ≤200% FPL	8.6	(7.3–10.0)	9.2	(8.0–10.6)
>200% FPL	6.9	(6.3–7.6)	7.1	(6.5–7.8)
Total	7.6	(7.1–8.2)	8.5	(8.0–9.0)

NOTES: ADHD, attention deficit hyperactivity disorder; CI, confidence interval; FPL, federal poverty level; NHIS, National Health Interview Survey; NSCH, National Survey of Children's Health.

[a] Other, non-Hispanic, includes American Indian/Alaska Native, Hawaiian/other Pacific Islander, and Asian. Persons categorized as multirace selected more than one race.

[b] The highest education in the household is based on the highest education of adults in the sample child's family for NHIS and on the education of parents or respondents (adults) for NSCH.

				Current ADHD (Parent Report)	
NHIS 2011 (N = 10,554)		NSCH 2007 (N = 78,042)		NSCH 2007 (N = 78,042)	
%	(95% CI)	%	(95% CI)	%	(95% CI)
12.0	(11.0–13.1)	12.3	(11.6–13.1)	9.6	(8.9–10.4)
4.7	(4.0–5.6)	5.3	(4.8–5.9)	3.8	(3.4–4.2)
2.1	(1.4–3.0)	1.5	(1.1–1.9)	1.1	(0.9–1.5)
8.4	(7.4–9.5)	9.1	(8.3–9.9)	7.5	(6.8–8.2)
11.9	(10.7–13.2)	12.4	(11.5–13.3)	8.8	(8.1–9.6)
5.6	(4.6–6.8)	5.4	(4.4–6.6)	4.0	(3.1–5.0)
8.8	(7.3–10.5)	10.0	(8.8–11.4)	7.7	(6.6–9.0)
10.1	(9.1–11.2)	10.0	(9.4–10.6)	7.6	(7.0–8.1)
5.5	(3.4–8.6)	13.0	(10.4–16.2)	10.2	(7.9–13.0)
4.1	(2.4–6.9)	4.0	(3.1–5.1)	3.0	(2.2–3.9)
7.7	(6.1–9.7)	8.5	(7.1–10.2)	6.6	(5.3–8.1)
7.5	(6.3–9.0)	11.8	(10.5–13.2)	8.7	(7.7–9.8)
8.8	(8.0–9.7)	8.1	(7.6–8.7)	6.3	(5.8–6.8)
8.7	(8.0–9.5)	9.2	(8.7–9.7)	7.1	(6.7–7.6)
4.7	(3.2–6.7)	6.3	(4.9–8.2)	3.5	(2.7–4.5)
7.5	(6.0–9.4)	8.8	(7.8–9.9)	7.0	(6.1–8.0)
8.7	(7.3–10.3)	9.3	(8.6–10.1)	7.1	(6.5–7.8)
10.3	(9.2–11.6)	10.3	(9.5–11.1)	7.7	(7.1–8.5)
6.0	(4.9–7.3)	6.6	(5.4–7.9)	4.8	(3.9–6.0)
10.5	(8.9–12.4)	11.1	(9.9–12.4)	8.7	(7.7–10.0)
6.7	(5.6–8.1)	9.7	(8.6–11.0)	7.2	(6.3–8.3)
8.3	(7.5–9.3)	8.0	(7.5–8.6)	6.1	(5.6–6.6)
8.4	(7.8–9.1)	8.9	(8.4–9.4)	6.8	(6.4–7.2)

[c] FPL is based on family income and family size and composition using federal poverty thresholds that are updated annually by the U.S. Census Bureau using the change in the average annual consumer price index for all urban consumers. For details, see: http://www.census.gov/hhes/www/poverty/index.html [October 2015].

SOURCE: Workshop presentation by Susanna Visser, June 2015, based on Centers for Disease Control and Prevention (2013). Mental health surveillance among children—United States, 2005–2011. *Morbidity and Mortality Weekly Report, 62* (Supplement 2).

least four teacher-reported symptoms. This difference underscores the importance of who is providing the information, especially in the case of direct assessments.

Visser concluded by saying that future comprehensive surveillance systems could benefit from consistent case definitions and validation of methodologies. To address SAMHSA's goals of producing estimates of serious emotional disturbance in children, combining mixed and multiple methods will be necessary, because no one system has everything that is needed, and combining methods can improve the data. Leveraging partnerships among agencies will also be important in developing better estimates.

IMPAIRMENT DATA FROM THE MEDICAL EXPENDITURE PANEL SURVEY

Benjamin Druss (Emory University) began his presentation by synthesizing some of the themes from previous presentations. Some of the questions that the workshop and the overall study are addressing are long-standing, fundamental issues related to the classification of diseases. Prior to 1952, there was little agreement regarding definitions of various disorders and impairments, and a common language for how to think about these issues only emerged gradually with the different releases of the DSM.

One issue that has been discussed is the need to think of these disorders and impairments as a continuum, rather than as categorical diagnoses, Druss said. There is a tendency to think that a person either has a disorder or does not have it. This kind of characterization is also linked to how mental health services are reimbursed. Furthermore, especially for children, it is important to understand their functioning in different contexts, such as school, home, and with friends. This issue underscores the importance of thinking about the role of the survey informant (the parent, teacher, or child providing the information). Another challenge that has been discussed is separating impairment from symptoms, but impairment is how illness is generally defined, and ultimately it is impairment that matters. Yet this topic, too, deserves further discussion, Druss noted.

From the perspective of research on services, one limitation of the available administrative data is that they only include people who are receiving services. The denominator of interest would be the people who have the illness. Druss described some of his research, which has focused on attempting to match rising rates of service use with the rates of impairment that are measured by a large-scale survey. Between 1995 and 2010, there was a doubling in mental health visits and a fivefold increase in the use of psychostimulants in the United States among children under 18.

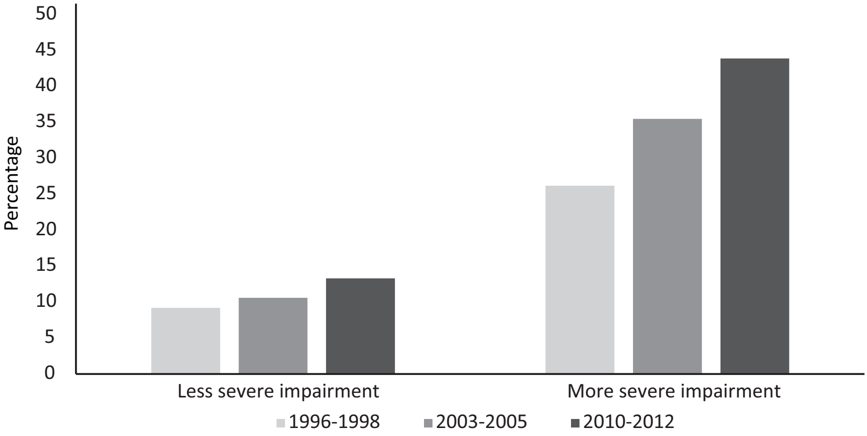


FIGURE 2-4 Percentage of children using any mental health services by impairment. NOTE: More severe odds ratio 2.2 (1.76-2.75); Less severe odds ratio 1.51 (1.35-1.72). SOURCE: Workshop presentation by Benjamin Druss, June 2015. Data from the Medical Expenditure Panel Survey.

It is important to understand whether these increases reflect a need that is now being better met (that is, whether there are more children with a demonstrated need who are being treated) or if there is an increase in the treatment of children with less serious symptoms. This question is also relevant because most treatments have only been tested among those with diagnosed conditions.

Druss described the work he and his colleagues, Mark Olfson (Columbia University) and Steven Marcus (University of Pennsylvania), conducted using data from the Medical Expenditure Panel Survey (MEPS).[5] The MEPS is primarily a service use survey, rather than an epidemiological survey, and it has included the Columbia Impairment Scale since the 1990s. As discussed earlier, the CIS is a measure of impairment across different domains. Although similar to other measures, the CIS also makes it difficult to tease some of the concepts apart.

Comparing data from 1996-1998, 2003-2005, and 2010-2012, the study found that, contrary to what one might expect, there was a modest decrease in impairment among children and adolescents from 12.5 percent in 1996-1998 to 10.5 percent in 2010-2012. At the same time, there was an increase in service use, and the MEPS data show that the rate of increase was more rapid among children with more severe impairments: see Figure 2-4. If impairment is a proxy for need for services, then this is

[5] Olfson, M., Druss, B.G. Druss, and Marcus, S.C. (2015). Trends in mental health care among children and adolescents. *New England Journal of Medicine, 372*, 2029-2038.

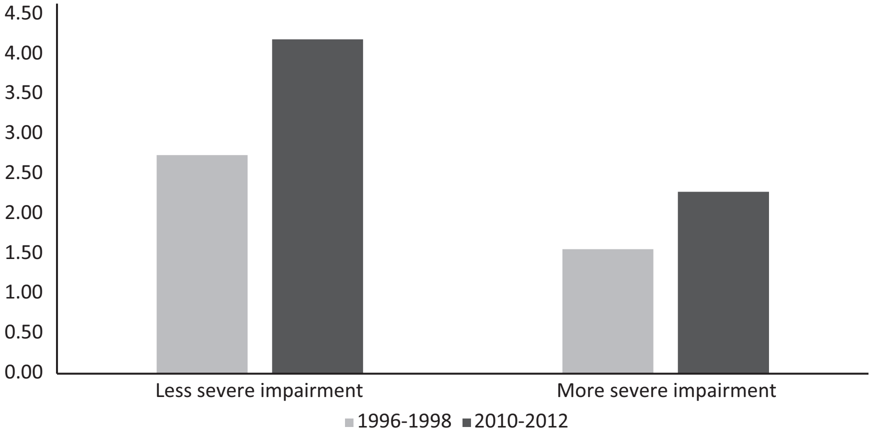


FIGURE 2-5 Number of children using any mental health services by impairment.
NOTES: 1.45 million increase in users with less severe impairment; 0.72 million increase in users with more severe impairment.
SOURCE: Workshop presentation by Benjamin Druss**,** June 2015. Data from the Medical Expenditure Panel Survey.

what one would want to see. However, the absolute numbers tell a different story. Between 1996-1998 and 2010-2012, there was a 1.45 million increase in the number of users of mental health services with less severe impairment and a 0.72 million increase in the number of users with more severe impairment: see Figure 2-5. In other words, while relative growth in treatment is greater among children with more severe impairment, absolute growth has been concentrated among those with less severe impairment.

Druss concluded by saying that epidemiological data can provide information about the need for services (the denominator), while service use data can provide the numerator for this type of research. The implications for SAMHSA, and possibly the NSDUH, are that including overall impairment data can be useful and may serve as a proxy for treatment need. Measuring disease burden alongside service use data can provide stronger policy relevance than either used alone.

3

Measurement Challenges for Population Surveys

LESSONS FROM CANADA

Peter Szatmari (University of Toronto) focused his presentation on current issues in psychiatric epidemiology of serious emotional disturbances in children, primarily in the areas of behavioral and developmental disorders and the measurement of impairment among children between the ages of 2 and 6. He also discussed his experiences with the Ontario Child Health Study (OCHS) to illustrate some of the challenges.

Szatmari said that there is widespread agreement in the literature on the prevalence estimates of many disorders, with or without impairment, and this is especially true in the area of disruptive behavior disorders—oppositional defiant disorder, attention deficit hyperactivity disorder (ADHD), conduct disorder—across developmental stages. A just-published meta-analysis found that there is relatively good agreement on the prevalence of mental disorders at around 13 percent.[1] The analysis also found that the prevalence rates would be approximately 17 percent higher if the criterion of impairment were not included. The availability of this research raises the question of whether another epidemiological study of child and adolescent psychiatric disorders is necessary. Szatmari pointed out that there are many unanswered questions that are relevant to mental health and addiction policies. Although SAMHSA's definition

[1] Polanczyk, G.V., Salum, G.A., Sugaya, L.S., Caye, A., and Rohde, L.A. (2015). Annual research review: A meta-analysis of the worldwide prevalence of mental disorders in children and adolescents. *Journal of Child Psychology and Psychiatry, 56*(3), 345-365.

of child serious emotional disturbance is limited to substantial impairment, Szatmari argued that, from a policy perspective, focusing on any disorder and any impairment, regardless of severity, is more meaningful. One would not want to limit treatment to only severe cancer or severe diabetes, and any mental disorder that causes impairment similarly requires treatment.

To better understand child serious emotional disturbance, more longitudinal data are needed on onset, chronicity, and course of disorder and impairment. For example, there are virtually no data available on the extent to which disorder and impairment follow similar or different pathways and how they are linked. Further work is also needed to link the available data to administrative databases. Linking epidemiological data to administrative databases, as discussed by Benjamin Druss (see Chapter 2), is likely to be very useful in informing mental health policies and guiding resource allocation for assessment and treatment. Finally, research is needed to enable a more nuanced understanding of modifiable risk factors for serious emotional disturbance in children.

Canada has had several initiatives to measure child serious emotional disturbance and impairment. One such initiative was the National Longitudinal Study of Children and Youth, which is now inactive, but included multiple waves, and eight cycles. The survey was focused on symptoms and impairment, rather than disorders. Because the study was funded by the government, confidentiality restrictions made it very difficult for researchers to use the data. A current initiative is the Early Development Instrument (EDI), which is turning out to be very useful. The instrument, which is completed by teachers for each child in the last year of kindergarten, measures “school readiness” on the basis of motor, language, social, and emotional-behavioral milestones. The data have been collected in most Canadian provinces, and they are now beginning to be used in many other countries around the world, with the support of the World Health Organization and UNICEF.

Szatmari discussed his experience with the 1987 OCHS and the current, 2014–2015, sequel to that study (OCHS-S). The study, which has a nested design, includes 25,000 children, in 13,500 households, in 180 neighborhoods across Canada. Up to three children are sampled in each family, which provides useful information about the familiality of disorders and a better understanding of what families face.

One of the main goals of the OCHS-S is to document the prevalence of mental disorders among children between the ages of 4 and 16. The team also wants to document child mental health needs and assess health system response by linking data on children with serious emotional disturbance to administrative databases, such as those for education and social services. The researchers also want to better understand the influence of poverty and income inequality on child serious mental disturbance and identify modifiable environmental influences, within a nested design.

In order to measure disorder, in 1983 the OCHS used a tool called the Survey Diagnostic Instrument. This was updated for the OCHS-S to make it congruent with the DSM IV. The disorders measured are ADHD, conduct disorder, oppositional defiant disorder, major depressive disorder, specific phobia, social phobia, separation anxiety disorder, generalized anxiety disorder, and substance use disorders. A randomly selected subset of the sampled children received a structured psychiatric interview, the Mini International Neuropsychiatric Interview for Children and Adolescents (MINI KID). The diagnostic instrument was completed by at least two informants for each child: parents, for all the children (4-16); the children themselves for those aged 12-16; and teachers for those aged 4-12.

Impairment is measured in multiple ways in the OCHS. The survey includes an independent measure of impairment, similar to the Strengths and Difficulties Questionnaire (SDQ) (see Chapter 2). The OCHS also includes the Brief Impairment Scale (BIS), and the MINI KID, which has impairment tied to each disorder. Impairment is defined across four domains: family, peers, school, and community.

The OCHS-S data will shed light on how the prevalence of mental disorders has changed since 1983, when it was 18 percent. Of great interest is the overlap between chronic medical conditions and mental disorders. In 1983, among children with chronic medical conditions and a functional limitation from the medical condition, rates of disorder were around 40 percent. This rate was 14 percent among children with no medical condition. Comparing families living above and below the poverty line, in 1983 children living in poverty experienced more mental health difficulties, and only 17 percent of those children with a disorder were using mental health services. Since then, more funding has been devoted to child health initiatives, and there are more programs available, so the new data are expected to provide important information for research and policy.

Szatmari said that the OCHS involves a large team across Canada, and the current study went through a very thorough process to collect stakeholders' input on the design. In addition to a comprehensive review of the literature, the team spent a lot of time talking to government policy makers about the mental health issues of concern to them and what questions they wanted answered. The team also conducted focus groups with clinicians and service providers, getting their input on the same issues. In hindsight, not enough time was spent on patient engagement, talking with children and families. The patient engagement literature is very impressive and reveals the important role patients can play in designing a study, formulating research questions, talking about measurement tools, and even in the interpretation of the results. Going forward, this area deserves a lot of attention in mental health and epidemiological surveys, Szatmari said.

