AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Psychopathology and Clinical Science

© 2023 American Psychological Association
ISSN: 2769-7541

https://doi.org/10.1037/abn0000805

Quantifying Skip-Out Information Loss When Assessing Major Depression Symptoms

Orla McBride[1], Jelle van Bezooijen[2], Steven H. Aggen[3, 4], Kenneth S. Kendler[3, 4], and Eiko I. Fried[2]

[1] School of Psychology, Ulster University
[2] Department of Psychology, Unit Clinical Psychology, Leiden University
[3] Virginia Institute for Psychiatric and Behavioral Genetics
[4] Department of Psychiatry, Virginia Commonwealth University School of Medicine

Large-scale mental health surveys screen participants for the presence of the core diagnostic criteria of a mental disorder such as major depressive disorder (MDD). Only participants who screen positive are administered the full diagnostic module; the remainder "skip-out." Although this procedure adheres faithfully to the psychiatric classification of mental disorders, it limits the use of the resulting survey data for conducting high-quality research of importance to scientists, clinicians, and policymakers. Here, we conduct a series of exploratory analyses using the Virginia Adult Twin Study of Psychiatric and Substance Use Disorders (VATSPSUD) data, a unique survey which suspended the skip-out procedure for assessing past-year MDD. Adult twins (N = 8,980) born between 1930 and 1974 were recruited from a multiple-birth record database established in 1980 and interviewed in mid-adulthood between 1987 and 1996. We compared the: (a) prevalence and levels of impairment of the diagnostic criteria (and disaggregated symptom items) of adults screening positive/negative and (b) patterns of associations between MDD diagnostic criteria (and disaggregated symptom items) under three conditions: (a) full data; (b) "skip-out" data substituted with zeros; and (c) "skip-out" data treated via listwise deletion. Important differences in the patterns of associations between diagnostic criteria and disaggregated symptom sets emerged which changed the statistical evidence regarding the dimensionality of the criteria/symptom items (i.e., Condition C). An ill-defined correlation matrix which was unsuitable for statistical analysis was produced (i.e., Condition B). Given the problems with these widely used approaches, we offer researchers and data analysts practical alternatives to using the skip-out procedure in future surveys.

General Scientific Summary

The use of "skip-out" procedures in diagnostic modules in national mental health surveys results in substantial proportions of missing data, causing theoretical and methodological challenges. Researchers need to think carefully about how best to treat this missing data because common methods to overcome this issue (e.g., complete case analyses, substitution with zero values) cause difficulties when analyzing this data using traditional statistical models.

Keywords: DSM, major depressive disorder, survey, screener questions, skip-out

Supplemental materials: https://doi.org/10.1037/abn0000805

For decades, large-scale psychiatric epidemiological surveys such as the Epidemiological Catchment Area (ECA) study (Klerman, 1986), the National Comorbidity Survey (NCS) (Kessler, 1994), and the National Epidemiological Surveys on Alcohol and Related Conditions (NESARC) (Hasin & Grant, 2015), have provided robust evidence as to the prevalence of, and the rate of treatment for, mental

Orla McBride https://orcid.org/0000-0003-3399-9466

An early draft of this work was presented at an internal research seminar at the School of Psychology, Ulster University, in 2020 and at the University of Oxford in 2021.

The VATSPSUD received ethical approval from the Institutional Review Board at Virginia Commonwealth University in accordance with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. The VATSPSUD was conducted between 1987 and 1996. Study participants were not asked for, and did not provide, their consent at the point of recruitment for the sharing of their sensitive data in line with current day open access practices. Study data, materials, and code are not available for sharing. This study was not preregistered.

We have no known conflict of interest to disclose.

Orla McBride served as lead for formal analysis, methodology, and writing–original draft. Jelle van Bezooijen served as methodology, visualization, writing–review and editing. Steven H. Aggen served as conceptualization, visualization, formal analysis, resources, data curation, writing–reviewing and editing. Kenneth S. Kendler served as supervision, methodology, investigation, conceptualization, writing–reviewing and editing. Eiko Fried served as conceptualisation, methodology, supervision, writing–review and editing.

Correspondence concerning this article should be addressed to Orla McBride, School of Psychology, Ulster University, Cromore Road, Coleraine, BT52 1SA, United Kingdom. Email: o.mcbride@ulster.ac.uk

disorders in the U.S. adult general population. A key design feature of these surveys is the administration of a validated diagnostic interview schedule to collect data on experiences of mental disorders as specified in psychiatric classification systems, such as the DSM (Robins & Cottler, 2004). Trained interviewers not only ask adults about their experiences of the clinical features of mental disorders, such as *depressed mood* and *anhedonia* in the case of MDD, but also assess for functional impairment and/or the presence of nonpsychiatric medical conditions that can have similar symptoms (Steinberg, 1994). This approach, whilst resource-intensive, is considered the "gold standard" in psychiatric epidemiological research, particularly when the main goal of the survey is to establish robust prevalence estimates of mental disorders in the general population (Kessler & Üstün, 2004). The superiority of this survey design has been highlighted by research which has demonstrated that substituting interviewer-led diagnostic interviews with self-report questionnaires for MDD (such as the Patient Health Questionnaire, PHQ-9) in large-scale surveys can result in overestimates of the prevalence of MDD by as much as 11.5% (Levis et al., 2020).

With respect to MDD, adults participating in surveys using diagnostic interviews are typically asked about their experiences of the core MDD criteria (i.e., *depressed mood* and *anhedonia*) with reference to a specific time frame (e.g., every day or nearly every day during a two-week period, or longer, in the past year), at the beginning of the diagnostic module (Spitzer et al., 1992). Adults endorsing neither episodes of *depressed mood* nor of *anhedonia* skip-out of the diagnostic module; the remainder are interviewed about their experiences of symptoms which operationalize seven additional diagnostic criteria for MDD (i.e., *weight/appetite changes*, *sleep disturbances*, *psychomotor changes*, *fatigue*, *guilt/worthlessness*, *concentration difficulties*, and *suicidal ideation*) during the same two-week (or longer) period when they experienced *depressed mood* and/or *anhedonia* (American Psychiatric Association, 1994). This information is then used to determine whether the clinical diagnostic threshold for MDD has been met. This approach adheres faithfully to the DSM nomenclature, which considers MDD as an episodic disorder characterized by temporally co-occurring symptoms that have an implied causal relationship with each other, and that the additional diagnostic criteria are not counted or considered unless they are part of the temporal and quasi-causal cluster of symptoms (American Psychiatric Association, 1994; Kendler et al., 2010). It also assumes that experiences of the additional diagnostic criteria in the absence of *depressed mood* or *anhedonia* are not necessarily indicators of MDD (e.g., *weight/appetite changes* due to overindulgence during a festive season; *fatigue* due to over-working; *sleep disturbances* due to childcare responsibilities, etc.) (Gruenberg et al., 2005).

From this perspective, imposing conditionality or a skip-out procedure in the diagnostic module has some justification because it (a) implements and operationalizes the clinically established hierarchy of the MDD diagnostic criteria (American Psychiatric Association, 1994), (b) maximizes fieldwork resources and reduces survey costs (Kessler & Üstün, 2004), and (c) reduces respondent burden by averting potentially lengthy diagnostic interviews for adults deemed not "at risk" of having experienced MDD based on their lack of endorsement of *depressed mood* and/or *anhedonia* (Kennedy, 2008). In recent years, however, behavioral and social scientists have been encouraged to, and have attempted to, exploit these rich survey data resources to address novel and important research questions on the conceptualization of psychopathology (Adelson, 2006; Caetano, 2015; Kessler et al., 2004; Kupfer et al., 2008). These questions may be theoretical in nature (e.g., how valid is the DSM's assumption of polythetic symptom criteria where "true" MDD requires the presence of one of the two core criteria for research on the nature of the depressive phenotype in the general population?) or simply descriptive (what is the prevalence of the additional diagnostic criteria for MDD such as *sleep disturbances* in the general population?). With a few exceptions (Hoffman, Steinley, Trull, Lane, et al., 2019; Robins & Cottler, 2004), there has been surprisingly little debate about how the use of the skip-out procedure may impact the secondary use of such survey resources. One pertinent issue is the scale of missing data which use of the skip-out procedure typically generates: for example, due to the skip-out in NESARC Wave 1, information on the additional diagnostic criteria for MDD such as *sleep problems* is missing for ~68% of the sample ($N = 29{,}430$) (Grant et al., 2003).

Researchers' understanding about the skip-out design is important because it will likely inform their approach to handling missing data on the additional diagnostic criteria. Missing data produced via the skip-out are not missing completely at random (MCAR), but missing at random (MAR) (Rubin, 1976); this means that the probability of missingness on the additional diagnostic criteria is linked to respondent characteristics that were assessed (i.e., the non-endorsement of the core MDD criteria). MAR data pose a considerable analytical challenge in epidemiological studies (Van der Heijden et al., 2006) and researchers tend to engage in two common strategies to deal with large blocks of missing data generated by the skip-out within diagnostic modules.

