Despite several decades of research, there remains disagreement as to whether there are unique differences in depression symptom expression between those with bipolar disorder and those with major depressive disorder. Discrepancy across studies may be attributed to many sources, such as heterogeneity across patient samples and research settings, variability in selection of symptoms to be evaluated, and lack of control for underlying depression symptom severity between patient groups. In this article, we briefly review the literature on differential symptom expression between bipolar and unipolar depression, and introduce item response theory (IRT) as one methodology that may be particularly useful in overcoming some of the challenges that have plagued existing research. We review some findings from our research program using an IRT approach to evaluate differences between bipolar and unipolar depression, and conclude with recommendations for future research in this area.
Bipolar Disorder

understanding them will aid in differential diagnosis and increase the likelihood that a patient will receive appropriate pharmacological and psychosocial treatments. Moreover, such differences, or lack thereof, may also be informative when adapting interventions for bipolar disorder from existing treatments for MDD (e.g. cognitive behavioral therapy).26

Is Bipolar Depression a ‘Different’ Depression?

Despite the growing literature focused on bipolar depression, there has been no clear consensus as to whether it can be differentiated from unipolar depression on the basis of symptoms alone. Some have reported a greater prevalence of atypical features in bipolar versus unipolar depression.22,23 However, others have failed to replicate these findings,26 and there is some evidence that bipolar depression might actually be marked by a greater prevalence of melancholic symptoms.28 Additional research suggests that individuals with unipolar depression may tend to endorse higher rates of anxiety and somatization in comparison with those with bipolar depression.27 However, compared with bipolar II depression alone, others have reported lower rates of anxiety and agitation in unipolar samples.20,28,29 Finally, there is some evidence that bipolar depression may be characterized by greater rates of psychosis than unipolar depression,21,26 although this finding may be limited to bipolar I depression only.26

Although suicidal ideation and behaviors may best be conceptualized as transdiagnostic,34 they are also Diagnostic and Statistical Manual of Mental Disorders, 4th Edition (DSM-IV) symptoms of a major depressive episode. Within this context, there is disagreement in the literature as to whether there are differences between bipolar and unipolar depression in rates of endorsement of suicidal ideation and behaviors. For example, in their review of the literature, Rihmer and Kiss concluded that “bipolar patients in general, and bipolar II subjects in particular, carry the highest risk of suicide.” However, published data do not fully support this assertion. Indeed, some33–35 have reported greater suicide risk in bipolar (I or II) disorder versus MDD, whereas others have reported the opposite effect, with greater suicide risk in MDD.41–43 Still others have failed to identify any diagnostic difference in levels of suicidal ideation, number of suicide attempts, or intent to die.44–46 With respect to bipolar II disorder specifically, some investigators have reported higher lifetime history of suicide attempts in bipolar II versus bipolar I disorder,47 whereas several studies have failed to demonstrate greater suicide risk in bipolar II relative to bipolar I disorder, with no differences reported between subtypes on measures of family history of suicide, suicidal ideation, or suicide attempts.48–50

Finally, in contrast to the literature that has emphasized potential diagnostic differences in depression symptom expression, some recent comprehensive reviews have concluded that bipolar and unipolar depression are more similar than different,51,52 and may not even be differentiable.53 In their two-illness model of bipolar disorder, Joffe et al.54 have suggested that bipolar illness may consist of two separate but inter-related disorders, mania and depression, the latter of which may be “no different from the broad range of depressive disorders that constitute unipolar depression.” In sum, data supporting a unique presentation of depression in bipolar disorder versus MDD remain equivocal, and there is a need for future research to address potential limitations of the extant literature in order to further clarify any differential phenomenology between the two conditions.