Szatmari also discussed his views on measuring impairment. He noted that it is important to distinguish impairment, distress, impact, and the burden associated with a disorder, as well as how these concepts differ from the quality of life. All of these concepts have slightly different nuances. Robert Goodman has argued that the measurement structure of impairment is stable, and it is very much defined by context: family, school, community.

It is also important to consider the relationship between symptoms and impairment, which was also discussed by Glorisa Canino (see Chapter 2). Cross-sectional studies indicate that impairment and symptoms overlap but are not identical. Unfortunately, not enough data exist from longitudinal studies. Some treatment studies provide experimental information on impairment and symptoms, but most focus on symptoms, and significant improvement in symptoms does not necessarily signify significant improvement in impairment. The research suggests that they are two separate constructs.

As Canino also described, Szatmari noted, there are two theoretical frameworks for measuring impairment. According to the DSM framework, symptoms cause impairment. In contrast, the framework of the International Classification of Functioning for Children and Youth (ICF-CY) is much more complex, and there is a bi-directional relationship between impairment, symptoms, and environment.[2] Szatmari argued that it is impossible to make causal inferences when it comes to symptoms and impairment, and it is best to be agnostic on this topic. Therefore, it is important to have an independent global measure of impairment. While the Children's Global Assessment Scale (CGAS) had a crucial role in the development of the field, it is an example of a measure that is not independent because it is "contaminated" by symptoms. Although this is a very important topic, the field overall has invested substantially more resources into the measurement of symptoms than the measurement of impairment.

Szatmari said that a very complicated measurement issue is that of comorbidity. If there are two disorders present, it is difficult to determine whether the second disorder is associated with any impairment in addition to the impairment that is associated with the first disorder. Another measurement issue, as was discussed earlier, is related to the role of informants. There are differences in informant reports not only for symptoms

[2] See Cramm, H., Aiken, A.B., and Stewart, D. (2012). Perspectives on the International Classification of Functioning, Disability, and Health: Child and Youth Version (ICF-CY) and occupational therapy practice. *Physical and Occupational Therapy in Pediatrics, 32*(4), 388-403.

but also for impairment. The Bella study, which measures mental health in children, adolescents, and young adults in Germany, found differences between parent and child reports, and research based on the original OCHS found differences between parent and teacher reports. Another issue is that if not all disorders are measured in a study, the result can be data on impairment but no data on disorder.

Szatmari said that he agreed with Canino that the ICF-CY framework is more nuanced. It allows for variation by sex/gender, culture, context (school, home, community), and so on. As Canino noted, functioning is a continuous concept, and cutoffs that determine impairment have not yet been validated. There is virtually no work in this area establishing what is and what is not impairment and how mild impairment differs from substantial impairment.

One important question that has been raised is whether measures of impairment need to be disorder specific. Szatmari argued that the impairment associated with ADHD can be different from the impairment associated with anxiety or mood disorders.

A related question is whether measures of impairment need to be developmentally specific and whether measures exist that are relevant at different developmental stages. Focusing on the very challenging issue of measuring impairment in the 2- to 6-year-old age group, Szatmari pointed out that symptoms or disorders have a very different impact on family, peers, school, and community at this developmental stage than they do for older age groups. For example, for the younger age group, caregiver burden is a very important measure of impairment, and the Parent Stress Index is one example of a scale available to capture that. There are also several other scales that are very good at measuring social relationships for infants.

Another concept that is very important to measure for the 2–6 age group is the acquisition of developmental competence in motor skills, language, and communication. Szatmari said that the Vineland Adaptive Behavior Scale is the best for this purpose. Two other very useful instruments are the Adaptive Behavior Scale and the Ages and Stages Questionnaire. The question is whether developmental assessments are the same as measuring impairment, and researchers in the infancy field would say that they measure the same concept, but there may be others who do not agree.

The EDI, discussed earlier, is a measure that captures school readiness, which is also very relevant for children between 2 and 6 and can be considered an aspect of impairment. A final measurement challenge noted by Szatmari as relevant for children younger than 6 is that, at this developmental stage, disruptive behavior disorders, autism spectrum disorders, and impairment are very difficult to disentangle. Because of that, it is best to measure all three.

In summary, Szatmari emphasized that impairment is important to measure because it drives the need for treatment. However, the measurement of impairment is complex, multidimensional, and influenced by multiple factors, not just disorder. The use of multiple informants is also necessary, and the role of teachers as informants would deserve further consideration. Furthermore, attempting to apply one construct of impairment to all child and adolescent developmental stages is an oversimplification. Ultimately, as Canino had noted earlier, the ICF framework is more useful when thinking about impairment than the DSM because it is more nuanced.

Ideally, Szatmari said, SAMHSA would count children with any diagnoses (not just specific diagnoses) and any impairment. This concept of serious emotional disturbance would be really useful for the next round of epidemiological studies. Including teachers as informants is very important, particularly when it comes to measuring impairment in addition to symptoms, because teachers in primary school know the children very well. Moreover, at least in the Canadian experience, collecting teacher data was not particularly challenging. More work is needed to develop instruments appropriate for each developmental stage, but having a global measure of impairment that is independent of symptoms and disorder makes a lot of sense: otherwise, the causal relationships are too entangled.

LESSONS FROM AUSTRALIA

Ian Hickie (University of Sydney) began his presentation by noting that he serves on the Australian Mental Health Commission and recently participated in a meeting at the Australian Academy of Health and Medical Sciences. The meeting included people from national health agencies and welfare and social science agencies, researchers, and others. At the meeting, the researchers had one view of what is needed, while those involved in policy making argued that the research never answers the questions that are most relevant to public policy. This is similar to SAMHSA's predicament, he said. The discussion needs to focus on the fundamental question of the purpose of specific programs and what kinds of measurement and instruments will deliver the kinds of information that are of use to those who are responsible for those programs.

It is important to note that there are some differences between surveillance activities and national health surveys, Hickie said. Surveillance methods are primarily applicable when the goal is to accurately measure changes in incidence or populations, or in regional variations, and the depth of the data is not a priority. Examples might be suicide prevention programs, or alcohol and drug use programs, if the intent is to enable a rapid public response. By contrast, the current focus of most national health surveys is resource allocation, and the Australian government clearly states that it is interested in return on investment. In certain ways, the Australian system is similar to the U.S. system, where taxes are collected by the federal government and returned to the states on the basis of resource allocation considerations. Because there is constant competition for these resources among health and social service agencies, key issues are the return on investment and the strategies of reducing impairment over time.

Referring to the prior discussions about instrumentation challenges, Hickie commented on the issue raised by Szatmari about the relationship between symptoms and impairment. In the DSM tradition, symptoms are asked first, followed by impairment. However, nearly all of the longitudinal studies in Australia show that the impairment has its onset earlier than a diagnosable disorder. Consequently, Hickie argued, instruments should ask about impairment first. In other words, the first question should be whether a child is impaired, and if so, what is the cause or what is the nature of the impairment. This is the opposite of the current common practice of asking whether the child has any symptoms, and then asking whether he or she has an impairment. The approach of asking about impairment first would pick up on a different set of problems than asking first about a symptom, and it better reflects the real world. The issue of the age of onset of the impairment further underscores that symptoms and impairment have different trajectories and likely respond to different sets of interventions.

Hickie noted that there are also positive aspects to the DSM system: for example, the early diagnostic interview schedules were very useful to Australian researchers in the 1990s, when several national surveys were introduced, many based on methods developed in the United States. These early efforts were very useful because the then-prevalent public perception in Australia was that mental disorders in children were rare and that substantial resource investments were not needed. New national data revealed the true magnitude of the problem and led to discussions of the concept of an unmet need. The related notion of a “met un-need” also emerged around the same time to describe services provided to those who may not really need them, according to a concept of optimization.

Hickie discussed research based on work done by the U.K. Government Office for Science, on the topic of the mental wealth of nations.[3] This research shows that early intervention is key. Although it seems obvious to want to provide the most services to those who are most severely impaired and have an obvious need for care, it is not clear whether that is the best return on investment in these situations. As Szatmari had discussed with regard to cancer care, in the past 30 years there has been increasing emphasis on early intervention: early diagnosis and treatment have reduced the progression of illness and the morbidity associated with cancer. In contrast, the concept of early intervention is not widely accepted in the field of child development, Hickie said, but to build the most mental capital, early intervention strategies are needed. And when looking for early evidence of deviations, they will be in impairment, not symptoms.

One challenge with using cross-sectional surveys in the area of mental health is that there is a high variability in testing on any particular day, especially for symptoms. Another challenge is that the data do not contain enough information about long-term implications. Furthermore, understanding prevalence is not enough. As Szatmari noted, surveys from across the world all show the same prevalence rates, and the surveys also consistently show that when impairment is factored in, the rates are lower. Getting the estimates right is easier for adolescents than younger children, but if the age of onset is obtained, those data can inform resource allocation and interventions.

Hickie described two initiatives in Australia focused on early intervention and risk factors. One is a population-based mental health promotion framework, centered around four topics: (1) individual brain development; (2) the family and social context of development; (3) the educational and experiential aspects; and (4) autonomy, social connection, and physical activity. The other initiative is centered around effective prevention and early and continuing intervention for mental health problems. It focuses on (1) early universal prevention (reducing risks, trauma, alcohol, drugs, inappropriate technology use) and reducing harms (suicide and self-harm); (2) early intervention for anxiety and depression; (3) continuing care supported by technology; and (4) building new service delivery platforms.

As in Canada, Australia also began placing a lot of emphasis on tracking early development, particularly on school readiness. Recently, the age of preschool screening was lowered from 4 to 3, and there is a desire to attend to problems early. For example, in the autism area, there is a lot of

[3] Beddington, J., Cooper, C.L., Field, J., Goswami, U., Hupper, F.A., Jenkins, R., Jones, H.S., Kirkwood, T.B.L., Sahakian, B.J., and Thomas, S.M. (2008). The mental wealth of nations. *Nature, 455*, 1057-1060.

focus on the period between the ages of 2 and 3. These initiatives have been very controversial because some have argued that they are driven by the pharmaceutical industry, but Hickie emphasized that the initiatives are driven by data from the early development indices.

Australia now also has an assessment, the National Assessment Program–Literacy and Numeracy (NAPLAN), which looks at educational assessment after entry to school, in school years 3, 5, 7, and 9. The plan is to use these assessments for research beyond just population mapping, including individual-level emotional tracking.

Hickie also noted that there has been interest in metrics such as disability-adjusted life years, which is a measure of overall disease burden, expressed as the number of years lost due to disability in the population. This measure can be helpful for understanding the persistence of emotional difficulties and the impacts on educational and social development over time.

Another development in Australia is the transition to a National Disability Insurance Scheme, which is trying to provide assistance to parents on the basis of individual need rather than diagnostic categories. An important component of this scheme is the magnitude of impairment and the amount of support that may be needed on an ongoing basis. The national surveys conducted since the 1990s have provided continuous data on service utilization, and young people continue to have the lowest rates of service utilization, despite funding invested into changing this. As Druss also noted in the previous session, the survey data on this topic are very useful for public policy because they can shed light on whether the programs are working. As Kathleen Merikangas also pointed out earlier, comorbidity is a major issue, which means that understanding the use of physical health services is also important.

Australian government surveys have the same limitations as surveys elsewhere, Hickie said, with not enough emphasis being placed on whether they are designed in a way that enables linkages to other data sources. There is increasing interest in this, and the goal should be to be able to link surveys not only to other sources of health data, but also to other information, such as education and welfare databases. So far, privacy concerns have limited the availability of longitudinal databases. To the extent that longitudinal data are available, these have generally been conducted by research agencies, rather than the government. The Brisbane Longitudinal Study of Adolescent Twins is one example of such research, which has provided useful data on the persistence of impairment.

Australia also faces challenges similar to other countries for telephone surveys, which are becoming increasingly difficult to field as more households abandon landlines. In-person interview visits are also challenging to conduct in Australia, Hickie said. However, online technologies turn out to be a very good way to collect in-depth information on health-related topics, and Australia has been experimenting with online sampling of young people as part of a large investment by the government in a cooperative for research on young people. One advantage of online data collection is that respondents are more likely to report certain things, such as suicidality and substance use, online than in a telephone interview. Online data collections can also be more flexible in accommodating modules of different questions administered to subsets of the sample. Tracking over time is also possible in online surveys.

Hickie said that a lot of emphasis has been placed in the field on standardizing instruments, but this focus assumes that data collections will continue to be based on the same types of approaches as they have been for the past couple of decades. Although the sampling challenges are substantial, especially if the goal is to produce nationally representative data, technological developments represent an opportunity for government agencies to rethink their approaches.

In conclusion, Hickie reemphasized the importance of the implications of age of onset and of obtaining accurate impairment data. He believes that beginning with impairment and then finding out the cause of the impairment would be a better approach to collecting these data than the traditional way of starting with disorder and then asking about impairment.

ADVANTAGES AND DISADVANTAGES OF THE STRENGTHS AND DIFFICULTIES QUESTIONNAIRE

Robert Goodman (Kings College, London) joined the workshop briefly by telephone to comment on the Strengths and Difficulties Questionnaire (SDQ). He noted that as one of the developers of the SDQ, he has a vested interest in it, and therefore a potential conflict. He said that the group should discuss the relative merits of the SDQ without his participation, but that he would briefly describe the potential usefulness of the instrument and answer any questions. He also noted, going forward, electronic use of the SDQ will require a license fee, while the paper versions will continue to be available free of charge.

Simply put, Goodman said, the SDQ works surprisingly well in studies of the general population, considering the complexity of the concepts measured. The instrument, which includes 25 items on psychological attributes, is relatively brief and can be used for children as young as 3. Despite its simplicity, the SDQ has good mathematical properties. For example, at least in the United Kingdom, the mean SDQ difficulties score can be used to predict the prevalence rate of mental disorders as judged by a much longer, independent diagnostic assessment.

The SDQ also includes an impact supplement and measures impairment with reasonable psychometric properties. In addition, what is more relevant to the public and policy makers, one can determine from the SDQ what proportion of parents who are reporting on their children believe that their child has a definite or severe level of difficulty.

Kathleen Merikangas asked Goodman to comment on how the SDQ could tap into meeting SAMHSA's goals of measuring both the DSM-equivalent diagnosis and impairment. Goodman replied that there is uncertainty about whether this would work in the United States. In the United Kingdom, there is a strong relationship between the mean SDQ total difficulty score in a particular population and the prevalence of DSM diagnoses in that population. For use in other countries, it would be important to start with a validation study, as one would with any other measures. There are datasets in the United States that could be used to do this testing, and if it turns out that it works, this would be a very powerful way statistically to look at the prevalence of DSM disorders in a particular sample. In terms of the SDQ impairment measure, which is simply asking respondents to judge how much the difficulty is interfering in everyday life or causing distress, it has psychometric properties that are as good as any.

Heather Ringeisen asked Goodman about whether he has done any research on the possibility of shortening the SDQ to a handful of essential items, for example, five or six items that would be well suited for use in time-sensitive national surveys. For example, the National Health Interview Survey includes a five-item version of the SDQ, and an additional impact item. Goodman replied that he has done this type of research and that he is a minimalist by nature, but in his view the short version of the instrument does not work well, either for one specific disorder or for predicting the presence of any mental disorder. Goodman added that he wishes that it would work, but the short version does not have enough sensitivity or specificity. However, he said, it would be possible to scale back the questions to just one—asking parents in a broader way whether they think their child has a problem or not. That approach would potentially be useful. However, between the one question and the current length of the SDQ, no other length really makes sense.

Susanna Visser noted that the Centers for Disease Control and Prevention has data from community-based studies that involved parents and teachers completing the full SDQ. The SDQ scores were used to identify children who would be invited for a follow-up diagnostic evaluation as part of a multistage design. The current sample includes approximately 1,000 families, and new families are being added. These data could be used to carry out a validation study on a U.S. population.

Canino asked Goodman to clarify why he used the word "impact" instead of "impairment." Goodman said that the DSM criteria specified both distress and impairment, and it seemed useful to link the two constructs together to retain the distress component. The impact portion of the instrument first asks whether the person has any difficulties related to concentration, emotion, behavior, or relationships. If there are difficulties, several follow-up questions are asked, including whether the difficulties cause distress and whether the difficulties interfere with everyday life in the key areas of family, learning, leisure, and peer relationships. The impact score is derived on the basis of the follow-up questions about distress and interference with everyday life.