The first strategy, which generally reflects psychiatry's conceptualization of MDD, is listwise deletion or complete case analysis (e.g., Carragher et al., 2009; Saito et al., 2010). This approach is problematic when the proportion of missingness is greater than 5% (Liu & De, 2015) because losing larger proportions of a sample may result in standard errors being substantially larger than they would be under missing data methods that preserved more of the available data. Moreover, when conducting criteria-level analyses using a listwise deletion approach, in the pairwise contingency table for binary coded present versus absent symptom criteria, the zero-zero cell for the *depressed mood* and *anhedonia* criteria pair will be empty (i.e., a structural zero in the sample) (Akande et al., 2017) since both criteria cannot be zero due to the skip-out procedure. This poses challenges for traditional statistical tests (e.g., chi-square analysis) (Finkler, 2010).

The second common strategy is to simply replace or substitute the missing values resulting from the skip-out with zeros. This approach is sometimes referred to as "imputation with zeros" (Huisman, 2009), but it is not a formal computational imputation approach such as multiple imputation (MI). It makes the strong assumption that if respondents have not experienced either of the core criteria for MDD, then they could not have experienced any of the additional symptoms and that the prevalence of these additional diagnostic criteria is likely to be low and not associated with meaningful or clinically relevant social, occupational, or functional impairment. Although this approach preserves the overall sample size for statistical analyses, substituting missing data on the additional diagnostic criteria with large numbers of zeros is problematic because it can induce spurious correlations among the (binary coded) depression criteria by distorting (possibly drastically) the 0–1 proportions used to obtain thresholds in the estimation of the tetrachoric correlations (Austin et al., 1998; Pierotti et al., 2017). In other words, because one is coding the seven additional criteria as zero simultaneously for many people, one induces positive correlations among these criteria, as well as positive correlations between the core criteria and skipped-out criteria (if core criteria are zero, then the others are also zero).

Borsboom et al. (2017) provide some evidence on this issue in a re-analysis of MDD symptom data from the National Comorbidity Survey Replication (NCS-R) reported by Forbes et al. (2017). Although the NCS-R implemented the skip-out procedure, Borsboom et al. (2017) examined the tetrachoric correlations of the additional depression symptoms (i.e., *weight problems*, *sleep problems*, *psychomotor problems*, and *fatigue*) in two ways when *depressed mood* and *anhedonia* were absent: (a) substituting with zeros (i.e., all skip-out symptoms with missing information were replaced with zeros) and (b) the additional symptoms were treated as missing data using pairwise deletion or complete case analysis. The results showed that under condition 1, the average correlation between the symptoms was 0.94, nearly 3 times as large as the average correlation of 0.33 under condition 2. Further, substituting the missing data with zeros resulted in a nonpositive definite correlation matrix, introducing formal problems for conducting statistical analyses on the data and/or the subsequent interpretation of such analyses (Lorenzo-Seva & Ferrando, 2021).

Widespread use of the skip-out procedure in national surveys, however, has meant limited opportunities to investigate fully the implications of this survey design feature. For example, it is currently not known: (a) how common it is to experience the additional diagnostic criteria of MDD in the absence of *depressed mood* or *anhedonia* in the adult general population?; (b) whether additional criteria experienced in absence of core MDD criteria are associated with meaningful or clinically relevant social/occupational functional impairment, and to what extent is important survey data being lost if important experiences of psychological distress (e.g., *suicidal ideation*) are only asked of those respondents endorsing the core MDD criteria?; and (c) how the nature of associations between the full set of MDD criteria (core and additional) in the adult general population may be impacted by employing different strategies to manage missing data produced by the skip-out, and what implications this may have for conducting statistical analyses on this survey data to address specific research questions?

Here, we address these three important research questions by conducting a descriptive secondary analysis of the Virginia Adult Twin Study of Psychiatric and Substance Use Disorders (VATSPSUD) (Kendler & Prescott, 2006). VATSPSUD is a comprehensive mental health survey that collected detailed information on MDD symptoms without implementing the skip-out procedure, and the data have the additional benefit that information on disaggregated symptoms for the additional MDD criteria is also available. To address research question one, we report the prevalence of the seven additional MDD criteria (as well as the disaggregated symptom items) for adults who experienced *depressed mood* and/or *anhedonia* compared to those who do not. To address research question two, we establish the prevalence of the level of impairment associated with the additional MDD diagnostic criteria (and disaggregated symptom items) among individuals who experienced *depressed mood* and/or *anhedonia* compared to those who do not. To address research question three, we examine the associations between MDD criteria (and disaggregated symptom items) under three conditions: (a) when full information for all survey respondents is present; (b) when missing data produced by the skip-out procedure is substituted with zeros; and (c) when listwise deletion of missing data produced by imposing the "skip-out" is employed. For each condition, we inspect the composition of the correlation matrices for the MDD criteria (and disaggregated symptom items), as well as the corresponding eigenvalues. Inspecting the nature and strength of correlations in the matrices is important to determine whether, and how, the pairwise associations between MDD criteria (and disaggregated symptom items) are impacted by traditional approaches to handling missing data produced by the skip-out. Eigenvalues are presented as a first line of descriptive empirical information about how dimensionality may be impacted by the different strategies for handling missing values resulting from using skip-outs. Eigenvalues are characteristic roots (Hoffman & Kunze, 1971) derived from a linear decomposition of the correlation matrices. Examination of the eigenvalues produced under each condition is valuable in determining whether the correlation matrices are positive-definite (Wothke, 1993). Using evidence from these results, we offer some potential best practice solutions for researchers with respect to overcoming missing data issues produced by the skip-out procedure in large national mental health surveys.

Method

Sample

Data examined in this study are from interviews administered to two related cohorts of twins from the population-based VATSPSUD (Kendler & Prescott, 1999, 2006). Briefly, the Virginia Twin Registry is a database of multiple births records occurring in the Commonwealth of Virginia held by the Virginia Department of Health Statistics since 1918, which was established in 1980. Contact details for twins were obtained by matching names and birth dates to state records, such as those of the Department of Motor Vehicles (Kendler & Prescott, 2006). The first cohort was a population-based sample of same-sex female-female (FF) twin pairs born between 1934 and 1974. The second cohort consisted of male-male/male-female twins (MF) pairs born between 1940 and 1974. FF twins were assessed four different times whereas the MF sample was interviewed twice. In this study, we used data from the face-to-face wave 1 interviews for the FF sample (FF1; $n = 2{,}162$, M age = 30.1 years, $SD = 7.6$) conducted between January 1987 and July 1989 and the first MF assessment conducted between March 1993 and October 1996 (MF1; $n = 6843$, 75% male; M age = 35.5 years, $SD = 9.2$, female M age = 35.4 years, $SD = 9.0$). The VATSPSUD received ethical approval from the Institutional Review Board at Virginia Commonwealth University.

Measures

Assessment of Major Depressive Disorder (MDD)

A structured psychiatric interview, based on the Structured Clinical Interview for DSM-III-R (Spitzer et al., 1987), was used to assess the prevalence of nine binary diagnostic criteria for MDD occurring during the last year: (a) core: *depressed mood*, *anhedonia* and (b) additional: *weight/appetite changes*, *sleep disturbances*, *psychomotor changes*, *fatigue*, *guilt/worthlessness*, *concentration difficulties*, and *suicidal ideation*. These criteria were operationalized by 14 disaggregated symptom items. Specifically, the weight/appetite change symptom was broken down into four items: (a) weight gain, (b) weight loss, (c) appetite increase, and (d) appetite decrease. Similarly, sleep problems (two items, one for insomnia, and another for hypersomnia) and psychomotor problems (two items, one for feeling slowed down and another for being restless) are aggregated to form the corresponding single symptom criteria used to determine diagnostic status. For example, if any of the four weight/appetite increase/decrease items is positive, the single weight/appetite change criteria variable will be positive (set to 1). These disaggregated symptoms are included and asked as separate items in the depression interview module and are then aggregated (i.e., collapsed over) to create the nine MDD criteria (refer to Table 1 for exact question wording).

Two important points with respect to the assessment of MDD in this study are worth noting. First, the SCID-DSM-III-R assessed for difficulties with thinking or concentration only; indecisiveness was not explicitly stated in the interview question (i.e., *Had trouble thinking or concentrating most of the time?*), which departs somewhat from the DSM-III-R specification (Cassidy et al., 1997). Second, the DSM-III-R MDD criteria differ slightly from those currently listed in DSM-5 with respect to suicidal ideation; specifically, participants in VATSPSUD were asked a broad question about suicidal ideation (i.e., *Thought a lot about death or about harming yourself?*) whereas DSM-5 assesses four aspects of suicidal ideation (i.e., thoughts of death, suicidal thinking, suicide attempt, and having a specific plan) (Uher et al., 2014).