4

Design and Estimation Options

LESSONS FROM THE NATIONAL SURVEY OF CHILDREN'S HEALTH

Michael Kogan (U.S. Department of Health and Human Services, Maternal and Child Health Bureau) provided an in-depth overview of the National Survey of Children's Health (NSCH).

The Maternal and Child Health Bureau (MCHB) has directed and funded two related surveys in the past, the NSCH and the National Survey of Children With Special Health Care Needs (NS-CSHCN), and these two surveys are being combined as part of an overall redesign. The NSCH was fielded in 2003, 2007, and 2011-2012, with the goal of producing national and state-based estimates on the health and well-being of children, their families, and their communities. Each sample included approximately 100,000 children. The NS-CSHCN was fielded in the intervening years, that is, in 2001, 2005, and 2009-2010, and its purpose was to assess the prevalence and impact of special health care needs among children and to evaluate changes over time. Each administration of the survey included approximately 40,000 children.

Both the NSCH and the NS-CSHCN were conducted in partnership with the Centers for Disease Control and Prevention and were fielded as a module of the State and Local Integrated Telephone Surveys. Both surveys began as random-digit dial surveys, limited to landlines, but recently began including a cell phone sampling frame, due to the challenges with landline-only telephone interviews. In the case of both surveys, the data are parent- and caregiver-reported.

Data from these surveys are used extensively by the federal government and others. One use is in Healthy People, a federal initiative that provides science-based, 10-year national objectives for improving the health of Americans. There are 15 Healthy People 2020 objectives that have been derived on the basis of the data from these surveys. These surveys have also been used for federal policy and program development in a number of areas. In cooperation with states, MCHB funds most of the maternal and child health services in the country through Title V block grants.[1] States are required to spend 30 percent of their funds on children's special health care needs, and the data from the surveys are used extensively in the Title V needs assessments.

The availability of state-level data from these surveys also enables other state-level planning and program development efforts. In addition, the data are used for scientific research. The MCHB also has systems and life-course indicators on a variety of topics and has published chart books devoted to mental health conditions.

Enabling easy access to the data has always been a priority, and the program has had an online data query system from its beginning. The query system is used extensively by researchers, policy makers, the media, and states.

The objective of the surveys was never limited to just health conditions. In addition to general health and specific health conditions, the surveys collect data about various aspects of children's lives, such as their families, their neighborhoods, and their health care experiences.

There are three versions of the instruments, for different age groups: birth to age 5, school-age children between the ages of 6 and 11, and adolescents between the ages of 12 and 17. The instruments have eight core content areas and include age-specific content. For example, as has been discussed earlier, school readiness has emerged as an important area in a number of countries, so an extensive set of questions was added on that topic.

The NSCH asks about the prevalence of a number of medical conditions, specifically, whether a doctor or other health care provider has ever told the parent or caregiver that the child had the condition. The list of conditions includes anxiety problems, depression, behavioral or conduct problems, substance abuse disorder, developmental delay, intellectual disability, speech or other language disorder, learning disability, attention deficit disorder or attention deficit hyperactivity disorder (ADD/ADHD), autism and autism spectrum disorder, and Tourette syndrome, as well as a fill-in-the-blank question for any other mental health condi-

[1] For more information on the Title V Maternal and Child Health Services Block Grant Program, see http://mchb.hrsa.gov/programs/titlevgrants/ [October 2015].

tion. The survey also asks about treatment, including the receipt of any treatment or counseling from a mental health professional, problems receiving needed mental health treatment or counseling, and several other treatment-related questions.

The survey also includes several questions about functioning. Among these are the frequency and degree of limitations and missed school days, as well as related items on parental concern about development or behavior, a developmental screening scale for those between 9 and 71 months (up to age 6), and school readiness.

As part of the redesign that is combining the NSCH and the NS-CSHCN, a major change is targeted at addressing the issue of declining response rates. When these surveys were first fielded, almost everyone had a land line. By 2014, more than 50 percent of households with children did not have land lines. Although the agency recently began including cell phones, cell phone households are much more difficult to survey, in part because there is no national registry, and they are more expensive to include. Consequently, the surveys simultaneously faced declining response rates and increasing costs. Due to these developments, the survey is currently being transitioned from a land line and cell phone sampling frame to an address-based sample. The first contact will be by email, and sample members will be encouraged to complete the survey online. They will also have the option to return the questionnaire by mail or call a telephone number if they prefer to complete the survey by phone.

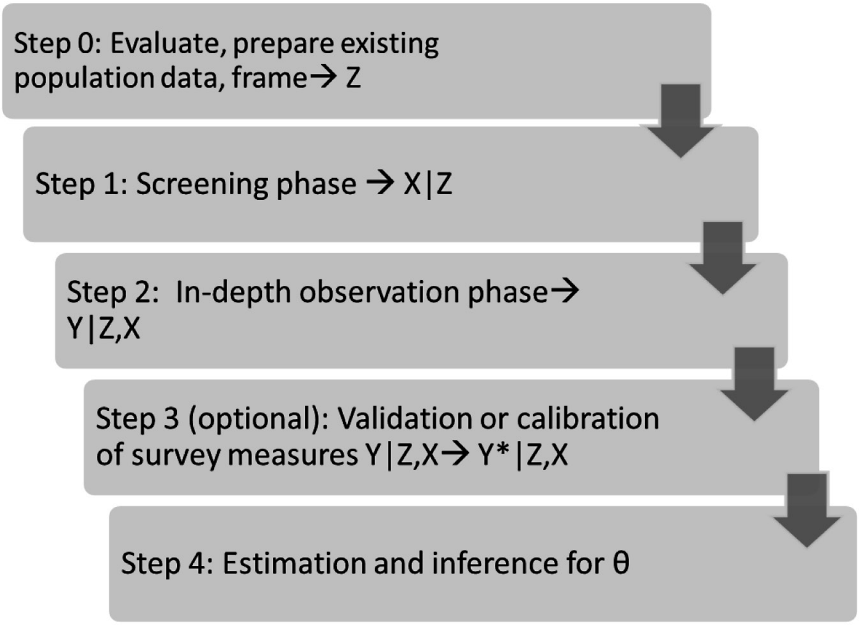
Finally, another goal of the redesign is to provide more timely data. In the past, each survey was conducted every 4 years. Because states would like to have the data more frequently, the new combined survey will be conducted annually, with a smaller sample. The redesigned survey is currently undergoing testing, including for mode effects, and MCHB anticipates that it will be ready to be fielded in the summer and fall of 2016, with a public data release scheduled for spring 2017.

CONSIDERATIONS IN MULTIPHASE STUDIES

Steven Heeringa (University of Michigan) discussed design and estimation considerations for multiphase studies of child serious emotional disturbance, primarily focusing on the statistical objectives, multiphase design choices, and measurement error. Earlier presentations and discussion had focused on the psychometric properties of some of the available measures, and Heeringa's talk focused on the statistical implications of the various levels of sensitivity and specificity associated with the measures. Heeringa also discussed examples of estimation and inference across a variety of settings in which existing data sources (longitudinal surveys, cross-sectional surveys, or administrative datasets) have varying degrees of richness.

From a statistical perspective, it is important to distinguish between two objectives for a screening study. One potential objective is simply to estimate prevalence and population size, in other words, to be able to report the number of affected children. However, invariably, when this type of data is collected, researchers want to turn to a second level of analysis, to conduct an in-depth subpopulation study, to better understand the characteristics of the population, the age of onset, the associated factors, the life course and development of the disorder, and so on. It is important to distinguish between these objectives because they have different optimal properties in terms of research design.

In multiphase studies, researchers are incrementally refining the information they have on the units of observation. Figure 4-1 illustrates a multiphase design framework.



Step 0: Evaluate, prepare existing population data, frame→ Z

Step 1: Screening phase → X|Z

Step 2: In-depth observation phase→ Y|Z,X

Step 3 (optional): Validation or calibration of survey measures Y|Z,X→ Y*|Z,X

Step 4: Estimation and inference for θ

FIGURE 4-1 Multiphase design framework.
NOTES: Z is existing population data, frame; X is phase 1 screening data; Y is phase 2 measurement of outcome of interest; Y* is validated, calibrated outcome of interest; θ is population parameter.
SOURCE: Workshop presentation by Steven Heeringa.

Heeringa said that the discussion during the workshop underscored the importance of beginning the process with an evaluation of what is already available (Step 0). A multiphase survey is expensive, and if one is needed, the best design optimizes inferential efficiency based on the existing data sources. The initial research will lead to variables *Z*, which are the existing data on the population or a large representative sample of that population.

The next step in the process is a screening phase (Step 1), to obtain additional screening data, potentially based on some of the screening measures that were discussed earlier in the workshop. This screening can be completed at relatively low cost for the screening observation, on the entire sample or on a set of observations.

A second-phase data collection is then carried out (Step 2), which is typically limited to a subsample due to high costs for each observational unit. At this stage, the measurement *X* is refined to get closer to the measurement of interest, *Y*, which might be a diagnosis, an index of severity, or an index of limitation. If *Y* is an imperfect measure, it can be further validated or calibrated (Y^*), for example, against a clinical standard. The final step is the estimation and inference of population attributes.

Heeringa noted several key cost considerations that come into play for multiphase designs. One is the expected prevalence rate in the target population. The expected prevalence rate for serious emotional disturbance will drive the sample size requirements to obtain certain levels of precision. If the aim also includes the second objective, he noted, which is to identify a subpopulation to study in depth as a domain of analysis or to follow longitudinally, the population prevalence for the characteristic of interest will determine the number of cases needed to ultimately achieve the appropriate sample size of eligible individuals.

In some cases, a first-phase screening (Step 1) might not be necessary, especially if sufficient information is available from the existing data used for a sampling frame. In these cases, a screening status could be assigned on the basis of the existing data. Most of the double sampling theory in the relevant literature is based on designs in which the cost for the screening is low relative to the cost of the in-depth observation phase (Step 2). This approach is often used in medical screening studies, when an in-depth clinical interview is used in the second phase.

The sensitivity (true positive rate) of the Step 1 screener is another cost factor because a high false positive rate—particularly if the objective is to identify a sample of individuals who meet the criteria of interest for further study—will necessitate starting with a larger sample size simply to observe a certain number of individuals who meet the criteria. Finally, whether a calibration study will be conducted is also a consideration.

The error factors complement the cost factors, Heeringa noted. Again, the prevalence of the target group is relevant because it drives the sampling variance. It also matters what the strength of association is between the data from the Step 1 screening and the data from the sampling frame, as well as between the final outcome measures of interest and both the sample frame data and the Step 1 screening data.

Another error factor is the specificity (true negative rate) of the screener, because in order to ensure coverage of all true cases in the population, a second-phase subsampling is needed of persons who are negative screens. A high false negative rate based on the Step 1 screener implies a need for variable weighting of Step 2 in-depth observations. When the subsampled negative screen cases are followed up, some of them will, in fact, turn out to be positive, and this means that variable weighting will be needed. Variable subsampling and weighting of Step 1 positive and negative screens increases the variance of population prevalence estimates and inflates variances of estimates for analyses of true subpopulation cases. Finally, if a calibration is needed, there is a potential for classification bias.

Table 4-1 illustrates the effects of sensitivity and specificity on the sampling variances of the estimates that are derived (or modeled) on the basis of screening data. The rows show the screening assessment from Step 1 or from a modeled assignment based on an administrative dataset. Some individuals are classified as negative screens and others as positive screens. For simplicity's sake, the table assumes a dichotomous measure, but the measure can also be continuous. Samples of individuals are drawn on the basis of the information from Step 1. Typically, a stratified subsample of the negative screens would be generated, and all of the positive screens would be included (sampled with a probability of 1. A follow-up is then conducted to obtain an observation on the true status or the proxy of the true status, *Y*. In the Table 4-1 example, the screener specificity and screener sensitivity are both 0.8.

TABLE 4-1 Measurement Example: Sensitivity/Specificity of Step 1 Screening

Step 1 Screening or Model Assignment	Step 2 Observed Status (Y)		
	NO (0)	YES (1)	Total
NO (0)	$P_{00} = 0.64$	$P_{01} = 0.04$	$P_{0+} = 0.36$
YES (1)	$P_{10} = 0.16$	$P_{11} = 0.16$	$P_{1+} = 0.64$
Total	$P_{+0} = 0.80$	$P_{+1} = 0.20$	$P_{++} = 1.00$

NOTES: True prevalence, 0.20; screener sensitivity, $P_{11}/P_{+1} = 0.8$; screener specificity, $P_{00}/P_{+0} = 0.8$.
SOURCE: Workshop presentation by Steven Heeringa, June 2015.

TABLE 4-2 Examples of Weighting Loss in Variance of Estimates of Population Prevalence Due to Step 2 Subsampling of Step 1 Negative Screens

Step 1 Positive Screens (f_{pos})	Step 2 Subsampling of Step 1 Negative Screens ($f_{\text{neg, sub}}$)	Percent Increase in Variance (p)
1.0	0.50	8
1.0	0.33	22
1.0	0.25	36
1.0	0.10	130

NOTE: True prevalence, 0.20.
SOURCE: Workshop presentation by Steven Heeringa, June 2015.

Heeringa noted that to compensate for the differential subsampling, conditional on the Step 1 determination, some weights are carried forward, and there is going to be a loss in the precision of the prevalence estimates. Table 4-2 provides examples of weighting loss in the variance of population prevalence estimates due to Step 2 subsampling of Step 1 negative screens. In the case of a first-stage screening with sensitivity of 0.8 and specificity of 0.8, if all of the positive screens are included in the follow-up and 50 percent of the negative screens are subsampled, there would be a relatively small increase in variance, approximately 8 percent. However, this subsampling rate would be more expensive due to the high fraction of Step 1 negative screens that require the higher cost Step 2 assessment. As the subsampling rate is lowered, the impact on the variance of the estimates starts to increase substantially, eventually reaching well over 100 percent in the case where Step 1 negative screens are subsampled at a rate of 1 in 10.

When researchers use these types of designs, there is a tendency to subsample the negative screens at a very low rate "just to check." But a very low subsampling rate for the negative screens is only reasonable when there is a very high degree of specificity to the Step 1 design and measurements, Heeringa said. Typically, a follow-up rate of one in four, one in three, or one in two is needed for the negative screens.

Table 4-3 shows formulas for the expected disposition of Step 2 eligible cases in a two-phase design. When thinking about the second objective of trying to identify a subsample of individuals for in-depth study, one example would be a longitudinal study of children to determine whether at a particular developmental stage they meet the criteria of a serious emotional disturbance. The variance loss is going to be a little different in longitudinal studies because the differential weights are carried forward.

TABLE 4-3 Formulas for Expected Disposition of Step 2 Eligible Cases in a Two-Phase Design

	Step 2 Expected Eligible Cases	
Step 1 Screening or Model Assignment	Expected Sample Size Eligible True Cases	Relative Design Weight Step 1 Is Equal Probability of Selection Method
NO (0)	$E(m_{01}) = \frac{n \cdot (1-P)}{K} \cdot (1-Spec)$	$W_i = \mathrm{K} = 1/f_{\text{neg, sub}}$
YES (1)	$E(m_{11}) = n \cdot P \cdot Sens$	$W_i = 1.0$

NOTES: $E(m)$ is expected sample size; W_i is relative design weight. See text for discussion.
SOURCE: Workshop presentation by Steven Heeringa, June 2015.

Moreover, Heeringa noted, once it is introduced at baseline, the weighting penalty on variance is persistent.

Table 4-4 shows examples of weighting loss in the variance of estimated means for a Phase 2 subpopulation sample. In this case, the variance is a function of both the subsampling rate for the Step 1 negative screens and of the specificity of the Step 1 screener. The table shows that there would be an increase of more than 50 percent in variance if the specificity was 0.8 and one in four of the negative Step 1 screens were sampled.

Heeringa also talked about a framework for how to think about the process of integrating survey data with existing data, such as information from administrative sources. The pyramid in Figure 4-2 illustrates this framework. The top of the pyramid is characterized by a state of relatively little existing information. There may be a small amount of information on the survey frame, but very little or no information about Step 1 measures and most likely no information at all on Step 2 measures. Further down the pyramid, there is increasingly more information from existing data sources, including not only data needed for sampling purposes, but also information that would make it possible to effectively prescreen cases on the basis of an administrative data source, and the data source might even contain some measures of *Y*. This might be the case, for example, if another survey had been completed in prior years.