During the interview, each member of a twin pair was interviewed separately by different interviewers and were asked whether they had experienced each of the 14 symptom items over the past year. If the participant responded "yes," they were asked if (a) they thought the symptom was the result of an illness or the taking of medication and (b) whether the symptom was severe enough that it interfered with their daily activities (interference item). Information on symptoms reported by participants which occurred due to an illness or medication was collected by interviewers, but only symptoms not reported to be due to medical illness or the taking of medications were coded

Table 1

*Comparison of Endorsement of Additional MDD Symptoms (and Associated Diagnostic Criteria) by Individuals Who Did or Did Not Experience Depressed Mood and/or Anhedonia (*N *= 8,980)*

		"Skip-out"—No depressed mood or anhedonia ("skip" subsample) $N = 5{,}685$ (63.3%)		Depressed mood and/or anhedonia ("complete" subsample) $N = 3{,}295$ (36.7%)		
MDD criterion	In the last year, have you had a time lasting at least 5 days, when you…	Occurred only (%)	Occurred and interfered (%)	Occurred only (%)	Occurred and interfered (%)	Chi-square, (*df*), *p*, *Cramer's V*
Weight/appetite changes	Had a significant decrease in your appetite?	2.0	0.4	18.8	9.4	1,332.203 (2) < .0001, $V = 0.385$
	Had a significant increase in your appetite?	2.5	0.2	10.1	3.5	409.930 (2) < .0001, $V = 0.214$
	Weren't trying to diet, when you lost a significant amount of weight (at least 2 lbs/week or 7 lbs total)?	0.5	1.0	3.4	12.1	665.560 (2) < .0001, $V = 0.272$
	Gained a significant amount of weight (at least 2lbs/week or 7lbs total)?	0.5	2.1	1.9	10.1	332.171 (2) < .0001, $V = 0.189$
Sleep disturbances	Had trouble sleeping nearly every night?	3.4	1.3	22.0	16.9	1,732.572 (2) < .0001, $V = 0.439$
	Slept considerably more than usual nearly every day?	0.8	0.4	7.7	7.0	659.858 (2) < .0001, $V = 0.271$
Psychomotor changes	Felt slowed down, that is moving or talking more slowly than is normal for you?	1.2	0.3	13.3	6.5	940.147 (2) < .0001, $V = 0.324$
	Were fidgety or restless most of the time?	3.7	0.6	26.9	7.9	1,477.417 (2) < .0001, $V = 0.406$
Fatigue	Felt tired or fatigued most of the time?	5.1	0.7	32.1	14.7	2,151.920 (2) < .0001, $V = 0.490$
Guilt/worthlessness	Were bothered by feeling worthless or guilty about things?	0.7	0.2	19.4	9.2	1,636.461 (2) < .0001, $V = 0.427$
Concentration difficulties	Had trouble thinking or concentrating most of the time?	1.2	0.5	15.7	12.6	1,458.715 (2) < .0001, $V = 0.403$
Death/self-harm	Thought a lot about death or about harming yourself?	0.1	0.1	6.7	4.1	604.989 (2) < .0001, $V = 0.260$

as being positively endorsed. Next, the interviewer enquired further as to which of the positively endorsed symptoms reported to have occurred during the year prior to the interview had temporally occurred together (i.e., that the set of symptoms was experienced at the same time rather than being scattered across the entire year, and the two core criteria needed to persist for at least 5 days). This was done to ensure that the symptom set used to determine diagnostic status formed a syndrome consistent with the DSM clinical definition for determining affected and unaffected MDD status.

Data Preparation

To utilize the full amount of information available for the individual MDD symptoms and associated levels of interference in the VATSPSUD interview (and in an attempt to mirror the SCID-5 assessment of major depressive episode symptoms), analyses were performed for the MDD diagnostic criteria using both binary and ordinal response scale coding. Inclusion of the interference information at the symptom level has not been utilized widely in analyses of the VATSPSUD data, and so our research team derived an operationalization of interference for this set of analyses. Specifically, all binary symptom item scales (i.e., present/absent) were extended by incorporating responses to the interference items to create new ordinal symptom items: did not occur (as part of a syndrome)—scored as “0”; occurred but no interference in daily life—scored as “1”; and occurred and interfered in their daily life (combined response options of “completely” or “a lot”)—scored as “2.” For *weight gain* and *weight loss* symptoms, a threshold cut off of ≥10 pounds was used to determine “inference”; similarly, for *sleep disturbances* symptoms, ≥ 4 hr of more or less sleep was used to create the interference category. Ordinal level diagnostic criteria operationalized by more than one item (i.e., appetite, sleep disturbances, and psychomotor changes) were generated as follows: a score of 0 across all symptoms assessing the criterion indicated absence of the criterion; any score of 1 but not 2 on any symptoms assessing the criterion indicated occurrence but no interference; and scores of 2 on one or more of the symptoms indicated the criterion was endorsed with interference in daily functioning.

Analytic Plan

Study aims were addressed via three analytic stages. In Stage 1, we determined the proportion of the sample who would have skipped out on the MDD diagnostic module, if this procedure would have been used in the VATSPSUD—that is, the proportion of respondents who endorsed neither depressed mood nor anhedonia. In Stage 2, we compared the endorsement patterns for the 14 disaggregated MDD symptoms items (ordinal scale) for individuals who would have skipped out of the MDD module (“skip” subsample), compared to those endorsing *depressed mood* and/or *anhedonia* (“complete” subsample). For Stages 1–2, frequencies, cross-tabulations, and chi-square significant tests and associated effect sizes (Cramer’s V) were conducted using Stata v15 (StataCorp., 2017). In Stage 3, we first compared the estimated tetrachoric correlation matrices for the nine binary MDD diagnostic criteria under three different conditions: (a) using complete data available in VATSPSUD (Condition A); (b) substituting the nine MDD diagnostic criteria values with zero for individuals who would have skipped out (Condition B); and (c) imposing a missing data structure for individuals who would have skipped-out (Condition C). Analyses were conducted using the *tetrachoric* command in Stata v15 (StataCorp., 2017). And second, we report the eigenvalues for each matrix for each condition to assess whether the correlation matrices are positive-definite, and to assess how the dimensionality of the MDD criteria (and symptoms) may alter across Conditions A–C. We focus on reporting findings with respect to the nine MDD diagnostic criteria (binary response scale) to relate to the large literature based on the presence or absence of the individual MDD diagnostic criteria (Aggen et al., 2005; Andrews et al., 2007; Chang et al., 2008; Geoffroy et al., 2018; Krueger & Finger, 2001; Reise & Waller, 2009; Zbozinek et al., 2012). Supplementary analyses were also conducted and reported on separately to exploit the richness of the VATSPSUD data. Specifically, the third stage of analysis for the nine MDD diagnostic criteria was run using the ordinal response scale and also for the 14 disaggregated MDD symptom items (using both binary and ordinal response scales) to test whether the findings were consistent across the different item scoring schemes (see online supplemental materials).

Data, Materials, and Code

Materials and analysis code for this study are not available.

Preregistration of Studies and Analysis Plans

This study was not preregistered.

Results

Stage 1

Two-thirds of respondents (63.3% of sample; $n = 5{,}685$, “skip” subsample) endorsed neither *depressed mood* nor *anhedonia* and would have skipped out and not been asked the symptom items used to operationalize the additional diagnostic criteria for MDD had this feature been implemented in the VATSPSUD interview. Of those reporting having experienced *depressed mood* or *anhedonia* in the past year (36.7% of the sample; $n = 3{,}295$, “complete” subsample), the most common endorsement pattern reported for these two core symptoms was to have experienced both symptoms (53.9%), followed by *depressed mood* only (36.6%) and *anhedonia* only (9.5%).

Stage 2

Symptoms items used to operationalize the additional diagnostic criteria for MDD were endorsed at statistically significant lower levels in the “skip” subsample compared to “complete” subsample (see Table 1). Overall, for the “skip” subsample, the most commonly endorsed symptom was *fatigue* (5.8%), the least commonly endorsed *suicidal ideation* (0.2%); in the “complete” subsample, these two symptoms were also the most and least commonly endorsed symptoms (46.8% and 10.8%, respectively). Important differences emerged between the two subsamples when considering symptom occurrence in more detail. In the “skip” subsample, the proportion of individuals reporting that the additional symptom interfered with daily life was lower for all symptoms. For example, for *fatigue*, only 0.7% in the “skip” subsample reported that this symptom interfered with daily life, compared to 14.7% in "complete" subsample. Similarly, *suicidal ideation* was associated with very minimal daily life interference in the "skip" subsample (0.1%) compared to "complete" subsample (4.1%). Medium to large effect sizes were evident for all comparisons (except for *increased weight*, which was small).

Stage 3

Tables 2–4 present the endorsement frequencies for the nine binary MDD diagnostic criteria and the correlation matrices (including all pairwise correlations, standard errors, thresholds, and eigenvalues) for each of the Conditions: A (complete data), B (additional diagnostic criteria set to zero), and C (additional diagnostic criteria set to missing). The top section summarizes the endorsement frequencies for each of the MDD criteria (0 = not present, 1 = present, and NA = missing). The next section presents the point estimates for the MDD inter-criteria tetrachoric correlations. Below that are the corresponding standard errors for these estimated correlations. The last two rows present the item threshold estimates and the eigenvalues for the full correlation matrix. The mean and median correlations for each Condition are summarized in Table 5. We used heat maps to visualize the strength of the pairwise correlations between the 9 binary MDD diagnostic criteria across Conditions A–C (Figure 1).

With respect to the endorsement frequencies, the sample size for *depressed mood* and *anhedonia* is constant across Conditions A–C (Tables 2–4), whereas these frequencies change for the additional diagnostic criteria. For example, for *weight/appetite changes*, the frequencies of "1s" (criterion present) drop from 1884 in Table 2 for Condition A (complete data) to 1,509 in Table 3 (additional criteria set to zero) when changing the responses of those who would have skipped out. This difference ($N = 375$) is added to the frequency of "0s" (criterion absent) in Table 3. Considering this example further, in Table 4, the frequency of positive endorsement ("1s") for *weight/appetite changes* remains at 1,509, but the number of "0s" has been dramatically reduced when the 5,686 "0s" are set as missing (Condition C).

When complete information was available as in Condition A (Table 2, Figure 1A), pairwise correlations between the diagnostic criteria ranged from 0.453 to 0.849 (mean = 0.652; median = 0.648). Substituting missing values with zeros for those who skip-out in Condition B (Table 3, Figure 1B) produced a narrower range of correlations, and the mean and median correlations were modestly higher (0.513–0.882; mean = 0.735, median = 0.736). Substituting the observed data with zeros in Condition B for those who would have skipped out resulted in inflated pairwise correlations when compared to Condition A. For example, the pairwise correlation between *depressed mood* and *fatigue* was 0.716 when complete information was available (Condition A) compared to 0.870 when reported data was treated as zero (Condition B). Compared to Conditions A and B, the range of correlations was wider in Condition C, which mimics the skip-out procedure (see Table 4, Figure 1C), and the mean and median correlations were lower (range 0.150–0.849; mean = 0.320, median = 0.295).

Inspection of the eigenvalues indicated some important differences across Conditions A–C (see Tables 2–4). Condition A had a

Table 2

Summary of Nine Binary MDD Diagnostic Criteria Endorsement Frequencies, Tetrachoric Correlations, Standard Errors, Item Thresholds, and Eigen Values (Condition A: Complete Data on Additional Diagnostic Criteria for MDD)

MDD diagnostic criteria	dm	li	wa	sp	pm	fa	gw	cc	td
				Frequency					
0	5,998	6,892	7,096	7,070	7,167	7,107	7,984	7,949	8,611
1	2,982	2,088	1,884	1,908	1,813	1,870	994	1,030	367
Missing	0	0	0	2	0	3	2	1	2
				Correlation					
dm	—	.849	.694	.737	.716	.716	.790	.729	.758
li	.849	—	.641	.674	.683	.660	.707	.705	.639
wa	.698	.641	—	.630	.609	.588	.574	.573	.512
sp	.737	.674	.630	—	.689	.694	.592	.616	.565
pm	.716	.683	.609	.689	—	.672	.625	.700	.557
fa	.716	.660	.588	.694	.672	—	.578	.614	.453
gw	.790	.707	.574	.592	.625	.578	—	.655	.717
cc	.729	.705	.573	.616	.700	.614	.655	—	.555
td	.758	.639	.512	.565	.557	.453	.717	.555	—
				SE					
dm	—	.008	.012	.011	.012	.011	.012	.013	.021
li	.008	—	.013	.013	.013	.013	.014	.014	.022
wa	.012	.013	—	.014	.014	.015	.017	.017	.025
sp	.011	.013	.014	—	.013	.012	.017	.016	.024
pm	.012	.013	.014	.013	—	.013	.016	.014	.024
fa	.011	.013	.015	.012	.013	—	.017	.016	.026
gw	.012	.014	.017	.017	.016	.017	—	.017	.019
cc	.013	.014	.017	.016	.014	.016	.017	—	.025
td	.021	.022	.025	.024	.024	.026	.019	.025	—
Threshold1	0.434	0.731	0.807	0.798	0.835	0.812	1.223	1.202	1.741
Eigenvalue	6.236	0.678	0.453	0.397	0.341	0.289	0.260	0.237	0.109

Note. dm = depressed mood, li = loss of interest/anhedonia, wa = any weight/appetite increase/decrease, sp = any sleep problems (insomnia and/or hypersomnia, pm = any psychometric problems agitation/retardation), fa = fatigue, gw = feelings of guilt or worthlessness, cc = inability to concentrate, td = thoughts of death or self-harm.

Table 3

Summary of Nine Binary MDD Diagnostic Criteria Endorsement Frequencies, Tetrachoric Correlations, Standard Errors, Item Thresholds, and Eigen Values (Condition B: Additional Diagnostic Criteria for MDD Set to Zero for "Skip" Subsample)

MDD diagnostic criteria	dm	li	wa	sp	pm	fa	gw	cc	td
				Frequency					
0	5,998	6,892	7,471	7,391	7,481	7,436	8,034	8,046	8,621
1	2,982	2,088	1,509	1,588	1,499	1,541	944	933	357
Missing	0	0	0	1	0	3	2	1	2
				Correlation					
dm	—	.849	.872	.882	.870	.870	.852	.833	.818
li	.849	—	.764	.773	.786	.767	.738	.766	.659
wa	.872	.764	—	.741	.729	.694	.669	.673	.585
sp	.882	.773	.741	—	.770	.768	.676	.695	.621
pm	.870	.786	.729	.770	—	.756	.708	.766	.626
fa	.870	.767	.694	.768	.756	—	.662	.695	.513
gw	.852	.738	.669	.676	.708	.662	—	.701	.734
cc	.833	.766	.673	.695	.766	.695	.701	—	.587
td	.818	.659	.575	.621	.626	.513	.734	.587	—
				SE					
dm	—	.008	.008	.007	.008	.008	.010	.011	.021
li	.008	—	.011	.011	.010	.011	.013	.012	.021
wa	.008	.011	—	.012	.013	.013	.016	.016	.023
sp	.007	.011	.012	—	.011	.011	.015	.015	.022
pm	.008	.010	.013	.011	—	.012	.014	.013	.022
fa	.008	.011	.013	.011	.012	—	.016	.015	.025
gw	.010	.013	.016	.015	.014	.016	—	.016	.019
cc	.011	.012	.016	.015	.013	.015	.016	—	.025
td	.021	.021	.023	.022	.022	.025	.019	.025	—
Threshold1	0.434	0.731	0.962	0.927	0.966	0.948	1.253	1.260	1.753
Eigenvalue	6.909	0.583	0.354	0.299	0.262	0.211	0.205	0.194	−0.018

Note. dm = depressed mood, li = loss of interest/anhedonia, wa = any weight/appetite increase/decrease, sp = any sleep problems (insomnia and/or hypersomnia, pm = any psychometric problems (agitation/retardation), fa = fatigue, gw = feelings of guilt or worthlessness, cc = inability to concentrate, td = thoughts of death or self-harm.

Table 4

Summary of Nine Binary MDD Diagnostic Criteria Endorsement Frequencies, Tetrachoric Correlations, Standard Errors, Item Thresholds, and Eigen Values (Condition C: Additional Diagnostic Criteria for MDD Set to Missing for "Skip" Subsample)

MDD diagnostic criteria	dm	li	wa	sp	pm	fa	gw	cc	td
				Frequency					
0	5,998	6,892	1,786	1,706	1,796	1,753	2,349	2,361	2,936
1	2,982	2,088	1,509	1,588	1,499	1,541	944	933	357
Missing	0	0	5,685	5,686	5,685	5,686	5,687	5,686	5,687
				Correlations					
dm	—	.849	.184	.207	.174	.150	.370	.278	.508
li	.849	—	.222	.224	.284	.219	.302	.371	.293
wa	.184	.222	—	.304	.295	.209	.284	.295	.262
sp	.207	.224	.304	10.000	.375	.361	.283	.322	.307
pm	.174	.284	.295	.375	—	.352	.360	.475	.327
fa	.150	.219	.209	.361	.352	10.000	.266	.330	.151
gw	.370	.302	.284	.283	.360	.266	10.000	.428	.554
cc	.278	.371	.295	.322	.475	.330	.428	—	.336
td	.508	.293	.262	.307	.327	.151	.554	.336	—
				SE					
dm	—	.008	.046	.045	.045	.046	.052	.052	.079
li	.008	—	.026	.026	.026	.026	.028	.028	.039
wa	.046	.026	—	.026	.026	.027	.027	.027	.035
sp	.045	.026	.026	—	.025	.025	.027	.027	.034
pm	.045	.026	.026	.025	—	.025	.026	.024	.034
fa	.046	.026	.027	.025	.025	—	.028	.027	.036
gw	.052	.028	.027	.027	.026	.028	—	.026	.029
cc	.052	.028	.027	.027	.024	.027	.026	—	.034
td	.079	.039	.035	.034	.034	.036	.029	.034	—
Threshold1	0.434	0.731	0.106	0.045	0.113	0.081	0.563	0.573	1.235
Eigenvalue	3.593	1.354	0.913	0.784	0.712	0.624	0.514	0.413	0.094

Note. dm = depressed mood, li = anhedonia, wa = any weight/appetite increase/decrease, sp = any sleep problems (insomnia and/or hypersomnia, pm = any psychometric problems (agitation/retardation), fa = fatigue, gw = feelings of guilt or worthlessness, cc = inability to concentrate, td = thoughts of death or self-harm.