Heeringa explained that the top box in Figure 4-2 is the domain of traditional survey designs. Double sampling involves selecting a probability sample based on the frame information. Based on the limited data available, a Step 1 screening is conducted to ascertain a set of preliminary information. That information is then used to perform stratified subsampling and to follow up with the subsample for in-depth data collection.

TABLE 4-4 Examples of Weighting Loss in Variance of Estimated Means for the Step 2 Eligible Subpopulation Sample

Step 1 Positive Screens (f_{pos})	Step 2 Subsampling of Step 1 Negative Screens ($f_{neg, sub}$)	Step 1 Screen Specificity, in Percent					
		1.0	0.9	0.8	0.7	0.6	0.5
		Percent Increase in Variance (p)					
1.0	0.50	0	11	13	12	11	10
1.0	0.33	0	30	34	33	30	20
1.0	0.25	0	50	56	54	50	46
1.0	0.10	0	180	203	194	180	265

NOTES: True prevalence, 0.20; sensitivity, 0.8.
SOURCE: Workshop presentation by Steven Heeringa, June 2015.

In some situations, more information may be suitable to incorporate into the survey design and inference from the survey data collection, which is often referred to as model-assisted survey sampling: this is shown in the second box of the pyramid. Survey-assisted modeling, shown in the third box in the pyramid is the direction many of the survey programs may take in the future. This box covers situations in which there is a large amount of existing data, such as administrative data, but the data do not contain all of the measures of association and covariates that may be needed: the survey is used to measure and "fill in" missing information in a model framework. Finally, the bottom box of the pyramid represents a pure modeling environment in which the existing data make it possible to model estimates directly through data mining and analytics.

Heeringa offered an example of a double sampling approach (top box of the pyramid)—the Flint Men's Health Study, which collected data to inform decisions about levels of prostate-specific antigen (PSA) among African-American men.[2] In Step 0, the only information available to the researchers was an area probability sample frame for Flint, Michigan, households. Step 1 involved screening a probability sample of households for African-American men over age 40, with some disproportionate sampling of census tracts to improve efficiency. As part of the same process, a health history interview was conducted and a blood sample was obtained

[2] Heeringa, S.G., Alcser, K.H., Doerr, K., Strawderman, M., Cooney, K., Medbery, B., and Schottenfeld, D. (2001). Potential selection bias in a community-based study of PSA levels in African-American men. *Journal of Clinical Epidemiology, 54*(2), 142-148.

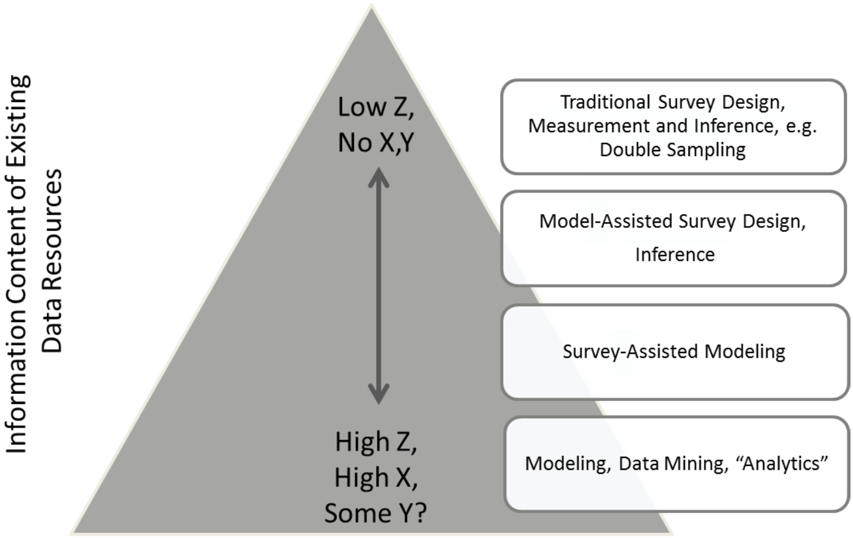

Information Content of Existing Data Resources
Low Z, No X,Y
High Z, High X, Some Y?
Traditional Survey Design, Measurement and Inference, e.g. Double Sampling
Model-Assisted Survey Design, Inference
Survey-Assisted Modeling
Modeling, Data Mining, "Analytics"

FIGURE 4-2 Integrating survey and administrative data: Adaptation to information content of available data.
SOURCE: Workshop presentation by Steven Heeringa, June 2015.

for a PSA test. In Step 2, stratified subsamples were selected, based on the PSA data that were obtained in Step 1, and the men who were selected into the sample were referred for further testing to determine probable cancer. The Step 3 validation involved an actual biopsy to confirm the findings from Step 2.

Heeringa next offered an example of a model-assisted survey design, the Aging Demographics and Memory Study (ADAMS).[3] For this project, more information was available in Steps 0 and 1. In Step 0 rich longitudinal data were available from the Health and Retirement Survey (HRS), which enabled the researchers to estimate a logit model of the probability of dementia on the basis of an external dataset. Given the existing information in the HRS and the externally derived model of dementia, the HRS panel frame was stratified by age, gender, and cognitive score. Step 1 involved screening and stratified subsampling for follow-up of the HRS panel, which was based on a stratification that used an externally estimated model relating probability of dementia to age, education level,

[3] Langa, K.M., Plassman, B.L., Wallace, R.B., Herzog, A.R., Heeringa, S.G., Ofstedal, M.B., Burke, J.R., Fisher, G.G., Fultz, N.H., Hurd, M.D., Potter. G.G., Rodgers. W.L., Steffans, D.C., Weir, D.R., and Willis, R.J. (2005). The Aging, Demographics and Memory Study: Study design and methods. *Neuroepidemiology, 25,* 181-191.

and cognition test scores. Step 2 involved in-home neurocognitive assessment and collection of medical records, which was followed by a consensus diagnostic conference review by an expert medical panel to assign a diagnosis category. Step 3 was a 2-year follow-up to refine probable or possible dementia into cognitive impairment with no dementia and dementia categories. These steps essentially involved direct estimation, and the key benefit of the existing cognitive information was the ability to use it in the HRS panel to establish strata for disproportionate sampling.

Heeringa also discussed how survey-assisted modeling was used as part of the same study. The analysis was not limited to the 850 cases who cooperated as part of the process described above, and for which the Z, X, and Y data were obtained. For approximately 10,000 cases of HRS panel members over the age of 70, Z and X data were available, including cognitive scores and covariate information from the HRS. For these cases, it was possible to use the relationships observed in ADAMS to draw strength from the data available on Z and X.

From the ADAMS survey data (training set), a “best” predictive model, $f(Y \mid Z,X)$ of the probability of dementia was estimated. The predictive model estimated from the multiphase ADAMS survey was then used to predict the probability of dementia for each eligible member of the larger HRS panel. Using the predicted or “imputed” probabilities of dementia generated for all eligible HRS panel members, estimates of population prevalence for dementia were produced using methods that properly reflected the uncertainty associated with the modeled values of Y^*.

Heeringa concluded the discussion of survey-assisted modeling by saying that predictive modeling approaches assign classification probabilities to all elements in the population frame (or weighted sample). Clinicians may always want a diagnosis, in other words a 0 or 1, but from a statistical perspective, a decision has to be made on whether to conduct the analyses on a probability scale or use the probabilities to impute discrete classifications. It is also important to ensure that inferences reflect the prediction (imputation) inherent in modeled values. Once the uncertainty associated with the predictor is incorporated, either through multiple imputation or some other method, these predicted probabilities can then be added to the data file and used not just as dependent variables, but also as independent variables. In the case of child serious emotional disturbance, once these probabilities are assigned to a child’s data record, for example, in the National Survey on Drug Use and Health (NSDUH), the data can be used both to analyze prevalence and to look at the impact on other outcomes for that particular individual.

Jonaki Bose (SAMHSA) asked Heeringa to comment on the number of predicted probability values needed. Heeringa replied that Bose’s question captures the uncertainty in the training model from Step 2, when the imputation is performed, or even if the second-stage investigation was used to set a cut point for everyone else in the first stage. Ideally, it is important to incorporate a measure of the uncertainty in setting that cut point. It can be treated as a missing data problem, where *X* is available for each case, and *Y* is only available for some cases. This is a structured missing dataset, so *Y* is imputed for everybody who has *X* but not *Y*. Heeringa added that he found that five imputations generally provided stability, but it depends on the fraction of the missing information. Some researchers are dealing with 5 or 10 imputations; others are recommending as many as 20 or 100.

Graham Kalton (Westat) commented on the question about whether to retain the probabilities, which was done in the example, or dichotomize the data. He noted that dichotomizing at the beginning of a multiple imputation would involve swapping between 0s and 1s, and that would contribute a lot of extra variance. Heeringa agreed that working directly with the probability scale is more efficient, but added that in practice the concept of a probability of having a disorder or holding an attribute is complicated to explain to researchers who want to use the data, and because of that a discrete classification may be needed.

Bose asked Heeringa whether he would consider the probability scale a true continuous measure. Heeringa replied that he would, at least in these cases. It would be important, however, to look at the distributional properties, in other words, at normality and censoring of the tails at 0 or 1 in the case of linear regression modeling.

MODEL-BASED ESTIMATES OF PREVALENCE IN A NATIONAL SURVEY

Heather Ringeisen and Jeremy Aldworth discussed research that RTI International conducted on behalf of SAMHSA to develop model-based methods for use in an existing national survey to estimate the prevalence of serious emotional disturbance in children and serious mental illness in adults. Ringeisen discussed the work related to serious emotional disturbance in children, and Aldworth described some of the technical details and lessons learned from both efforts.

The Child Serious Emotional Disturbance Pilot Study

To provide some background on the model-based estimation project for children, Ringeisen said that in 2006, SAMHSA convened an advisory panel that made several recommendations about what a study to estimate child serious emotional disturbance might look like. At that time, it was decided that the short version of the Strengths and Difficulties Questionnaire (SDQ) included in the National Health Interview Survey would be used to generate a score to be used as a predictor variable. The NHIS version of the SDQ is different from the one described by Goodman (see Chapter 3) because it is a five-item version that also includes an additional impact item that asks parents to report whether the child has minor, moderate, or severe difficulties. Because this is a standardized tool that offers a continuum of scores and at least a proxy indicator for impairment, this instrument and the NHIS appeared to be the best option for the project at that time. The added benefit of the NHIS was that it could collect this information for children as young as 4.

On the basis of the recommendations of the advisory panel, SAMHSA, and later the National Institute of Mental Health (NIMH), funded a pilot study with the primary objective to explore the development of a model-based procedure to estimate serious emotional disturbance in children, using data collected on the NHIS. The study was a collaboration with the National Center for Health Statistics (NCHS). RTI International conducted the pilot study in partnership with Duke University.

Ringeisen explained that the pilot study method was based on developing predictive models in which the five-item SDQ and a sixth impact item from the SDQ were the independent variables that predicted serious emotional disturbance, which was then determined on the basis of "gold standard" clinical interview data. The sample for the pilot study included two groups, both taken from participants in the final three quarters of the 2011 NHIS and the first quarter of the 2012 NHIS: one group was the parents of children aged 4-17; the other was youth aged 12-17 who reported about themselves. The sample was limited to children whose parents completed the NHIS interview in English, provided complete contact information and SDQ responses, and indicated the child had no history of intellectual disability, developmental delay, autism, or Down syndrome (in order to dovetail with the SAMHSA definition of serious emotional disturbance in children). The sampling strata were defined by NHIS SDQ scores and then sampled proportionally to the size of the standard error of a proxy measure of child mental disorder distributed across the strata.

Of the 1,187 identified parent respondents, 217 (18 percent) completed interviews. Interviews were also conducted with the youth group, which resulted in 78 cases where both a parent and a child interview were completed. From among the remaining cases, 195 were ineligible due to a competing study using the NHIS sample, 277 were not locatable, and 239 were not contactable. An additional 200 refused to complete the interview or ended participation before completing it.

A low response rate was one of the main challenges for the study. One of the reasons for the low rare was the 18-month delay between the NHIS data collection and the pilot study, so by the time the follow-up study was conducted many sample members were not locatable. In addition, several cases were pulled from the study after the initial work of preparing the sample was completed because they were needed for another project. The final response rate of around 18 percent was not high enough to give the researchers sufficient confidence in the statistical results, and therefore they are not included in her presentation, Ringeisen said. Rather, she focused on lessons learned and possible ideas for moving forward.

The delays in fielding the study were primarily due to negotiations about ethical and security concerns that are related to these types of studies. One issue was whether it was appropriate to interview children about mental health issues by telephone. Initially, the intent was to conduct telephone interviews with both parents and children in a broad age range, but feedback from the relevant institutional and ethics review boards resulted in limiting the child interviews to those between the ages of 12 and 17. There were also questions about whether it would be advisable and operationally feasible to report individual results to the participants. NCHS generally prefers to share survey results with the respondents by providing a summary of their data at the end of the interview or at a later time, but this approach is not typical in the mental health field. Finally, some of the delay was associated with developing an appropriate protocol for handling distressed child respondents during this type of telephone-based study. The negotiations were hindered by the lack of a clear precedent, Ringeisen noted.

Once the data collection could commence, a "gold standard" clinical interview was used as part of the study to ascertain serious emotional disturbance status. The instruments used were a shortened version of the Child and Adolescent Psychiatric Assessment (CAPA) for children age 8 and over and the Preschool Age Psychiatric Assessment (PAPA) for children between the ages of 4 and 7. The short version of the instruments included five modules: depression, anxiety, attention deficit hyperactivity disorder (ADHD), oppositional defiant disorder, and conduct disorder. The instruments also included an incapacities module that assessed impairment for all endorsed symptoms (absent, partial, or severe). Serious emotional disturbance status (positive or negative) was determined on the basis of responses to the five CAPA/PAPA modules as well as scores on the incapacities module. These instruments were the only diagnostic interviews that were designed to be complementary across the wide age range included in the study. However, Ringeisen said, using two versions of the instrument had implications for statistical modeling (which were discussed later by Aldworth).

Using the data available, a series of analyses were conducted to determine the issues associated with modeling the children's status using independent variables from the NHIS. There was no apparent consensus in the field about the appropriate cut point for impairment, so the decision was made to use the CAPA incapacities section to model three different possible definitions, based on three different impairment cut points: (1) presence of any disorder assessed and any partial or severe impairment rating, (2) presence of any disorder assessed and at least three partial or one severe impairment rating, and (3) presence of any disorder assessed and any severe impairment rating. The first definition was the most general, and the third one was the most restrictive, while the middle one was similar to the definition in the National Comorbidity Study Adolescent Supplement. As would be expected, Ringeisen said, the different cut points resulted in very different prevalence rates.

The models examined various ways to use the NHIS SDQ to predict child serious emotional disturbance. One approach was based on the total five-item SDQ score, which Goodman had earlier said that he does not think works very well. Another approach was based on the five individual SDQ item scores, both with and without the SDQ impact item. The idea was to try to determine which combination would perform best at predicting serious emotional disturbance status on the basis of the CAPA and PAPA interviews. The researchers also examined the impact of age on the calibration models, Ringeisen said, but due to small sample sizes, the results of these exploratory analyses were ambiguous.

As had been discussed earlier during the workshop, research indicates that obtaining more than one report on these topics is very useful, but due to institutional review board concerns, telephone interviews were only possible with children 12 and older. This approach meant that cases in the age groups where two reports were available had a higher probability of a positive serious emotional disturbance status. In addition, Ringeisen reminded the workshop participants that the data that included only parent interviews were collected with two different instruments, the PAPA for children aged 4-7 and CAPA for children aged 8-11. The CAPA was also the instrument used for the oldest age group (ages 12-17). Thus, from a statistical perspective, reconciling different age-dependent measures and definitions was also difficult.

Ringeisen said that there are several issues that still need resolution before attempting to implement model-based estimation of serious emotional disturbance in children. One issue is that there are difficulties with matching reference periods between screening and diagnostic assessments. The *Federal Register* definition of child serious emotional disturbance requires the presence of a past year mental disorder (see Chapter 1), and there are a number of diagnostic assessments that ask about the past year. The SDQ asks about behaviors exhibited in the past 6 months, while the CAPA and PAPA ask about symptoms in the past 3 months. Thus, there was a mismatch between the screening and clinical interview reference periods in the pilot study.