Table 5

Summary of Pairwise Correlation Range, Mean, and Median Correlations Obtained From Conditions A–C Using MDD Symptom Item and Diagnostic Criteria Sets Using Binary and Ordinal Response Scales

	Condition A Complete data on additional MDD criteria/items			Condition B Additional MDD criteria/items set to Zero for "skip" subsample			Condition C Additional MDD criteria/items set to missing for "skip" subsample		
MDD diagnostic criteria/symptom set	Correlation range	Mean	Median	Correlation range	Mean	Median	Correlation range	Mean	Median
9 binary MDD diagnostic criteria	0.453 to 0.849	0.652	0.648	0.513 to 0.882	0.735	0.736	0.150 to 0.849	0.320	0.295
9 ordinal MDD diagnostic criteria	0.468 to 0.839	0.635	0.639	0.517 to 0.839	0.704	0.710	0.272 to 0.839	0.390	0.354
14 binary MDD symptom items	−0.087 to 0.849	0.504	0.517	0.027 to 0.870	0.594	0.598	−0.276 to 0.849	0.231	0.219
14 ordinal MDD symptom items	−0.084 to 0.839	0.494	0.517	0.027 to 0.843	0.567	0.579	−0.269 to 0.839	0.275	0.267

well-conditioned invertible matrix (all positive eigenvalues) and a unidimensional structure (indicated by one large positive eigenvalue of 6.236). Condition B, although similar to Condition A in terms of unidimensionality (i.e., one large positive eigenvalue of 6.909), produced an ill-condition matrix with a negative eigenvalue. Condition C produced a positive-definite matrix but pointed toward two underlying dimensions, as indicated by two eigenvalues greater than one (3.592 and 1.453). This points to the fact that important psychometric properties such as unidimensionality are affected by different ways to deal with skip-out data. It is important to emphasize that what could be considered as trivial recoding of the item data introduced alternations in the patterning of the inter-item correlations.

In Figure 2, we plot the estimated pairwise correlations between the nine binary MDD diagnostic criteria with corresponding symmetric lower and upper standard error boundaries for each of the Conditions (A in blue; B in red; C in green). Dashed vertical lines group the correlations for each of the three conditions for each MDD diagnostic criteria. The only overlap in the tetrachoric point estimates and standard error boundaries across conditions is for *depressed mood* and *anhedonia*, which is as expected given the data is identical for these criteria across conditions. When comparing the complete data (Condition A, blue) to the replacement with zeros for skip-outs (Condition B, red), there is some overlap in the point estimates and standard error boundaries for "anhedonia and worthlessness/guilt," "anhedonia and suicidal ideation," "worthlessness/guilt and suicidal ideation," and "concentration difficulties and suicidal ideation"; however, there was no overlap between Condition C (green) or Conditions A/B in terms of the point estimates and standard error boundaries.

For the replication of Stage 3 (see online supplemental materials), polychoric correlation matrices were estimated for the nine diagnostic criteria ordinal scale that incorporates the interference information. Tetrachoric and polychoric correlations were also estimated for the 14 binary/ordinal disaggregated symptom item sets. Generally, when comparing the results for the 14 MDD symptom item set to the nine diagnostic criteria, regardless of whether binary or ordinal response scales are used, the range of correlation estimates tended to widen, with reductions in the mean and median correlations for each Condition A–C. Using the nine ordinal MDD diagnostic criteria set produced lower mean and median correlations for Conditions A and B but showed increased mean and median correlations for Condition C (see Table 5). Well-conditioned invertible matrices for the binary symptom item sets were produced for Conditions A and C but not Condition B; for the ordinal criteria and symptom item sets, similar issues for the condition of the matrix for Condition B were not observed. Evidence pointing toward unidimensionality emerged when the ordinal diagnostic criteria were analyzed for Conditions A and B, but not for Condition C. Evidence of multidimensionality was evident for all three conditions using the binary and ordinal symptom item sets.

Discussion

Hoffman, Steinley, Trull, and Sher (2019) recently cautioned that "researchers often utilize data without fully considering the structure of the diagnostic instrument and the form it imposes on resulting data" (p. 79). Inspired by this thesis, our paper aimed to evaluate the impact of using different approaches to manage and deal with complex skip-out design procedures in epidemiological survey data. We argue that widespread and routine use of the skip-out procedure in psychiatric epidemiological surveys designed primarily to assess the population prevalence of mental disorders, such as MDD, in accordance with established psychiatric classification systems, introduces limitations to the data which make it problematic to use the data to pursue other research aims.

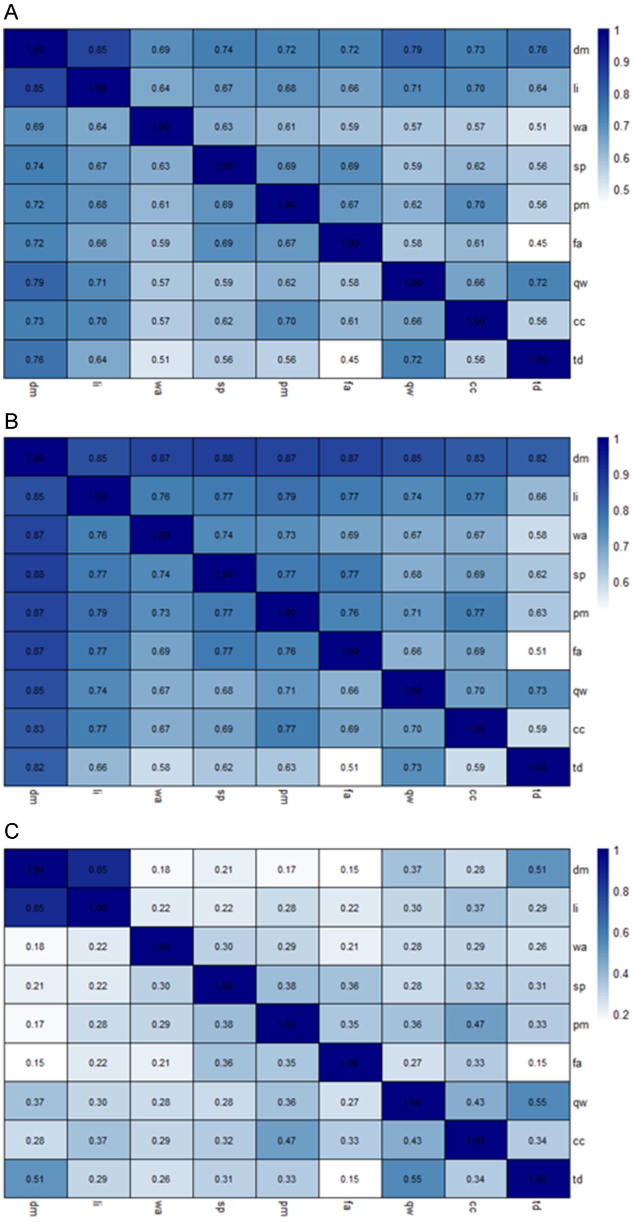
Our findings can be summarized succinctly. Approximately two-thirds (66.3%) of the population-based VATSPSUD sample did not endorse having experienced *depressed mood* or *anhedonia* for a two-week period during the last year and would have skipped-out of the MDD diagnostic module had this design feature been imposed. This proportion is consistent with previously surveyed samples that imposed the skip-out for MDD (e.g., 2001–2002 NESARC, 68%). *Depressed mood* was the most prevalent symptom among those participants who endorsed either or both of the core symptoms (only 9.5% of individuals reported *anhedonia* in the absence of *depressed mood*), which is consistent with evidence suggesting that *anhedonia* is endorsed at a considerably lower frequency than *depressed mood* (Buckner et al., 2008). Overall, it appears that there is nothing unusual about the prevalence of these core MDD criteria in the VATSPSUP sample compared to other national surveys.

We showed that there was a moderate to strong association between experiencing at least one of the core MDD criteria and levels of endorsement of each of the additional symptoms, as well as associated levels of interference in daily functioning. Together, these findings appear to (a) provide broad support in favor of the DSM's core tenet as to the centrality of *depressed mood* and *anhedonia* (substantive argument) and (b) justify the implementation of the skip-out procedure in surveys designed to assess the prevalence of MDD in accordance with psychiatric classification systems (statistical argument).

We see two reasons to argue against accepting each of these two conclusions. First, regarding the substantive argument, we counter that this is the case for any symptoms/criteria when all symptoms/

Figure 1

Pairwise Correlation Range for Nine MDD Binary Diagnostic Criteria Obtained From Conditions A–C



Note. (Panel A) Complete data, (Panel B) additional symptoms recoded as zero for "skip" subsample, and (Panel C) additional symptoms treated as missing for "skip" subsample. Strength of correlation indicated by darker color. See the online article for the color version of this figure.

criteria are positively inter-correlated, of course. For example, when we compare participants without *sleep disturbances* relative to those with *sleep disturbances*, we find significantly higher symptom frequencies on all other eight symptoms in the latter group: this is a necessary result that follows from a positive manifold of positively inter-correlated items. Second, while symptom endorsement and interference were lower in those who did not meet either core criterion of MDD, they were far from zero and may provide meaningful epidemiological and clinical insights. As a result, substituting missing data with zeros is problematic when the goal of the research study is to extend our nosological understanding of symptom patterns beyond those directly tied to current clinical definitions of MDD-affected status.