Another challenge, as several of the speakers noted, is that there is no consensus for how serious impairment should be measured or defined as an explicit cut point within particular measures. In the pilot study, three definitions were examined, each varying in the degree of severity required to meet the definition of serious impairment. These definitions each resulted in very different prevalence estimates across age groups, Ringeisen noted.

There are also several instrument limitations. The 1993 *Federal Register* definition is designed to cover children from birth to 18 years, but complementary diagnostic and impairment tools do not exist to measure the presence of an impairing mental disorder across this entire age span. There are also instrument limitations with regard to updates for DSM-5, because most clinical interview tools have not yet released updates, and some instruments will not have a DSM-5 update available. In addition, Ringeisen said, there are no studies that directly compare the impact of administration mode (e.g., telephone or in-person administration) for the leading lay-administered child diagnostic interviews, such as the CIDI, the Diagnostic Interview Schedule for Children, or CAPA.

Future research of this type will need to have sufficiently large sample sizes for each age group to support age-specific models. Additional thought will also need to be given to reconciling discontinuities between the three age groups (4-7, 8-11, and 12-17) and their implications for estimating serious emotional disturbance across ages. The ramifications of different types and numbers of reporters (parents, children, teachers) for statistical modeling also need to be better understood.

Aldworth provided additional technical details about the model-based estimation effort for serious emotional disturbance in children. He began by listing the assumption underlying the method. The first assumption is that the there is a "gold standard" measure that is based on a diagnosis from a clinical interview, and that this measure has yes/no values. To use a model-based method, which is cheaper than administering the clinical interview to everyone in the sample, it is assumed that the gold standard measure is the truth, even if the truth is not measurable or knowable. If there is any deviation between the gold standard measure and the truth, this is not accounted for using this method.

The steps in a model-based method to estimate child serious emotional disturbance can be described as follows:

- Administer a predicting scale, such as the SDQ, to all eligible respondents in the main survey.
- Select a subsample from among the eligible respondents in the

main survey and administer the clinical interview to obtain the gold standard measure for the subsample.
- For the subsample, use a logistic regression model to “match” the gold standard measure to the scale administered in the questionnaire. The gold standard measure is the response variable, and the scale from the questionnaire that was administered is the predictor variable. Other covariates associated with serious emotional disturbance can also be included in the model.
- Select a cut point to dichotomize predicted probabilities of serious emotional disturbance (yes/no) based on model.
- Extrapolate the model (and cut point) to all eligible respondents in the main survey to determine serious emotional disturbance status for each respondent, and obtain prevalence estimates, using the full power of the main survey

Aldworth noted that the purpose of the model-based method is to provide prevalence estimates overall and within subgroups, and not to provide an individual diagnosis. Consequently, minimizing bias is crucial, while minimizing the error rate is not as important. For more nuanced analyses, the error rates may become more important. In contrast, for individual diagnostic tests, the needs are very different and depend on the disease. For example, in the case of a disease such as Ebola, a false negative might be a death sentence, and so it is critical for the false negative rate to be near zero, irrespective of the false positive rate. In other words, in this case, bias is not as important.

The Adult Serious Mental Illness Study

Aldworth also discussed the work RTI has performed to model serious mental illness in adults. The project, referred to as the Mental Health Surveillance Study (MHSS), used data from the NSDUH, which included the Structured Clinical Interview for the DSM-IV. The methodology was similar to that used by RTI to estimate child serious emotional disturbance, but while the child study required three distinct serioius emotional disturbance measures that varied by age group and led to the decision to develop three different statistical models by age, the serious mental illness study involved a single measure for the entire adult age span.

Aldworth said that one challenge with the adult project was the effect of the weights. Optimizing the design to select a subsample led to large weights because the weights from the subsample were multiplied with the weights from the full NSDUH. One concern related to large weights is that they may have undue influence on determining the position of the cut point. The approach may work well in a subsample and produce an unbiased estimate, but it could become a problem when extrapolated to a main survey, which is used to produce the prevalence estimates. A more practical concern was that a large weight straddling the neighborhood of the cut point could make it difficult to equalize the false positive and false negative counts, resulting in bias in either direction, depending on the location of the cut point. To counteract some of these problems, Aldworth said, the research team modified the subsample design by "reversing" selected undersampling and oversampling in the main survey. This technique helped reduce that weight variability.

In addition to the impact of the weights, sample size limitations represented the other major challenge in the project to model serious mental illness in adults. The size of the subsample affected the gold standard estimates, as well as the model-based analyses and estimates. For the gold standard estimates, the measure is assumed to be truth for the selected respondents. However, Aldworth noted, if the subsample size is too small, the gold standard prevalence estimates may be subject to large design-based sampling error, particularly at the subgroup level. A model is used to match the gold standard prevalence estimates made available by sample design, but a model cannot address design-based sampling error, regardless of how good the model is.

A small subsample size can also affect model-based analyses and estimates. The weighted model-based analyses and estimates are subject to the same design-based sampling error, again, particularly at the subgroup level. The model-based bias and classification error are statistics that play an important role in determining the model, and they are conditional on the subsample prevalence estimates. If the subsample is small, the bias and classification error could also be affected. In addition, Aldworth noted, a small sample size could also affect regular model-based errors. There may not be enough data to identify the best model, resulting in unmeasured bias at the subgroup level, and the standard errors associated with the beta estimates of the model may be larger.

In the MHSS study, after 1 year of data collection, the researchers had a sample size of 750, and they developed a two degrees of freedom model using the Kessler Psychological Distress Scale (K6) and the World Health Organization Disability Assessment Schedule (WHO-DAS) as predictors. By the end of 5 years, data had been collected on 5,000 cases, which allowed determination of an improved five degrees of freedom model, including an age variable, past year major depressive episode, and past year suicidal thoughts, in addition to the K6 and WHO-DAS data. The age variable allowed for a substantial reduction in bias within age groups. Adding past year major depressive episode and past year suicidal thoughts resulted in large reductions in classification error. In other words, this is an example of a better specified model, in which the model error was reduced.

Aldworth noted that model misspecification error and model error will be transferred to the error in the prevalence estimates. However, these errors are not accounted for in the standard errors associated with model-based prevalence estimates obtained from the main study data, which only take the design-based sampling error into account. This is especially a problem with smaller sample sizes, where the errors that are not accounted for are larger, so more work is needed to address these technical difficulties.

Another challenge discussed by Aldworth was the unintended consequence associated with adding covariates. Although the model was improved by adding past year major depressive episode and past year suicidal thoughts, problems arose in any joint analyses with serious mental illness and the two new predictive variables. For example, trying to determine the proportion of adults who had both a major depressive episode in the past year and a serious mental illness tended to result in overestimates. One way to address this problem is to formulate the final model taking into account the list of variables assumed to be important in terms of planned joint analyses and create alternative models. The alternative versions of the final model can then exclude the covariates required for joint analyses from the model, as needed. The RTI team is working on this for the adult mental health project, and it could also be considered for the child study.

Aldworth concluded by saying that although predicted probabilities are theoretically continuous in a (0,1) interval, in practice there are only as many distinct values as there are in all combinations of the predictor variables. For example, if there is one predictor variable in a model with six levels, then there will be only six distinct predicted probabilities possible. If there are too few distinct predicted probabilities, gradations between them may be too coarse to identify a minimum-biased cut point. In practice, the cut point can occupy any value within the interval between gradations (consecutive predicted probabilities), which leaves some indeterminacy if the interval is large.

Next Steps

After Aldworth's presentation, Ringeisen summarized some of the next steps needed, particularly for prevalence estimates of child serious emotional disturbance. She said that a priority is to develop a well-operationalized definition of the concept, amenable to estimation in a national study. Such a definition should include recommended cut points for various age groups on available impairment measures. Operationalizing existing measures of impairment with concrete, developmentally grounded and culturally sensitive anchors is also needed to increase the accuracy of assessments. Tools are especially needed to assess impairment in the youngest children, she noted. The research community as a whole needs to build agreement around the best predictive and gold standard measures to use in estimating the national and state prevalence of serious emotional disturbance in children from birth to 18 years. As discussed, existing datasets could be leveraged to conduct analyses that can assess the power of various candidate tools for prediction in statistical models.

It would be useful to better understand how to address the seam effects that result from change in instrument type across the age span and in the required number of informants by child age. It would also be useful to have more information on the effects of data collection modes, in the context of both data quality and cost efficiency. Finally, it will be important to address challenges associated with the varying reference periods that exist between the *Federal Register* definition of child serious emotional disturbance, predictive tools embedded in candidate national surveys, and "gold standard" clinical assessment tools.

Kalton asked whether it would be necessary to include the respondent's state of residence as one of the variables in the model, such as the one used in the MHSS, in order to accomplish the goal of producing state-level estimates. Aldworth replied that it is yet to be determined whether there is bias at the state level. If there is, it might be possible to create a few state categories, instead of adding all 50 to the model. Bose added that given the sample size limitations, the assumption had to be made that the model would hold for all states. Ideally, there would be a large enough sample size to examine potential biases at the state level, but that may not be practical due to cost considerations. She added that the model that was used for producing estimates of serious mental illness in adults was found to be acceptable by SAMHSA and data users.

Dean Kilpatrick (Medical University of South Carolina) asked the group whether it would be worthwhile to reexamine whether the clinical interviews conducted by trained mental health professionals are truly the gold standard, as is generally assumed. Research on mode effects shows that the perceived anonymity associated with self-administered questionnaires often results in different responses. Bose replied that if the decision is made to use model-based estimation, a discussion of the gold standard would definitely be needed. Szatmari commented that he and his colleagues used a questionnaire that was widely criticized for not having the same credibility as a clinical interview, but when the correlates of classification by questionnaire were compared with the correlates of classification by clinical interview, the results were the same. In other words, both methods led to the conclusion that attention deficit hyperactivity disorder (ADHD) is associated with similar characteristics, including boys and poverty. The children may not be exactly the same, but the conclusion was that the two methods have the same construct validity. Bose added that fitness for use is an important consideration. The estimates of serious mental illness have an approximately 4-5 percent misclassification rate, and some users are worried about that. SAMHSA would like guidance on the advantages and disadvantages of structured lay interviews in comparison with semistructured lay interviews and with clinical interviews.

Canino said that if the goal is not only to produce prevalence rates but also to measure DSM disorders with substantial impairment, the choice of measures is more limited. Bose agreed that if a model is used one is not really measuring those concepts, but instead using a proxy. In these cases, face validity is very important. Canino added that in this sense the definition could be interpreted as meeting the criteria for a disorder or a probable disorder.

SMALL-AREA ESTIMATION OF PREVALENCE

Alan Zaslavsky (Harvard Medical School) discussed a small-area estimation project to estimate the prevalence of serious emotional disturbance among children in schools.[4] Small-area estimation is useful when sample sizes are inadequate to produce direct estimates. The project involved using a short scale to collect detailed domain data, and a limited amount of data collected based on a calibration survey with a validated longer instrument. Models were developed relating the instruments at the individual and school levels, and the relationships were used to build predictions from the short scale for school-level prevalence rates and individual-level screening.

One of the measures used for the study was the CIDI, which is used for the assessment of mental disorders and is administered by trained lay interviewers. The version used in this study was the adolescent version, which includes both a child interview and a parent questionnaire. As others mentioned in the first half of the workshop, whether parent reports or child reports are more accurate depends in part on the disorder. For example, parents can observe ADHD, while children may be unaware of it. The CIDI contains most of the information required to determine serious emotional disturbance. What is missing is an assessment of func-

[4] Li, F., and Zaslavsky, A.M. (2010). Using a short screening scale for small-area estimation of mental illness prevalence for schools. *Journal of the American Statistical Association, 105*(492), 1323-1332; Li, F., Green, J.G., Kessler, R.C., and Zaslavsky, A.M. (2010). Estimating prevalence of serious emotional disturbance in schools using a brief screening scale. *International Journal of Methods in Psychiatric Research*, June 19, Suppl. 1, 88-98.

tioning, but it can be imputed from the CIDI diagnosis and other items. In the study Zaslavsky described, this was done based on a model estimated from 347 clinical validation interviews, using the Children's Global Assessment Scale (CGAS), which were conducted as part of the study.

The other measure included in the study was the K6. The K6 can also be administered by lay interviewers or self-administered and includes six items focused on internalizing disorders. Because the scale is geared toward adults, it addresses the most common internalizing disorders among adults, such as depression and anxiety. In order to supplement the K6 for an adolescent population (ages 14-18), Zaslavsky said, the researchers considered 18 additional items for possible inclusion, including some screeners and some questions about symptoms of behavior disorders. These additional measures focused on disorders with earlier onset, and externalizing disorders, such as oppositional defiant disorder (ODD), conduct disorder, and ADHD. After testing, five items were selected to predict the disorders that are not very well predicted by the K6, and these included screeners for ADHD, ODD, and intermittent explosive disorder, as well as two personality items: "I can stay out of trouble" and "I have a strong temper."

Zaslavsky said that the predictive power of the K6 scale increased when the five items were added (K6+5). In the case of mood disorder and anxiety disorder, the area under the receiver operating characteristic curve increased slightly, from .77 to .81 for mood disorder and from .73 to .75 for anxiety disorder. However, in the case of any behavioral disorder, the increase was considerably larger, from .67 to .82. For serious emotional disturbance with an associated behavioral disorder, the increase was from .53 to .78.

Data collected in the National Comorbidity Survey Adolescent Supplement (NCS-A), discussed by Merikangas (see Chapter 2), was used to develop the methodology to estimate the small-area prevalence of serious emotional disturbance among schoolchildren in the United States. For this study, the school-based component of the NCS-A was used, which is based on a stratified national probability proportional to size sampling method. As a result of a highly complex model that allows for the replacement of schools that refused to participate, the number of schools included was 320. However, for this study, data from 282 schools were used, with at least 10 students per school. The number of adolescents included in the study was 9,244, which represented a 74.7 percent participation rate. The parent response rate, conditional on a child response, was 83.7 percent.

Zaslavsky explained that he and his colleagues fitted a Bayesian bivariate multilevel regression model with correlated effects for the probability of serious emotional disturbance and the augmented K6 (K6+5) score at the individual and school levels. For two continuous outcomes (Y_1, Y_2), the following two-level bivariate random effects model was assumed:

$$Y_{ijm} = \mathbf{X}_{ijm}\beta_m + v_{im} + e_{ijm}\ ,$$

where i is cluster (school, neighborhood, etc.); j is individual; m = 1, 2 = measure (K6 or CIDI); X is covariates; v is cluster-level random effect; and e is individual-level random effect. The random effects for the two measures are related at both individual and school levels: $(v_{i1}, v_{i2})' \sim N(0,\Sigma_v)$, and $(e_{ij1}, e_{ij2})' \sim N(0,\Sigma_e)$.

The outcome can be treated as a dichotomized outcome, in which case the Y_{ij2} is the latent variable that determines whether the dichotomized variable is 1 or 0 (whether serious emotional disturbance is present or not):

$$Y_{ij3} = I\left\{Y_{ij2} > 0\right\},$$

where Y_1 is the screener score, Y_2 is Φ(P(SED)), and Y_3 is the presence of serious emotional disturbance.

Alternatively, Zaslavsky said, it is possible to make use of the probability obtained from the CIDI as the outcome. Either of these approaches works in the model, but using a dichotomized outcome means throwing away some information in the data about which of the "noncases" are close to "caseness" and which ones are far from it.

The covariates available were age, sex, race, and ethnicity, as well as age of school entrance in order to distinguish children who entered school at an early age from those who entered when they were older. Information on the schools included whether it was a public or private school and the size (fewer than 50 teachers or 50 or more teachers).

Figure 4-3 illustrates the model relationships. The top level in the diagram shows the school-level effect, and it includes both the K6+5 and the serious emotional disturbance diagnostic procedure. They are related by a covariance matrix that provides information on how strongly they are correlated with each other. In other words, the school's mean level of children with a serious emotional disturbance and the school's mean level of K6+5 are related to each other, but they are not the same thing. Below that, there are the student-level effects, which are distinct for each individual: they are also related to each other, but distinct. Children who have a high score on the K6+5 will tend to have a high probability for serious emotional disturbance, but those are only related through a covariance. The covariate effects are shown on the sides of the diagram, Zaslavsky noted. The covariates that contribute to the responses of the student on the K6+5

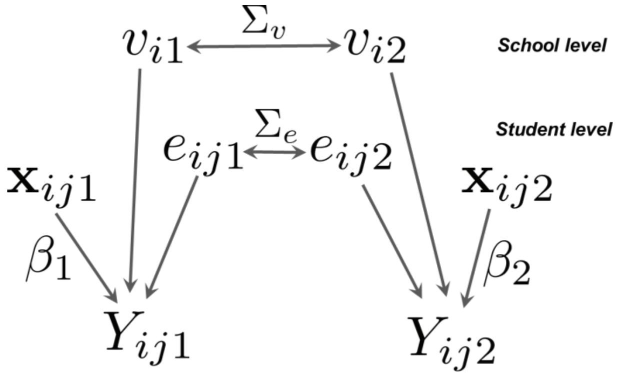


FIGURE 4-3 Model diagram for estimating serious emotional disturbance among children in schools.
NOTE: See text for definitions of variables.
SOURCE: Workshop presentation by Alan Zaslavsky, June 2015.

are on the left, and the covariates (possibly the same) that contribute to the responses on the diagnostic interview are on the right.

The coefficients and variance components in the model were estimated from the data that were available from the 282 schools, and Zaslavsky next discussed those data. Table 4-5 shows variances, using continuous outcomes. The left column shows the school-level variances, and those are variances on the scale of the 0 to 44 coding of the results from the K6+5 scale. The right column shows the individual-level variances for the same quantities.

The individual-level variances are 15 to 20 times larger than the school-level variances, which is not surprising, because the emotional status of different children in the same school can vary greatly, Zaslavsky said. The last line of the table shows the correlation between the K6+5 variation and the serious emotional disturbance variation, and the correlation for the school-level component is very high (.845). This correlation means that although the school's overall mean for serious emotional disturbance and K6+5 scores are not perfectly related to each other, they are strongly correlated, which is a measure of the validity of the K6+5 scale as an approximation or proxy for serious emotional disturbance. The correlation at the individual level is much smaller (.544), but still fairly substantial. Finding much stronger correlations at an aggregate level than at the individual level is not unusual in surveys, Zaslavsky noted.

Figures 4-4 through 4-6 show several different predictions. Figure 4-4 shows an out-of-sample prediction, which involves collecting data from

TABLE 4-5 Variances of K6+5 and Serious Emotional Disturbance Data at the School and Individual Levels

Variance	*v* (School Level)	*e* (Individual Level)
Variance for K6+5 (σ_1^2)	0.019	0.371
Variance for serious emotional disturbance (σ_2^2)	0.037	0.597
P (Correlation)	0.845	0.544

NOTE: See text for discussion.
SOURCE: Workshop presentation by Alan Zaslavsky, June 2015.

a subsample of students in the school and using those data to make a prediction for the remainder of the school. If K6+5 data are collected for a subsample of the students, it is possible to get an approximate inference for the school-level effect on the K6+5 scale, knowing that the mean of the individual-level effects tend to average out with large samples (as noted above, the correlation was .845). For another group of students who were not included in the survey, the prediction, v_{i2}, was estimated, shown at the top of the diagram. One source of error is introduced because the students who were surveyed using the K6+5 have some individual variation around the school mean, ē, but with a large enough sample size, that variation is reduced. An additional amount of variation originates from the fact that the K6+5 mean for the whole school and the serious emotional disturbance mean for the whole school are correlated but not perfectly related. And finally, Zaslavsky noted, there is some additional variation associated with the prediction for the children who were not in the original sample. The average of that variation will tend to be close to zero with large samples.

Figure 4-5 shows a naive individual-level prediction. If the school information was not available, it would be possible to use the individual K6+5 data and random effects to predict the serious emotional disturbance scores for the same children, ignoring the clustering. This scenario flattens out the model and lumps together school-level variation and individual-level variation within the school. The correlation across the top of the diagram is close to the .544 of the individual-level effects because most of the variation for a single individual is at the individual level. The school-level variation does not carry much weight. This might not work well, Zaslavsky said.

Figure 4-6 shows an in-sample prediction, which involves collecting the K6+5 data for a sample of children, estimating their individual- and

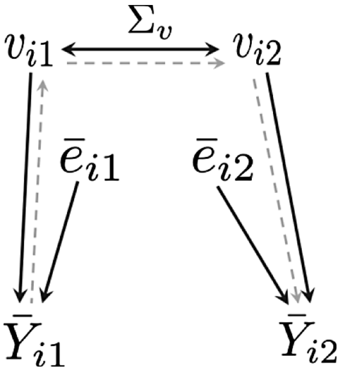


FIGURE 4-4 Out-of-sample prediction.
NOTE: This scenario involves collecting K6+5 measures in the school subsample and predicting serious emotional disturbance scores for the remainder of the school.
SOURCE: Workshop presentation by Alan Zaslavsky, June 2015.

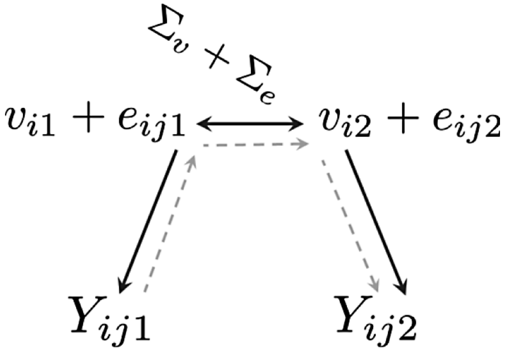


FIGURE 4-5 Individual-level prediction (naive).
NOTE: This scenario involves collecting K6+5 measures for individuals and predicting serious emotional disturbance scores for the same individuals; clustering is ignored.
SOURCE: Workshop presentation by Alan Zaslavsky, June 2015.

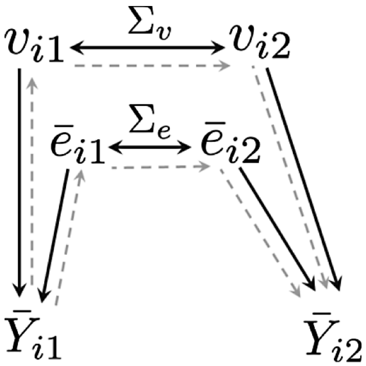


FIGURE 4-6 In-sample prediction.
NOTE: This scenario involves collecting K6+5 measures for a sample and predicting serious emotional disturbance scores for the same sample. The design with sampling within school combines in- and out-of-sample prediction.
SOURCE: Workshop presentation by Alan Zaslavsky, June 2015.

school-level mean effects on the K6+5, and then predicting serious emotional disturbance levels for the same sample of children based on the individual- and school-level effects on the serious emotional disturbance evaluation. In particular, this describes the situation in which every child in the school is administered the K6+5 scale and then serious emotional disturbance rates are estimated for the same group of children. This approach will enable a prediction of what the children's scores would have been on the diagnostic evaluation. Again, if the sample size is large enough, the mean individual-level variation will tend to be close to zero. The correlations will be mostly determined by the correlation between K6+5 and serious emotional disturbance at the school level, which was very high.

The prediction for individuals, that is, predicting children's serious emotional disturbance scores on the basis of their K6+5 scores, is a special case of in-sample prediction. However, the school-level random effect has a substantial effect on predictions for an individual, so it is helpful to make use of the information that a given child is in a school with a certain overall level of serious emotional disturbance, in addition to making use of the child›s K6+5 score. For example, a child with an average score on the K6+5 in a school that is one standard deviation above the median school-level random effect will have a probability of serious emotional disturbance of 12.7 percent; in a school that is one standard deviation below the median school-level random effect, the child will have a probability of serious emotional disturbance of 6.3 percent.

Zaslavsky discussed several limitations of the study. He acknowledged that there are measurement concerns related to the scales used, including the augmented K6, as well as the CIDI used as a proxy for serious emotional disturbance, which required some imputation of the clinical assessment of impairment. The use or development of other possible short scales could have been investigated. Furthermore, it is not clear whether these items would function in the same way in a different context. For example, the K6+5 was part of a lay interview that asked about a full range of DSM diagnoses, and it is possible that these measures would perform differently if they were included in a school health survey for which the context was not mental health.

Another limitation was that although the validation sample was fairly large in terms of these types of interviews, it was still not large enough to provide very precise estimates of the parameters of these models, particularly the variance/covariance parameters. More school-level covariates would have also been useful, especially if assignment policies and practices lead to higher rates of serious emotional disturbance in some schools.

There were also problems with the model fit at the high end, which Zaslavsky said has been seen in some of the other models. If the goal is to rank schools or even to just try to get a rough idea of which schools have the most children with needs, this may not be a very big problem. However, it could be an issue when estimating absolute prevalence rates.

Zaslavsky mentioned that he and his colleagues did another study in which they used neighborhood characteristics (poverty, homelessness, racial/ethnic composition, stability, religion, urbanicity) as covariates. This approach was more similar to that of some of the studies described by the other speakers, but the covariates available were weaker. The study also included some additional school-level data, based on a principal questionnaire, on items such as frequency of depression reports, attacks, and fights.

Zaslavsky concluded by saying that it is technically possible to combine a short screening scale with a calibration survey, and the calibration survey can be something that is practical to administer. The short screening scale could be integrated with the collection of school health information or some other activity, and it could even be done online. In the case of his study, the scale only included 11 items, so it would be feasible to carry it out for a large number of children, in a large number of schools.

In summary, Zaslavsky noted that there was substantial improvement relative to a purely synthetic model based on demographics. In addition, the multilevel model was also better than a single-level regression from the K6+5 mainly because the multilevel model adjusted the coefficients on the basis of the amount of data available for each school. This approach provides more efficient estimates when the amount of data available varies across schools. With the parameter estimates in this study, a sample of 100 children would be fairly close to what is needed to obtain the best possible information from a school, so this approach may be feasible to implement in most schools.

5

Key Themes and Possible Next Steps

This chapter summarizes the workshop's two discussion sessions, which followed the morning and afternoon presentations. The workshop chair, Kathleen Merikangas, noted several themes that had emerged from the talks on measurement. One of these is that a focus on the two domains, disorder and impairment, is really critical. However, the practice of collecting data on individual disorders with specific diagnostic criteria, which has characterized epidemiological research of mental disorders in the past couple of decades, is unlikely to help advance the goals that SAMHSA has laid out. A different approach suggested by Ian Hickie was that in community studies researchers could start with asking about impairment, on the basis of a definition of impairment that is orthogonal to symptoms. Once impairment is ascertained, information could be collected about the major classes or sources of those impairments. This approach would be a major shift from the current tradition that builds symptoms into impairment and infers causality, based on the DSM framework, which is also difficult to apply to general population studies. The shift would also mean focusing on the rates of impaired children, rather than rates of disorders, such as depression, anxiety, and attention deficit hyperactivity disorder (ADHD). The question of course, is related to implementation: What is the best way to do a survey that focuses on impairment?

Another theme, Merikangas noted, is the importance of informants. From the perspective of survey design and implementation, an immediate concern is the cost associated with obtaining data from as many as three informants (e.g., parent, child, teacher). The question is whether there are practical ways of reducing costs. There was also a lot of discussion of the instruments that could be used for measuring impairment. The emphasis was on differences between different developmental stages and how developmental norms, such as school readiness, are relevant. More work is needed to determine which instruments to use and on the optimal use of informants.

Merikangas said that the limitations of cross-sectional data were also discussed. Thinking in terms of 12-month impairment rates is very difficult. The lack of longitudinal studies also limits the understanding of the effectiveness of the medical care and services that are provided. Other countries do longitudinal studies, and it is a shame that this has not been done in the United States, on at least subsamples.

Another topic participants noted is the need to involve many different voices. Merikangas referred to the presentation by Peter Szatmari, about how stakeholders can be involved from the early stages in the development of a data collection to assure that the measures will work in a population study. Some of the groups to think about include epidemiologists, policy researchers, service providers, and even representatives of the public. This early inclusion of a wide range of stakeholders would be another big shift in thinking in the area of mental health data collections.

Jonaki Bose noted that there were two threads in the discussion. One is related to addressing SAMHSA's immediate need to provide state-level estimates of serious emotional disturbance in children, based on a specific definition. For example, if it is possible to develop a one-time model, assume relative stability over time, and identify some predictive variables that each state is likely to have, it might serve SAMHSA's needs well for the purposes of resource allocation. The other thread is a broader discussion of how to collect data that provide a more complete picture of child serious emotional disturbance beyond what is required by the legislation, and it might be helpful to distinguish between these two areas of discussion.

Dean Kilpatrick recalled that someone once said that mental illness was anything that was significantly disturbing to oneself or anyone else and brought a person to the attention of mental health professionals. This is a tongue-in-cheek definition, but it suggests that there are two things that are important to measure for any diagnosis. One is what might be considered distress, which could include any behaviors or problems that are disturbing to oneself. The other is what is disturbing to other people, which tends to be better captured by impairment. Impairment is indeed really important, but distress is important as well, and including it might be a good way of screening. In other words, if one is distressed, that information could be followed up with questions about why.

Susanna Visser said she wondered whether asking about impairment first would underestimate prevalence rates. She said that if children are appropriately treated, they might not show up in impairment assessments. For example, approximately 80 percent of the children with ADHD are medicated for the disorder. Furthermore, a comparison of the rates of agreement between children being treated for ADHD under Medicaid and children who meet diagnostic criteria showed that only one-third of the children who are taking medication for the disorder meet the diagnostic criteria.

In addition, Visser noted, most instruments do not provide instructions on whether to report symptoms and impairment on or off medications, and this is typically not accounted for in epidemiological studies. When she and her colleagues asked parents whether they were reporting about their children when on or off medication, half of the respondents said that they were reporting off medication or they were reporting based on a mix of what is observed on and off medication. Those who said that they were thinking primarily of the child on medication were asked whether they could report about symptoms and impairment off medication, and the researchers found that not everyone could. To some extent, ability to report off medication impairment depends on the disorder and type of medication that is used. For some externalizing disorders, the effects of taking the medication are more readily noticeable than for some of the internalizing disorders. Moreover, some parents cannot report about children's symptoms off medication because the children are on the medication all of the time.

Nora Cate Schaeffer (University of Wisconsin–Madison) asked whether impairment that is actually caused by medication is a concern and whether there is a way to also factor this in. Glorisa Canino replied that it depends on the definition of impairment and the measure used. Most impairment instruments do not ask about impairment due to symptoms or due to medications, instead focusing on the functioning of the child. She added that in her presentation she was advocating for an impairment measure that is not tied to symptoms, following the definition of the International Classification of Functioning for Children and Youth (ICF-CY). However, if the goal is to simply produce state-level prevalence estimates, it might be possible to use a shorter scale, such as the Strengths and Difficulties Questionnaire (SDQ), even if it does not meet all of the criteria in the definition in the legislation.

Canino also pointed out that the situation in which a large percentage of children with a disorder, such as ADHD, are being treated with medication is primarily characteristic of the United States, particularly in the white population. This is not the case in other countries, or among minority populations in the United States. In other countries, perhaps one in three or one in four children with serious emotional disturbance might be treated.

Hortensia Amaro (University of Southern California) commented that the emphasis on impairment is important from a policy perspective because it focuses on earlier interventions. This perspective could improve outcomes before children are at a stage when they have a positive diagnosis.

Ian Hickie offered a cautionary comment about the data available from administrative records, based on his experience with ADHD and depression data in Australia. Typically, the rates of treatment in these datasets are very high, with 80-90 percent of the people with a diagnosis receiving medication. It is likely that this is due to a diagnosis being assigned after a decision is made about treatment. These rates are very different from the rates that are obtained from population surveys, which Canino said are likely to show that 20-30 percent of the children with serious emotional disturbance are receiving treatment. Surveys also show that it is often children with the most severe impairment who are not getting treatment.

Another challenge pointed out by Hickie is that although many of the interventions improve function, they do not completely eliminate symptoms or impairment. In the case of ADHD, symptoms tend to greatly improve with treatment, but educational and social difficulties persist. Consequently, continuous measurement is useful, and a focus on impairment is important, he said. The symptom measure may have gone way down on the ADHD scale, but the impairment in the area of education may still be quite considerable. There are many developmental areas for which prescribing medication is not sufficient, and investments into educational and social supports are necessary in order to eliminate impairments.