Second, regarding the statistical argument, our analyses demonstrated that, at first glance, substituting skip-out missingness with zeros produces patterns of association, and similarly structured correlation matrices, to the complete data as opposed to when listwise deletion (or complete case data) is analyzed. Consistent patterns of results emerged across the binary and ordinal diagnostic criteria analyses, as well as the binary and ordinal symptom item analyses. That said, some important differences warrant attention. Specifically, substituting skip-out missingness with zeros, when using the binary diagnostic criteria or symptoms item set, turns a well-defined tetrachoric correlation matrix into one that is ill-defined and unsuitable for statistical analysis (Wothke, 1993). This makes it a problematic option for dealing with this type of missing data. Further, missingness introduced by the imposed skip-out results in considerably lower mean and median correlations and alters the patterning of correlations so as to change the statistical evidence regarding the unidimensionality of the underlying symptom items/criteria. Thus, researchers interested in using national mental health survey data to examine, for example, the phenotype of major depression in the adult population, or psychometric properties of corresponding scales, are likely to arrive at radically different conclusions depending on their data preparation choices and analytic strategy.

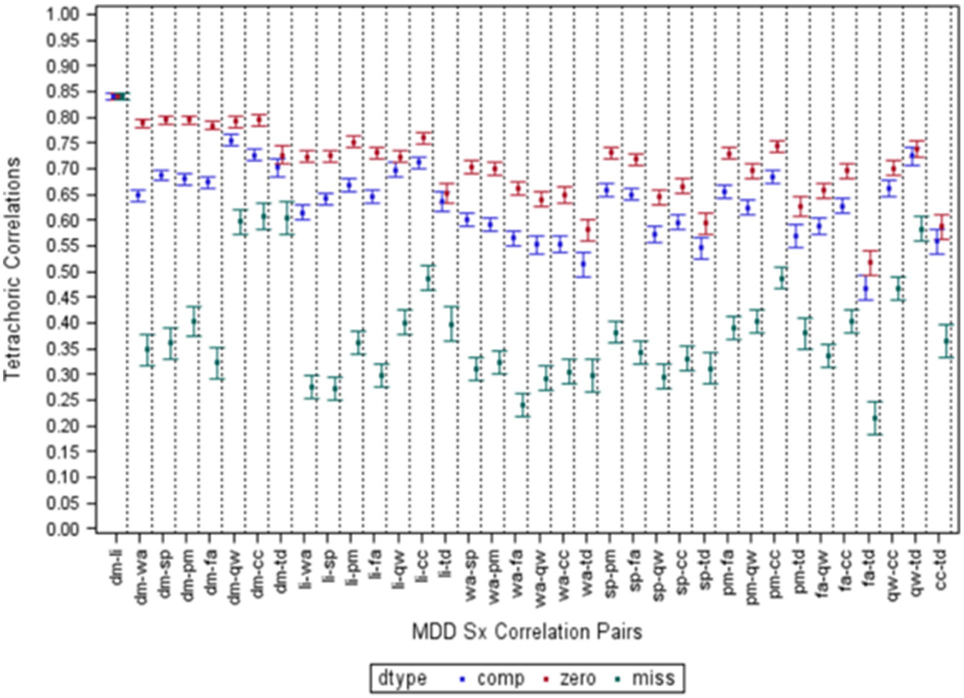
Before offering some points of consideration for survey methodologists, data analysts, researchers, and other stakeholders who have a vested interest in collecting and analyzing high-quality MDD (or other mental disorders) data in mental health surveys, we point to several limitations of our study. First, the analyses presented here are descriptive and no a priori hypotheses were pre-registered; our findings require replication using data from comparable surveys not implementing the skip-out. Second, as previously noted, the VATSPSUD assessment for MDD was in accordance with the DSM-III-R/DSM-IV. The assessment of a few criteria (e.g., *suicidal ideation*) differ somewhat from the current DSM-5 specification of MDD. As a result, we were not able to examine in any further detail experiences such as suicidal attempts for survey respondents. Third, since interference data at the symptom level was not routinely used in analyses of the VATSPSUD data, our team opted to impose quite stringent objective criteria to assess for interference with respect to the additional MDD criteria (i.e., four or more hours of disturbance in sleep patterns; ten or more pounds of weight loss or gain). Although there was no existing reliability and validity evidence for scores from this operationalization, we considered that these thresholds captured meaningful levels of impairment as opposed to minor fluctuations in typical weight changes or sleep routines. It is possible, however, that smaller changes in weight or, in particular, shorter durations of disruption to normal sleep patterns may be associated with clinically-relevant levels of daily interference.

Existing Survey Data Resources

Current guidance exists as to how researchers should handle survey data that is MAR. For example, best practice dictates that when data are MAR and when levels of missing data are high $> 10\%$, researchers should only use multiple imputation (MI) or full information

Figure 2

Summary of Pairwise Correlation Point Estimate and Standard Error Boundaries for 36 Pairwise Correlations for Nine MDD Binary Diagnostic Criteria Obtained From Conditions A–C



Note. Condition A (blue) complete data, Condition B (red) additional symptoms recoded as zero for "skip" subsample, and Condition C (green) additional symptoms treated as missing for "skip" subsample. Overlapping standard error boundaries indicates no statistically significant difference between the Conditions in terms of the strength of the pairwise correlation. See the online article for the color version of this figure.

maximum likelihood (FIML) estimation to managing missing survey data appropriately (Little et al., 2014). Multiple imputation involves making a copy of the original dataset and replacing missing data with plausible estimates of what the data would have been, had the participant been asked the question and provided an answer in the survey (Little & Rubin, 1987). MI introduces complexities to the data analysis because typically 20 to 100 imputations are required to recover the missing values in most cases, and the process involves fitting multiple replicates of the statistical analysis and pooling the results before drawing inferences (Graham et al., 2007). Little et al. (2014) recommend that FIML be considered as a simpler alternative to MI when the statistical model can accommodate maximum likelihood estimation (e.g., multilevel modeling or structural equation modeling).

An important caveat, however, is that conducting MI or FIML using only data that is available on symptoms gathered using the skip-out can introduce biases when analyzing the symptom-level data. We posit that a more defensible approach is that MI/FIML approaches for skip-out data would be better informed by obtaining information from comparable sample(s) that did not use skip-outs for those respondents who did not endorse *depressed mood* or *anhedonia*. Fortunately, the availability of surveys, such as the VATSPSUD, and other more recently conducted studies that did not implement the skip-out procedure when assessing for the occurrence of symptoms of mental disorders in the last year (e.g., Kaiser et al., 2020), provide an opportunity to conduct methodological work to help inform this approach to MI/FIML in the future. This is an important area for future research to explore.

Future Mental Health Surveys

Designing diagnostic modules in large-scale mental health surveys so that all sampled respondents are asked about the presence or absence of additional symptoms, including associated level of impairment or interference with functioning, for all mental disorders assessed in a survey according to current classification system guidance would provide the optimal item-level data for researchers and analysts. However, this approach is likely to be impractical for survey methodologists charged with designing, conducting, and ultimately paying for data collection since requiring each respondent to provide a response to each item will increase the burden for participants as well as adding to the overall project costs. Moreover, this type of approach to measurement is generally only feasible when the time reference for the occurrence of mental disorders in a survey relates to the last year (as opposed to lifetime) in order to reduce recall bias (Patten, 2003).

Leading experts in the missing data field (Little et al., 2014) and applied researchers alike (Hoffman, Steinley, Trull, & Sher, 2019) advocate that more careful consideration of the "not-missing-by-design" or planned missing data design in survey research is warranted to help mitigate issues with missing data, including those associated with the use of the skip-out. This type of experimental survey design affords researchers an efficient way to maximize data collection from respondents, while simultaneously providing a degree of control with respect to survey timings, costings, and respondent burden (Imbriano & Raghunathan, 2020; Rhemtulla & Little, 2012). Although planned missing data designs have been recommended for decades (Shoemaker, 1973), there have been calls in recent years for additional research to both improve the design of such methods and to increase their implementation in epidemiological studies (Peytchev & Peytcheva, 2017; Rioux et al., 2020).

One specific approach is the Split Questionnaire Survey Design (SQSD; Raghunathan and Grizzle (1995), which uses matrix sampling to administer the survey. Applying this design to the context of the MDD diagnostic module in large surveys, the full diagnostic module (i.e., the "long questionnaire") would be randomly administered to a proportion of the sample, regardless of the respondent's endorsement of *depressed mood* and *anhedonia*, to generate a complete dataset. We advocate that this long questionnaire should follow the process used in the VATSPSUD, that is: (a) assess for the occurrence of each symptom within a specific time frame (e.g., last year); (b) determine, in detail, changes in complex behaviors such as sleeping patterns, weight changes, or psychomotor activities (i.e., to measure both increases and decreases in these behaviors); (c) identify the levels of functional impairment associated with any symptoms experienced; and (d) enquire about the temporal ordering of the symptoms during the chosen time frame.