Neil Russell said that in the context of SAMHSA's goals, the point made by Visser was very important, and the issue of how treatment affects a researcher's ability to measure prevalence should be investigated further. It would be a shame if the definition and measurement would lead to a large proportion of children being missed because they were being treated and asymptomatic or functioning better. From a policy perspective, the goal would be to distinguish between children who are impaired and children who are being treated but had been impaired prior to treatment. It is not clear how big of a methodological issue this is, but it may have a large impact on resource allocation. A Venn diagram would be particularly useful, showing how symptoms, function, and treatment map onto each other.

Merikangas said that this overlap issue is the reason that the National Comorbidity Survey Replication Adolescent Supplement (NCS-A) is focused on obtaining information about problems over the course of a lifetime. If the focus is limited to the present, a child might be on medication, or something else in the child's life might be affecting the immediate perception. This approach can also provide some information about the history of the problem, albeit the historical aspects of the data tend to be less reliable.

Benjamin Druss commented that several of the speakers mentioned that the prevalence rates are going to be different depending on what definitions are used. There was no discussion of whether workshop participants believe that the rates vary substantially over time, and if so, how quickly they are likely to change and why. This is one of the clinical and policy questions that needs to inform the methodological discussion, particularly about whether these data collections need to be carried out annually.

Merikangas replied that this question can be addressed separately for diagnosis and impairment. The term "surveillance" is sometimes used (based on infectious disease models) when the goal might be to monitor whether there are rapid changes. Using the same framework, sometimes a reporter calls and wants to know whether there is "an epidemic of depression." However, there is no evidence that the rates of many of the mental health disorders change substantially over time. The ups and downs are slight, she said.

The other question is whether impairment levels change as a function of more treatment, and there are good data to indicate that this is the case. This finding means that to the extent there is an interest in disorders, it may be possible to estimate classes of disorders more broadly than has been done in surveys, such as the NCS-A, which included operationalizations of all of the diagnostic criteria, questions about symptoms, and duration. Merikangas added that Alan Zaslavsky's research seems to indicate the same thing, which is that there are differences among the schools, but the differences are not necessarily attributable to the disorders. The impairment is absorbing much broader contextual influences. Merikangas added that she would want to move away from doing prevalence surveys in which the instruments are developed to tap specifically into the DSM and focus instead on getting an index of the disorder, severity, and impairment duration.

Michael Davern (NORC at the University of Chicago) noted that SAMHSA emphasized a need to produce prevalence estimates, and that there are some covariates that they would also like to have. He asked whether his understanding is correct and whether focusing on impairment can help address that mandate. Russell replied that SAMHSA does have a legislative mandate and a definition, but if there is a way to obtain these numbers using an approach that starts with impairment, then that would be useful to know. Canino added that previous discussions implied that a proxy of the DMS disorders might also be acceptable.

Merikangas said that she did not interpret the legislative mandate as a requirement to produce prevalence rates of each DSM disorder. Unless the focus is shifted from disorders to children, some children will be counted more than once, for example, if they have both depression and ADHD. Collecting information about a child's impairment before asking about the causes eliminates this problem and simplifies the process.

Furthermore, if the data about impairment and disorders are collected more efficiently, it would be possible to allocate more resources on a better understanding of regional differences or oversampling of some demographic groups in some years. This more in-depth data could lead to more informed decisions about treatment.

Bose asked for clarification on the assumption that prevalence rates might not change quickly enough to warrant annual data collections. Merikangas replied that if the question was the rate of impairment and finding out the cause of the impairment in each state, then perhaps annual surveys could be done. The question is whether it would be possible to move away from collecting comprehensive information on the full spectrum of DSM disorders. In other words, how much effort is worth investing into each of these, when research indicates that the rates remain rather stable for some of these conditions over years. Zaslavsky's approach appears to be based on similar reasoning, because he did not use the prevalence estimates from the NCS-A: he just combined all of the disorders to determine whether there is any disorder present or not. Bose agreed that even a one-time study that uses good predictors for modeling serious emotional disturbance would be an improvement over what is available now.

Kilpatrick agreed with the views that changes should not be expected from one year to another for the mental disorders discussed. Over time, more effective screenings and treatments should make a difference, as they did for more traditional health conditions, such as high blood pressure, but the rates should not be expected to change annually. However, even if SAMHSA concludes that annual data collections are not needed, it does not mean that a large, comprehensive baseline study is not justified. There are many relevant topics in the areas of both child and adult mental health that have not been measured or have not been measured well, such as exposure to potentially traumatic events and other stressors.

Szatmari also agreed that perhaps annual cross-sectional data collections are not needed, but he argued that there is a great need for longitudinal data and that following children for 6 months or 1 year would be very useful, because although the rates do not change, the children contributing to those rates change. He also noted that discussion and agreement are needed on what disorders need to be measured. In other words, what is needed in order to understand a child and what can be excluded. The basket could include important topics that have not been discussed, such as substance use, asthma, and epilepsy.

Hickie commented that one unintended consequence of the focus on prevalence is that once it becomes clear that these conditions do not change very much, the result is less frequent surveys because no differences are being observed. Cross-sectional impairment rates might also be stable, but the duration and the consequences of impairment could change dramatically as a result of better services. The duration of the impairment matters a lot, particularly in childhood, and, from a policy perspective, reducing that duration is very important.

Hickie said that one problem with perceptions about mental health is that diseases are not cured, but the reality is that interventions do change consequences dramatically. If the duration of impairments is shortened for children, their likelihood of success in education and their long-term success in employment increases substantially. However, one limitation associated with the impairment scales now available is that it is not known how they predict future functioning, in the crucial transition from school to higher education and employment. This further supports the need for frequent measurement, which can capture these types of issues and better reflect the benefits of services.

Graham Kalton reminded the participants that SAMHSA has a relatively straightforward mandate. While the data collection may be complicated to implement, the requirements are fairly clear. The discussions about how one could expand on the mandate are interesting and could contribute to the development of a comprehensive research agenda, but they also have cost implications, so the priority should be to consider what would satisfy the requirement to measure serious emotional disturbance and produce state-level estimates at some regular intervals. Replacing annual data collections with the longitudinal follow-ups would be more expensive, even if the number of surveys is the same, because longitudinal surveys have a distinct set of challenges.

Kalton added that one design option that may be worth considering and has not been discussed is to collect annual data on smaller samples than what is needed to produce the estimates and then average them over several years, as is done, for example, in the American Community Survey and the National Crime Victimization Survey. The advantage of this approach is that it does not alternate between a large data collection in one year and no data collection in other years. The operations and costs are spread out. Jeremy Aldworth noted that the approach of aggregating small chunks of sample over the years was used in the Mental Health Surveillance Study. An advantage of this method is that it is possible to make modifications over the course of the years as the data come in and new information becomes available.

Hickie said that the traditional methods tend to involve face-to-face interviewing, which is very expensive and also has many limitations. However, new technologies are more cost effective and could enable more frequent data collections. Kalton replied that those in the survey business have been working on developing methods that would enable better use of these new technologies, such as online data collections, but the question is whether there is a way to do these with a good representative sample, rather than with people who sign up and are essentially "professional survey takers." Hickie said that starting with a representative sample was important, but there are various ways for integrating the two methods, one of which was described in the presentation by Michael Kogan.

Merikangas asked Canino to share her thoughts on the discussions that were centered around impairment throughout the workshop and the possibility of not collecting full DSM diagnoses every year, while still being able to provide estimates that meet the criteria of child serious emotional disturbance. Canino said that based on the mandate, what needs to be measured is DSM disorders and impairment. Although the issues surrounding the measurement of DSM disorders have been generally resolved over the years, the measurement of impairment remains a challenge. Impairment is important for determining treatment and for predicting service use, and it is amazing that so much effort has been invested into developing a classification system for disorders and measures of disorders and so little effort has been invested into discussing what is most relevant for the families that are affected, which is the functioning of their children. The ultimate goal with any treatment is to see an improvement in a child's functioning.

Canino recalled the perspective of child psychiatrist Michael Rutter on the difference between impairment in children and impairment in adults. Rutter said that if a child has a disorder and as a consequence of that the child drops out of school, he (or she) might not be able to get a job and might have difficulties for the rest of his life, even if he gets better or no longer has the disorder. In other words, the consequences of the childhood disorder are much more severe and are much longer lasting than those for adults. Because of that, it is crucial to get the measurement right.

When the Columbia Impairment Scale was developed, the youth version of the International Classification of Functioning did not exist. Canino, who worked with Hector Bird (Columbia University) on its development, noted that the definition used for youth was much simpler than any for adults, limited to a decrease in function. It was understood that functioning was a continuum that went from high functioning to low functioning and impairment. It was also understood that impairment was context related. However, the conceptualization of impairment in the ICF-CY is not as clear: it refers to a decrease or deterioration in the functioning of a child that is expressed as a consequence of a health condition but is separate from the health condition. The ICF-CY also acknowledges that the disability can be manifested both at an individual level and at a societal level. Furthermore, it takes into consideration the developmental level of the child, which is extremely important because being impaired at the age of 3, when the child is in a family context and the impairments are related to that context, is different from being impaired as an adolescent. The instrument that is used or developed has to be one that considers all of these aspects. It is also important to note that this definition takes into consideration both mental and physical disorders to understand the construct of impairment for an individual.

Canino noted that the ICF-CY also discusses substantial impairment. This is important because for the purposes of resource allocation, the definition has to be limited in scope. Hickie argued that the opposite should be happening, with more focus on providing services to children who are not severely impaired yet, which would be ideal. Perhaps that is realistic in Australia, Canino said, but in the United States the discussion has not focused on that concept as yet.

Bose commented that if data are collected on both impairment and disorder, and there are people who have the beginning of impairments and perhaps no diagnosis of disorder, it would be possible to also conduct analyses focused on those groups, regardless of what the definition is for funding purposes. Canino agreed that there would be great advantages to being able to conduct those types of analyses. She added that she concurs with the perspective that impairment comes first and that it is not always a consequence of a disorder.

The instrument that comes closest in Canino's view to meeting the current goals of measuring impairment is the World Health Organization Disability Assessment Schedule (WHO-DAS) for children. Canino said that the child WHO-DAS still requires psychometric validation, but it is working in Nigeria, which is an indication that it could work in the United States. She added that it is important to conduct the analyses for the United States, and SAMHSA could commission such work. The first step would be to analyze the data already available on adolescents. If that looks promising, the next step would be to conduct research to determine how it works with younger children, using the instrument that has been adapted from the adult WHO-DAS. Research would also be useful to develop severity scores or cutoff scores for various instruments and to reduce the length of the instruments.

Visser said that the issue of differential psychometrics as a function of ethnicity and race would also need to be the subject of research. At the moment, it does not appear that race and ethnicity could be added to a model, and that is a limitation. For example, there is a large percentage of Puerto Ricans in New York and a large percentage of Mexican Americans in Texas, and those two groups have different health-seeking behaviors, as well as different rates of reporting symptoms and impairment. If state is added to the model, these differences could become a problem. The goal would be to find an impairment measure that has strong psychometrics properties within racial and ethnic subgroups because there will be no other way to account for these types of differences if a modeling technique is used.

Merikangas added that another research project that SAMHSA will need to undertake is to evaluate the role of informants by developmental level. She asked Canino to clarify how one would measure a decrease in functioning. Is this an area where longitudinal data would be most useful? Canino replied that parents can be asked how their children's functioning compares with that in the previous year.

Davern said that some of the speakers made a convincing case from a research perspective that prioritizing impairment would be better. If the goal is to use these estimates for funding allocations, then one result of focusing on the prevalence of impairment rather than the prevalence of disorder could mean that states that are successful at reducing impairment would get less funding than they would need on the basis of the prevalence of disorders over time. This potential result could also be an issue for a modeling approach, depending on whether state is included in the model.

Visser noted that the states are very experienced with these types of challenges and have creative ways of addressing them. Many of the contracts they fund are performance based. Based on the experience in other areas, such as infant mortality, as the focus shifts to the final few percent, it is recognized that those are more difficult to address and more funding is needed.

Kalton said that, in general, it may be premature to think about specific methods, such as modeling. It is important to first think through the overall design. One of the first questions is whether it would be feasible to have a good instrument that could be administered to children on an annual basis, possibly on the internet, with the purpose of measuring the relative levels of serious emotional disturbance across the states. If this is possible without the need for follow-up clinical interviews, it would be the most important parameter to establish first. Including clinical interviews is expensive, even if there is modeling involved, and it would be difficult to do on a reasonable scale. Perhaps what would be useful is an instrument that could be administered regularly and would only be subjected to some validation work occasionally and then modified on the basis of that validation work, as needed.

Kalton added that there are a number of other smaller issues that deserve further discussion. One is that the current definition seems to include a criterion of sufficient duration and a reference to the past year. It is not clear how this would fit with the available scales, which seem to have varying reference periods. Another issue is the role of comorbidity, which came up in several presentations, but it is not yet completely clear as to how comorbidity and impairment interrelate and what the consequences of such interrelationships might be.

In addition, Kalton said, it will be important to get a good handle on the implications of the child's age and what is feasible to do with different age groups. In particular, what can be done with those under the age of 6, and what can be done with children in the 0-3 age group? Furthermore, how important is it to have data from three different informants—the child, the parent, and the teacher—and would it be possible to develop an approach that does not require that? Some of these activities would be enormously costly and some might be mutually exclusive, such as combining an internet data collection with a complicated design that requires multiple informants.

When grappling with these ideas, Kalton said, perhaps the fitness-for-use perspective that was mentioned in the presentations would be useful to keep in mind. Perhaps the goal is not perfection, but to identify a measure that is reasonable and acceptable and charts the variation across states so that funding allocations can be based on it. Canino mentioned that there is no consensus on where to place impairment cut points. Perhaps this does not matter if the goal is to allocate a fixed amount of funding across states. But if the overall amount of funding available depends on the rates, then that is a different scenario.

Canino addressed Kalton's point about the number of informants by saying that for pragmatic purposes, such as the data collection that SAMHSA may be undertaking, reasonable decisions can be made on the basis of the age of the child. For example, for children under 12, she would only do parent interviews. For children 13 and older, she would collect data from the child only. For specific purposes—for example, if a good measure of ADHD is needed—she would include the child and the teacher. This approach would not meet the requirement for obtaining research funding from the National Institute of Mental Health (NIMH) because NIMH reviewers would expect all three informants to be included, but it would be a practical solution for SAMHSA.

Following up on Kalton's question about whether the "gold standard" clinical interview would always need to be included, Bose asked the participants for their views on a hypothetical design that would be based on a one-time cross-sectional study that included the Composite International Diagnostic Interview (CIDI, which is administered by lay interviewers), and also included potential predictor variables, such as the full SDQ, a short scale that Goodman has tested for his purposes, or some other set of predictors. What would be the flaws of a model for which the outcome would be serious emotional disturbance based on the CIDI score for the age group that the CIDI covers and the predictor variables? Would a design along these lines pass the laugh test? Also, what sample size would be required? If data were available for 10,000 children who completed the CIDI, would it be possible to develop a predictive model? And what sample size would be sufficient for calibration? Typically, it appears that very small sample sizes, such as 10 percent or less, are used for calibration, but would it be adequate to administer a clinical version to 40 cases for each disorder?

Merikangas said that a similar approach was used as part of the NCS-A in about four sites. The test included oversamples of children with certain disorders that were more difficult to measure, such as mania. A lot of work was done very inexpensively on a small sample, which was later used to calibrate the broader instrument for larger samples. A recent Canadian study also used a similar approach, but the data from the study were not yet available for this workshop.

Steven Heeringa noted that Ronald Kessler and his colleagues have used this calibration approach for the CIDI-SC mental health diagnoses in the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS).[1] The samples might be 400 or 500, and there might be 40 positive screens for each diagnosis. That particular calibration, represented by the calibration from Y to Y^* in his model, will have an extreme amount of variability. Heeringa said that, in his understanding, the way some of the other studies have used this approach was to determine what criteria in the Y variable should be used to produce the most reliable estimate, that is, as a projection of true value, rather than to develop an adjustment model. In the case of the study described by Aldworth, Heeringa commented that the standard errors of the coefficients in the logit model and, therefore, the variance of any predictions from that model are going to be a function of the strength of the relationships, as well as the relative

[1] Kessler, R.C., Santiago, P.N., Colpe, L.J., Dempsey, C.L., First, M.B., Heeringa, S.G., Stein, M.B., Fullerton, C.S., Gruber, M.J., Naifeh, J.A., Nock, M.K., Sampson, N.A., Schoenbaum, M., Zaslavsky, A.M., and Ursano, R.J. (2013). Clinical reappraisal of the Composite International Diagnostic Interview Screening Scales (CIDI-SC) in the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS). *International Journal of Methods in Psychiatric Research*, *22*(4), 303-321.

sample sizes in Step 1 and Step 2 (see Chapter 4). Adjusting 10,000 cases on a model fitted to 40 cases might be difficult.