For the remainder of the sample, the diagnostic module would assess the two core symptoms and a smaller number of additional symptoms, ideally in a way that each possible pair of questions is observed in the partial dataset (Peytchev & Peytcheva, 2017). Assignment to conditions is random so that the unobserved part of the questionnaire for any individual can be treated as MCAR (Little et al., 2014; Raghunathan & Grizzle, 1995). Since MCAR produces no bias in the estimated parameters of a given statistical model, the resulting data could then be analyzed appropriately using MI or FIML.

It is challenging to determine an optimal SQSD in surveys in such a way in which it both avoids the potential loss of important information and also maximizes the success of subsequent approaches taken to handle the missing data in the partial dataset. However, researchers armed with information about the inter-relationships between all variables/survey items, which can be obtained from analysis of complete datasets, are in a better position to plan and implement this design more effectively (Adıgüzel & Wedel, 2013; Imbriano & Raghunathan, 2020; Raghunathan & Grizzle, 1995; Vriens et al., 2001). Similar to the need for additional work to be conducted on sample data not employing the skip-out to help inform MI/FIML approaches for existing survey resources, novel methodological work is now required to help inform the development of optimal SQSD designs for future mental health surveys.

A related planned missing design option is to address how the issue of the skip-out may be achieved through an adaption of the two-method design (Little & Rhemtulla, 2013). In this planned missing data approach, a gold-standard measure is administered to a random subsample of survey participants, and a biased (typically self-report) measure is administered to the entire sample. The resulting data can be modeling using a latent factor approach where both the gold-standard measure and the self-report measure load on a common factor (capturing the variance common across both measures), and the self-report measure also loads on a bias factor (capturing the variance that is shared only among the indicators of the self-report after conditioning on the common factor) (see (Graham et al., 2006); Rioux et al. (2020)).

The benefit of this approach is that when the data are modelled using the latent factor approach, the regression parameters are found to be more valid when compared to using a self-report alone, and provide better power given that it would be typical to administer the "gold standard" measure to a smaller sample (Rioux et al., 2020). Despite the potential usefulness of this approach, survey researchers considering this experimental design would still face challenges with respect to the traditional skip-out approach in the "gold standard" measure (i.e., the structured clinical interview adopting the DSM diagnostic approach) used in large-scale mental health surveys, and power is less of an issue given the sample sizes typically recruited for these types of surveys. Nevertheless, this option and others outlined here warrant further consideration and exploration by researchers interested in exploring avenues to improve the collection of robust epidemiological data in future large-scale mental health surveys. This will help ensure that valuable survey resources are maximized to address important nosological and psychopathological research questions going forward.

References

Adelson, R. (2006). Nationwide survey spotlights US alcohol abuse. *Monitor on Psychology*, *37*(1), 30. https://www.apa.org/monitor/jan06/alcohol

Adıgüzel, F., & Wedel, M. (2013). *Split questionnaire design for massive surveys*. In *59th ISI world statistics conference, Hong Kong*.

Aggen, S. H., Neale, M. C., & Kendler, K. S. (2005). DSM Criteria for major depression: Evaluating symptom patterns using latent-trait item response models. *Psychological Medicine*, *35*(4), 475–487. https://doi.org/10.1017/S0033291704003563

Akande, O., Li, F., & Reiter, J. (2017). An empirical comparison of multiple imputation methods for categorical data. *The American Statistician*, *71*(2), 162–170. https://doi.org/10.1080/00031305.2016.1277158

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.).

Andrews, G., Brugha, T., Thase, M. E., Duffy, F. F., Rucci, P., & Slade, T. (2007). Dimensionality and the category of major depressive episode. *International Journal of Methods in Psychiatric Research*, *16*(S1), S41–S51. https://doi.org/10.1002/mpr.216

Austin, E. J., Deary, I. J., Gibson, G. J., McGregor, M. J., & Dent, J. B. (1998). Individual response spread in self-report scales: Personality correlations and consequences. *Personality and Individual Differences*, *24*(3), 421–438. https://doi.org/10.1016/j.paid.2005.10.018

Borsboom, D., Fried, E. I., Epskamp, S., Waldorp, L. J., van Borkulo, C. D., van der Maas, H. L., & Cramer, A. O. (2017). False alarm? A comprehensive reanalysis of "evidence that psychopathology symptom networks have limited replicability" by Forbes, Wright, Markon, and Krueger (2017). *Journal of Abnormal Psychology*, *126*(7), 989–999. https://doi.org/10.1037/abn0000306

Buckner, J. D., Joiner, T. E., Jr., Pettit, J. W., Lewinsohn, P. M., & Schmidt, N. B. (2008). Implications of the DSM's emphasis on sadness and anhedonia in major depressive disorder. *Psychiatry Research*, *159*(1–2), 25–30. https://doi.org/10.1016/j.psychres.2007.05.010

Caetano, R. (2015). A decade after NESARC: What has it told us? *Addiction*, *110*(3), 375–377. https://doi.org/10.1111/add.12627

Carragher, N., Adamson, G., Bunting, B., & McCann, S. (2009). Subtypes of depression in a nationally representative sample. *Journal of Affective Disorders*, *113*(1–2), 88–99. https://doi.org/10.1016/j.jad.2008.05.015

Cassidy, F., Murry, E., Forest, K., & Carroll, B. J. (1997). The performance of DSM-III-R major depression criteria in the diagnosis of bipolar mixed states. *Journal of Affective Disorders*, *46*(1), 79–81. https://doi.org/10.1016/S0165-0327(97)00084-0

Chang, S. M., Hahm, B.-J., Lee, J.-Y., Shin, M. S., Jeon, H. J., Hong, J.-P., Lee, H. B., Lee, D.-W., & Cho, M. J. (2008). Cross-national difference in the prevalence of depression caused by the diagnostic threshold. *Journal of Affective Disorders*, *106*(1–2), 159–167. https://doi.org/10.1016/j.jad.2007.07.023

Finkler, A. (2010). *Goodness of fit statistics for sparse contingency tables*. https://hal.archives-ouvertes.fr/hal-00490383v2

Forbes, M. K., Wright, A. G., Markon, K. E., & Krueger, R. F. (2017). Evidence that psychopathology symptom networks have limited replicability. *Journal of Abnormal Psychology*, *126*(7), 969–988. https://doi.org/10.1037/abn0000276

Geoffroy, P. A., Hoertel, N., Etain, B., Bellivier, F., Delorme, R., Limosin, F., & Peyre, H. (2018). Insomnia and hypersomnia in major depressive episode: Prevalence, sociodemographic characteristics and psychiatric comorbidity in a population-based study. *Journal of Affective Disorders*, *226*, 132–141. https://doi.org/10.1016/j.jad.2017.09.032

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science*, *8*(3), 206–213. https://doi.org/10.1007/s11121-007-0070-9

Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods*, *11*(4), 323–343. https://doi.org/10.1037/1082-989X.11.4.323

Grant, B. F., Dawson, D. A., Stinson, F. S., Chou, P. S., Kay, W., & Pickering, R. (2003). The Alcohol Use Disorder and Associated Disabilities Interview Schedule-IV (AUDADIS-IV): Reliability of alcohol consumption, tobacco use, family history of depression and psychiatric diagnostic modules in a general population sample. *Drug and Alcohol Dependence*, *71*(1), 7–16. https://doi.org/10.1016/S0376-8716(03)00070-X

Gruenberg, A. M., Goldstein, R. D., & Pincus, H. A. (2005). Classification of depression: Research and diagnostic criteria: DSM-IV and ICD-10. In J. Licinio, & M.-L. Wong (Eds.), *Biology of depression: From novel insights to therapeutic strategies* (pp. 1–12). Wiley-VCH. https://doi.org/10.1002/9783527619672.ch1

Hasin, D. S., & Grant, B. F. (2015). The National Epidemiologic Survey on Alcohol and Related Conditions (NESARC) waves 1 and 2: Review and summary of findings. *Social Psychiatry and Psychiatric Epidemiology*, *50*(11), 1609–1640. https://doi.org/10.1007/s00127-015-1088-0

Hoffman, K., & Kunze, R. (1971). *Linear algebra* (pp. 122–125). Prentice-Hall.