The Aging Demographics and Memory Study (ADAMS) described by Heeringa included dementia assessments that cost approximately $1,000 per person for the sample members that were in the *Y* measurement. The ratio was 857 assessments and about 8,000 to 9,000 cases, and there was no guarantee that it would produce a model that functioned as well as it did. But the researchers had very good initial information to be able to stratify the Step 2 sample and improve the efficiency of estimating the model. Because of the amount of information that was available for the stratification, the precision of the estimates of prevalence was high even from direct estimation, and it was much better than would have been possible on the basis of a naïve sample of a similar size. For the next ADAMS data collection, the researchers are not going to include the assessments because they have concluded that they have enough information on that particular step. Heeringa and his colleagues were able to refine the screening items to obtain these predictions of dementia through that step.

Kilpatrick commented that if the ultimate purpose is to establish the relative numbers of cases per state for the allocation of funding, then even if the lack of clinical calibration resulted in elevated prevalence rates—and it is not clear that it would—the relative burden per state would presumably still be correct. He also wondered whether more of the state- or substate-level data available from the U.S. Census Bureau could be used for modeling.

Schaffer noted that there was a lot of discussion of gold standards, and perhaps interviewer effects deserve more attention. When diagnostic-style interviews are included, there is probably only a small number of trained interviewers, and it is likely that there is a high interviewer variance. It is worth thinking about how to take this into account even if there are not enough interviewers to estimate it. At the minimum, attention should be paid to how the cases are assigned so that a few great interviewers do not end up dominating an entire dataset.

Visser described one of her ongoing studies with colleagues as a context for some cost considerations. Their Project to Learn about ADHD and Youth Mental Health (PLAY-MH) study is a multistage design that includes a minimum of 5,000 children at each study site. The sites receive $250,000 each year for 2 years to do screenings of the entire school district and then pull out a sample. On the basis of simulations, the researchers determined that they needed 650 diagnostic interviews to be able to estimate prevalence of 2 percent or greater. Getting those 650 interviews has not been easy, even with incentives for the families, the schools, and even the teachers who are conducting the screenings ($4 per screen). If SAMHSA is considering some school-based data collections, in a few sites or even nationally, a partnership with the Centers for Disease Control and Prevention might be useful to leverage mutual resources.

Russell asked for clarification on the idea of starting the measurement with impairment. How would the disorders be ascertained in this case and how does one ensure that the definition of child serious emotional disturbance is being met? At the minimum, SAMHSA would need to be able to say that the process resulted in a fairly decent proxy measure. Merikangas replied that the main difference would be the different order, but some disorder information would still be collected. The exact data that would go into the basket and the process would need to be decided. Hickie agreed that the disorders would not be skipped and clarified that according to current methods, one has to have the disorder in order for an impairment to be obtained. The idea would be to separate the two and allow for impairment information to emerge, independent of the disorder. It would still be possible to obtain the prevalence rate, but impairment data would not be limited to those with a disorder.

Kalton reminded the group about the discussion related to the role of treatment and that it will be important to capture the disorder information even if a child is successfully being treated and the impairment has been reduced or eliminated. Hickie agreed that the disorder questions should not be skipped, regardless of the level of impairment. Visser cautioned that the issue of how exactly the disorder information will be captured deserves very careful consideration, especially if the intent is to do model-based estimation.

Appendix A

Workshop Agenda

WORKSHOP ON INTEGRATING NEW MEASURES OF SERIOUS EMOTIONAL DISTURBANCE IN CHILDREN INTO SAMHSA'S DATA COLLECTION PROGRAMS

National Academy of Sciences
Keck Center, Room 201
500 Fifth Street, NW
Washington, DC 20001
June 11, 2015

8:30–8:45 **WELCOME AND INTRODUCTIONS**

Kathleen Ries Merikangas, *Workshop Chair, National Institute of Mental Health*
Constance Citro, *Director, Committee on National Statistics*

8:45–9:00 **SAMHSA'S GOALS AND CHALLENGES RELATED TO MEASURING SERIOUS EMOTIONAL DISTURBANCE IN CHILDREN**

Neil Russell, *Director, Division of Surveillance and Data Collection, Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration*

9:00–10:30 **OVERVIEW OF EXISTING DATA ON SED IN CHILDREN**

U.S. National Data on Prevalence of Mental Disorders

Kathleen Ries Merikangas, *National Institute of Mental Health*

The Need for Measuring Impairment in Functioning for Assessing SED
Glorisa Canino, *University of Puerto Rico*

Ongoing Federal Child Mental Health Surveillance Systems
Susanna Visser, *Centers for Disease Control and Prevention*

Trends in Mental Health Impairment and Treatment in Children and Adolescents
Benjamin Druss, *Emory University*

10:30–10:40 **COFFEE BREAK**

10:40–12:10 **INTERNATIONAL PERSPECTIVES ON CHALLENGES RELATED TO MEASURING CHILD SED IN POPULATION SURVEYS**

Diagnostic and Impairment Issues: Behavior and Developmental Disorders
Peter Szatmari, *University of Toronto, Canada*

Diagnostic and Impairment Issues: Mood and Anxiety Disorders
Ian Hickie, *University of Sydney, Australia*

Discussion of the Strengths and Difficulties Questionnaire
Robert Goodman, *King's College, London*

12:10–12:30 **FLOOR DISCUSSION**
Kathleen Ries Merikangas, *Workshop Chair, National Institute of Mental Health*

12:30–1:30 **LUNCH TO CONTINUE MORNING DISCUSSION**
Third Floor Atrium

1:30–3:40 **DESIGN AND ESTIMATION OPTIONS FOR MEASURING CHILD SERIOUS EMOTIONAL DISTURBANCE**

The National Survey of Children's Health and Behavioral Health Measures
Michael Kogan, *Maternal and Child Health Bureau, Health Resources and Services Administration*

Design and Estimation Considerations in Multiphase Studies
Steven Heeringa, *University of Michigan*

Estimating the Prevalence of Child SED within a National Survey: Pilot Study Experiences
Heather Ringeisen and Jeremy Aldworth, *RTI*

Small-Area Estimation of Prevalence of SED in Schools
Alan Zaslavsky, *Harvard University*

3:40–4:00 **COFFEE BREAK**

4:00–5:30 **DISCUSSION OF THE MEASUREMENT AND ESTIMATION OPTIONS**
Kathleen Ries Merikangas, *National Institute of Mental Health*
Glorisa Canino, *University of Puerto Rico*
Michael Davern, *NORC at the University of Chicago*
Graham Kalton, *Westat*

Appendix B

Biographical Sketches of Steering Committee Members and Speakers

JEREMY ALDWORTH (*Speaker*) is senior research statistician at RTI International where he works on the National Survey on Drug Use and Health for the Substance Abuse and Mental Health Services Administration. He previously worked at Biogen Idec, Eli Lilly and Company, and the Institute for Commercial Forestry Research. He has a Ph.D. in statistics from Iowa State University**.**

GLORISA CANINO (*Member, Steering Committee*) is director of the Behavioral Sciences Research Institute at the School of Medicine at the University of Puerto Rico. Her research is focused on cross-cultural child and adult psychiatric epidemiology; service utilization patterns and barriers to care faced by Latino children and adults; and instrument psychometrics, particularly as they relate to the adaptation and translation of instruments to the Latino culture. She is a recipient of the Rema Lapouse Award of the American Public Health Association, which is given to outstanding scientists in the area of psychiatric epidemiology. She has a Ph.D. in psychology from Temple University.

MICHAEL DAVERN (*Member, Steering Committee*) is senior vice president and director of the Public Health Research Department at NORC at the University of Chicago. Previously, he was an assistant professor of health policy and management and research director of the State Health Access Data Assistance Center at the University of Minnesota School of Public Health. His expertise includes survey research, public health data, linking surveys with administrative data and Census Bureau data, and the use of these data for policy research, simulation, and evaluation. He has a Ph.D. in sociology from the University of Notre Dame.

BENJAMIN DRUSS (*Speaker*) is a professor and Rosalynn Carter chair in mental health in the Department of Health Policy and Management at the Rollins School of Public Health at Emory University, where he also directs the Center for Behavioral Health Policy Studies. His research focuses on improving physical health and health care among people with serious mental disorders. He has received a number of national awards for his work, including the Health Services Research Senior Scholar Award from the American Psychiatric Association and the Armin Loeb Award from the Psychiatric Rehabilitation Association. He has an M.P.H. from Yale University and an M.D. from New York University.

ROBERT GOODMAN (*Speaker*) is professor of brain and behavioral medicine in the Child and Adolescent Mental Health Clinical Academic Group and the Department of Child and Adolescent Psychiatry at King's College in London. He is the developer of the Strengths and Difficulties Questionnaire and the Development and Well-Being Assessment. His research interests include the psychiatric consequences of chronic neurodevelopmental disorders. He has a Ph.D. degree, is a member of the Royal College of Physicians, and was awarded a fellowship of the Royal College of Psychiatrists.

STEVEN HEERINGA (*Speaker*) is a research affiliate at the Population Studies Center and a senior research scientist at the Survey Research Center, both at the University of Michigan. He is a member of the faculty of the University of Michigan Program in Survey Methods and the Joint Program in Survey Methodology. He has contributed as a consulting statistician to a number of international research programs and ongoing survey data collection projects, including the Army Study to Assess Risk and Resilience in Servicemembers, the Health and Retirement Study, the Monitoring the Future Study, the National Study of American Life, and the South African Stress and Health Study. He has a Ph.D. in biostatistics from the University of Michigan.

IAN HICKIE (*Speaker*) is a senior principal research fellow of the National Health and Medical Research Council of Australia, a professor of psychiatry at Sydney Medical School, and the executive director of the Brain & Mind Research Institute at the University of Sydney. He was one of Australia's first national mental health commissioners. He is particularly interested in youth mental health and the prevention of and early intervention in emerging mood disorders, including optimizing treatments for young people with emerging mood disorders. He is a fellow of the Academy of Social Sciences in Australia and of the Royal Australian & New Zealand College of Psychiatrists and a member of the Order of Australia. He has a doctorate of medicine from the University of New South Wales.

GRAHAM KALTON (*Member, Steering Committee*) is chair of the board and senior vice president of Westat. Previously, he held positions at the University of Maryland, the University of Michigan, the University of Southampton, and the London School of Economics. His research interests are in survey sampling and general survey methodology. He was a cofounder of the Joint Program in Survey Methodology at the University of Maryland. He is a past president of the International Association of Survey Statisticians, past chair of the American Statistical Association's section on survey research methods and the Royal Statistical Society's social statistics section, and a member of the Council of the Royal Statistical Society. He is also a fellow of the American Statistical Association and a national associate of the National Academies. He has a Ph.D. in survey statistics from the University of Southampton.

MICHAEL KOGAN (*Speaker*) is director of the Office of Epidemiology, Policy and Evaluation for the Maternal and Child Health Bureau at the Health Resources and Services Administration and project director of the National Survey of Children's Health and the National Survey of Children with Special Health Care Needs. Previously, he was a senior epidemiologist with the Centers for Disease Control and Prevention and National Center for Health Statistics. He has also held adjunct academic appointments at the University of Alabama at Birmingham and Harvard University and is a regular lecturer at Georgetown University. His research focuses on pediatric and perinatal epidemiology, over-the-counter medication use among children; racial and ethnic disparities in birth outcomes; the consequences of underinsurance; and both the prevalence of autism spectrum disorder (ASD) and the health care experiences of families with ASD children. He is a recipient of the Advancing Knowledge Award from the Coalition for Excellence in Maternal and Child Health Epidemiology and of the Administrator's Award for Excellence from the Health Resources and Services Administration. He has a Ph.D. in epidemiology from Yale University.

KATHLEEN RIES MERIKANGAS *(Chair, Steering Committee)* is senior investigator and chief of the Genetic Epidemiology Research Branch in the Intramural Research Program at the National Institute of Mental Health. Previously, she was a professor of epidemiology and public health, psychiatry, and psychology and director of the genetic epidemiology research unit in the Department of Epidemiology and Public Health at Yale University. Her research focuses on studies of the patterns and components of familial aggregation of mental disorders and familial mechanisms for comorbidity of mental and medical disorders; identification of early signs and risk factors for psychiatric disorders among high- and low-risk youth using prospective longitudinal high-risk studies; and large-scale population-based studies of mental disorders, including high-risk designs and prospective longitudinal research. She has a Ph.D. in chronic disease epidemiology from the University of Pittsburgh.

HEATHER RINGEISEN (*Speaker*) is director of the Children and Families Research Program at RTI International. She serves as a co-investigator on the National Survey of Child and Adolescent Well-Being II and as the project director of a calibration study to determine an estimate of childhood serious emotional disturbance within the National Health Interview Survey. Her research focuses on children's mental health services research, with an interest in the nonspecialty mental health service systems. Previously, she served as chief of the Child and Adolescent Services Research Program at the National Institute of Mental Health, where she directed a research program that examined the quality, organization, and financing of services for children with mental disorders. She is a recipient of the policy fellowship from the Society for Research in Child Development. She has a Ph.D. in clinical child psychology from Auburn University and is a licensed clinical child psychologist.

NEIL RUSSELL (*Speaker*) is director of the Division of Surveillance and Data Collection in the Center for Behavioral Health Statistics and Quality at the Substance Abuse and Mental Health Services Administration. His areas of expertise include behavioral health statistics and epidemiology; basic and applied research in behavioral health data systems and statistical methodology; as well as surveillance and data collection. He has a Ph.D. in sociology from Arizona State University with a focus in survey research.

PETER SZATMARI (*Speaker*) is chief of the Child and Youth Mental Health Collaborative at the Centre for Addiction and Mental Health and the Hospital for Sick Children in Toronto, as well as director of the Division of Child and Youth Mental Health at the University of Toronto. His work has been in the field of autism spectrum disorder (ASD), including studying the longitudinal course of this disorder and its genetic causes. He is the founding director of the Canadian Autism Intervention Research Network, a patient-oriented research network in early intervention in

ASD. He has also consulted with government agencies in Canada, the United States, and the United Kingdom. He has an M.D. and an M.S. degree from McMaster University.

SUSANNA VISSER ***(Speaker)*** is the acting associate director of science for the Division of Human Development and Disability of the National Center on Birth Defects and Developmental Disabilities at the U.S. Department of Health and Human Services Centers for Disease Control and Prevention. Her work focuses on the epidemiologic study of neurobehavioral and mental health conditions of children, including ADHD, Tourette syndrome, and autism. Her expertise includes the analysis of longitudinal survey data covering developmental trends across the life span and population-based survey data. She currently serves as the committee epidemiologist for the American Academy of Pediatrics ADHD diagnostic and treatment guidelines committee and participates in technical expert panels for two national surveys sponsored by the Maternal and Child Health Bureau. She has a Ph.D. from the School of Public Health of the University of Illinois at Chicago.

ALAN ZASLAVSKY (*Speaker*) is professor of health care policy (statistics) in the Department of Health Care Policy at Harvard Medical School. His methodological research interests include surveys, census methodology, microsimulation models, missing data, hierarchical modeling, small-area estimation, and applied Bayesian methodology. His health services research focuses primarily on developing methodology for quality measurement of health plans and providers and understanding the implications of these quality measurements. He also works on analyses for the World Mental Health Surveys and for the Study to Assess Risk and Resilience in Servicemembers of the U.S. Army. His interests include methodology for measuring racial and ethnic disparities in care and determining their causes, quality measurement for pediatric hospital care, and national opinion research on health policy issues. He is a fellow of the American Statistical Association, an elected member of the International Statistical Institute, and a national associate of the National Academies of Sciences, Engineering, and Medicine. He has a Ph.D. in applied mathematics from the Massachusetts Institute of Technology.