Hoffman, M., Steinley, D., Trull, T. J., Lane, S. P., Wood, P. K., & Sher, K. J. (2019). The influence of sample selection on the structure of psychopathology symptom networks: An example with alcohol use disorder. *Journal of Abnormal Psychology*, *128*(5), 473–486. https://doi.org/10.1037/abn0000438

Hoffman, M., Steinley, D., Trull, T. J., & Sher, K. J. (2019). Estimating transdiagnostic symptom networks: The problem of “skip outs” in diagnostic interviews. *Psychological Assessment*, *31*(1), 73–81. https://doi.org/10.1037/pas0000644

Huisman, M. (2009). Imputation of missing network data: Some simple procedures. *Journal of Social Structure*, *10*(1), 1–29. https://doi.org/10.21307/joss-2019-050

Imbriano, P. M., & Raghunathan, T. E. (2020). Three-Form Split Questionnaire Design for Panel Surveys. *Journal of Official Statistics*, *36*(4), 827–854. https://doi.org/10.2478/JOS-2020-0040

Kaiser, A. J., Funkhouser, C. J., Mittal, V. A., Walther, S., & Shankman, S. A. (2020). Test-retest & familial concordance of MDD symptoms. *Psychiatry Research*, *292*, Article 113313. https://doi.org/10.1016/j.psychres.2020.113313

Kendler, K. S., Muñoz, R. A., & Murphy, G. (2010). The development of the Feighner criteria: A historical perspective. *American Journal of Psychiatry*, *167*(2), 134–142. https://doi.org/10.1176/appi.ajp.2009.09081155

Kendler, K. S., & Prescott, C. A. (1999). A population-based twin study of lifetime major depression in men and women. *Archives of General Psychiatry*, *56*(1), 39–44. https://doi.org/10.1001/archpsyc.56.1.39

Kendler, K. S., & Prescott, C. A. (2006). *Genes, environment, and psychopathology: Understanding the causes of psychiatric and substance use disorders*. Guilford Press.

Kennedy, S. H. (2008). Core symptoms of major depressive disorder: Relevance to diagnosis and treatment. *Dialogues in Clinical Neuroscience*, *10*(3), 271–277. https://doi.org/10.31887/DCNS.2008.10.3/shkennedy

Kessler, R. C. (1994). The National Comorbidity Survey of the United States. *International Review of Psychiatry*, *6*(4), 365–376. https://doi.org/10.3109/09540269409023274

Kessler, R. C., Berglund, P., Chiu, W. T., Demler, O., Heeringa, S., Hiripi, E., Jin, R., Pennell, B. E., Walters, E. E., & Zaslavsky, A. (2004). The US National Comorbidity Survey Replication (NCS-R): Design and field procedures. *International Journal of Methods in Psychiatric Research*, *13*(2), 69–92. https://doi.org/10.1002/mpr.167

Kessler, R. C., & Üstün, T. B. (2004). The World Mental Health (WMH) survey initiative version of the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI). *International Journal of Methods in Psychiatric Research*, *13*(2), 93–121. https://doi.org/10.1002/mpr.168

Klerman, G. L. (1986). The National Institute of Mental Health—epidemiologic catchment area (NIMH-ECA) program. *Social Psychiatry*, *21*(4), 159–166. https://doi.org/10.1007/BF00583995

Krueger, R. F., & Finger, M. S. (2001). Using item response theory to understand comorbidity among anxiety and unipolar mood disorders. *Psychological Assessment*, *13*(1), Article 140. https://doi.org/10.1037/1040-3590.13.1.140

Kupfer, D. J., Regier, D. A., & Kuhl, E. A. (2008). On the road to DSM-V and ICD-11. *European Archives of Psychiatry and Clinical Neuroscience*, *258*(S5), 2–6. https://doi.org/10.1007/s00406-008-5002-6

Levis, B., Benedetti, A., Ioannidis, J. P., Sun, Y., Negeri, Z., He, C., Wu, Y., Krishnan, A., Bhandari, P. M., & Neupane, D. (2020). Patient Health Questionnaire-9 scores do not accurately estimate depression prevalence: Individual participant data meta-analysis. *Journal of Clinical Epidemiology*, *122*, 115–128.e1. https://doi.org/10.1016/j.jclinepi.2020.02.002

Little, R., & Rubin, D. (1987). *Statistical analysis with missing data* (15). Wiley.

Little, T. D., Jorgensen, T. D., Lang, K. M., & Moore, E. W. G. (2014). On the joys of missing data. *Journal of Pediatric Psychology*, *39*(2), 151–162. https://doi.org/10.1093/jpepsy/jst048

Little, T. D., & Rhemtulla, M. (2013). Planned missing data designs for developmental researchers. *Child Development Perspectives*, *7*(4), 199–204. https://doi.org/10.1111/cdep.12043

Liu, Y., & De, A. (2015). Multiple imputation by fully conditional specification for dealing with missing data in a large epidemiologic study. *International Journal of Statistics in Medical Research*, *4*(3), 287–295. https://doi.org/10.6000/1929-6029.2015.04.03.7

Lorenzo-Seva, U., & Ferrando, P. J. (2021). Not positive definite correlation matrices in exploratory item factor analysis: Causes, consequences and a proposed solution. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(1), 138–147. https://doi.org/10.1080/10705511.2020.1735393

Patten, S. B. (2003). Recall bias and major depression lifetime prevalence. *Social Psychiatry and Psychiatric Epidemiology*, *38*(6), 290–296. https://doi.org/10.1007/s00127-003-0649-9

Peytchev, A., & Peytcheva, E. (2017). Reduction of measurement error due to survey length: Evaluation of the split questionnaire design approach. *Survey Research Methods*, *11*(4), 361–368. https://doi.org/10.18148/srm/2017.v11i4.7145

Pierotti, M. E., Martín-Fernández, J. A., & Barceló-Vidal, C. (2017). The peril of proportions: Robust niche indices for categorical data. *Methods in Ecology and Evolution*, *8*(2), 223–231. https://doi.org/10.1111/2041-210X.12656

Raghunathan, T. E., & Grizzle, J. E. (1995). A split questionnaire survey design. *Journal of the American Statistical Association*, *90*(429), 54–63. https://doi.org/10.1080/01621459.1995.10476488

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology*, *5*(1), 27–48. https://doi.org/10.1146/annurev.clinpsy.032408.153553

Rhemtulla, M., & Little, T. D. (2012). Planned missing data designs for research in cognitive development. *Journal of Cognition and Development*, *13*(4), 425–438. https://doi.org/10.1080/15248372.2012.717340

Rioux, C., Lewin, A., Odejimi, O. A., & Little, T. D. (2020). Reflection on modern methods: Planned missing data designs for epidemiological research. *International Journal of Epidemiology*, *49*(5), 1702–1711. https://doi.org/10.1093/ije/dyaa042

Robins, L. N., & Cottler, L. B. (2004). Making a structured psychiatric diagnostic interview faithful to the nomenclature. *American Journal of Epidemiology*, *160*(8), 808–813. https://doi.org/10.1093/aje/kwh283

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Saito, M., Iwata, N., Kawakami, N., Matsuyama, Y., Ono, Y., Nakane, Y., Nakamura, Y., Tachimori, H., Uda, H., & Nakane, H. (2010). Evaluation of the DSM-IV and ICD-10 criteria for depressive disorders in a community population in Japan using item response theory. *International Journal of Methods in Psychiatric Research*, *19*(4), 211–222. https://doi.org/10.1002/mpr.320

Shoemaker, D. M. (1973). *Principles and procedures of multiple matrix sampling*. Ballinger.

Spitzer, R., Williams, J., Gibbon, M., & First, M. (1987). *Structured clinical interview for DSM-III-R (SCID-P, 4/1/87)—patient version* (Vol. 11). New York State Psychiatric Institute.

Spitzer, R. L., Williams, J. B., Gibbon, M., & First, M. B. (1992). The structured clinical interview for DSM-III-R (SCID): I: History, rationale, and description. *Archives of General Psychiatry*, *49*(8), 624–629. https://doi.org/10.1001/archpsyc.1992.01820080032005

StataCorp. (2017). *Stata data analysis statistical software: Release 15*. StataCorp LP.

Steinberg, M. (1994). *Interviewer's guide to the structured clinical interview for DSM-IV dissociative disorders (SCID-D)*. American Psychiatric Association.

Uher, R., Payne, J. L., Pavlova, B., & Perlis, R. H. (2014). Major depressive disorder in DSM-5: Implications for clinical practice and research of changes from DSM-IV. *Depression and Anxiety*, *31*(6), 459–471. https://doi.org/10.1002/da.22217

Van der Heijden, G. J., Donders, A. R. T., Stijnen, T., & Moons, K. G. (2006). Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: A clinical example. *Journal of Clinical Epidemiology*, *59*(10), 1102–1109. https://doi.org/10.1016/j.jclinepi.2006.01.015

Vriens, M., Wedel, M., & Sándor, Z. (2001). Split-questionnaire designs: A new tool in survey design and panel management. *Marketing Research*, *13*(2), 14–19.

Wothke, W. (1993). Nonpositive definite matrices in structural modeling. In K. Bollen, & J. S. Long (Eds.), *Testing structural equation models* (pp. 256–293). Sage.

Zbozinek, T. D., Rose, R. D., Wolitzky-Taylor, K. B., Sherbourne, C., Sullivan, G., Stein, M. B., Roy-Byrne, P. P., & Craske, M. G. (2012). Diagnostic overlap of generalized anxiety disorder and major depressive disorder in a primary care sample. *Depression and Anxiety*, *29*(12), 1065–1071. https://doi.org/10.1002/da.22026.

Received November 9, 2021
Revision received October 24, 2022
Accepted November 3, 2022 ■