Limitations of the Extant Literature

Indeed, the mixed findings reviewed above may be attributed to methodological limitations of prior research. First, the large majority of studies have relied on clinical versus community samples of individuals, some of whom were recruited from inpatient settings or had been enrolled in clinical efficacy trials that use narrow inclusion criteria.22 In addition, with few exceptions, sample sizes have generally been small, thus rendering statistical parameters unstable.27 Another source of inconsistency across studies may be attributed to type I error, as most studies have performed a large number of comparisons without corresponding alpha-level corrections, thereby increasing the risk of falsely interpreting chance observations as indicative of reliable cross-diagnostic differences.

It should also be noted that several studies have combined individuals with bipolar I and bipolar II depression for purposes of comparison against unipolar depression,22,23,33,41 which may account for inconsistencies when bipolar subtypes have been evaluated independently of one another. In addition, the majority of this research has focused narrowly on atypical or other feature specifiers, and has not comprehensively addressed potential differences across all DSM-IV depressive symptoms. Although such clinical feature specifiers may, indeed, be useful in differentiating between bipolar and unipolar depression, such a focus limits an understanding of how the core symptoms of depression operate in these two disorders.

Perhaps most importantly, several studies have not controlled for overall symptom severity in their comparisons across groups.27 It is therefore unclear whether any differential symptom expression reported in the literature is due to true phenomenological differences between bipolar and unipolar depression, or whether such differences are instead reflective of greater overall depression severity in one group versus another. This latter point is especially critical in the context of heterogeneity of sample selection in the published literature (i.e. inpatient versus outpatient, clinical versus community), and in light of findings that both atypical features and bipolar II disorder may be associated with greater depression severity. For example, the higher risk of suicide in bipolar II versus bipolar I disorder that has been reported in some research may be better accounted for by greater depression severity in the bipolar II samples that were evaluated. Consistent with this argument, Cooke et al. reported no differences between MDD, bipolar I, and bipolar II disorders on a number of depression course characteristics when comorbidity (i.e. a proxy for illness severity) was accounted for in the analysis.

Given the limitations reviewed above, we have argued that methods based on item response theory (IRT) might be particularly useful in overcoming some of the challenges that have plagued existing research comparing bipolar and unipolar depression. Although a comprehensive overview of IRT is beyond the scope of the current review, we briefly describe this methodology in the paragraphs that follow and review its potential strengths and limitations as applied to the study of diagnostic group differences in the expression of depressive symptoms. We also
review findings from our research program evaluating such differences using an IRT approach, and conclude with recommendations for future research in this area.

**Item Response Theory**

Initially derived from the testing and assessment literature, IRT\(^{50}\) is based on the notion that an individual will provide a particular response to an item or symptom query given her/his underlying severity on the latent trait being measured. As applied to our research, for example, we have relied on IRT modeling to examine the likelihood that a respondent will endorse a particular symptom of depression (i.e. an item response) given her/his particular level of depression severity (i.e. the latent trait).\(^{51-53}\) This approach departs from models in which levels of the latent trait are defined primarily by the total number of positive responses to a specific set of symptom questions, and instead frees individual symptoms from any prescribed sets.

Given this focus on the symptom rather than the scale, IRT facilitates a common metric across many measures of the same underlying construct, and provides an opportunity for increasingly efficient measurement. These innovations have been leveraged into computer adaptive testing,\(^{48}\) in which the same latent construct can be assessed reliably and validly with wholly distinct sets of symptoms. An additional advantage of anchoring to a common latent trait is the capacity to evaluate qualitative differences in symptom reports from different subgroups with the same level of the latent trait (e.g. depression severity). For example, IRT as applied to the evaluation of differential item functioning (DIF)\(^{56}\) can be used to examine differences in symptom expression in bipolar versus unipolar depression.

**Item Response Theory Assumptions**

IRT models estimate the parameters of a mathematical function, typically logistic in nature, to determine the probability of a particular item response at a particular level of the latent trait. For dichotomous items such as symptom queries, this probability is modeled using the item response curve (IRC). One primary assumption of IRT is that responses to items or symptom queries are a function of individual variation along a single underlying dimension. As applied to the latent trait of depression severity, for example, this assumption is meaningful for both theoretical and statistical reasons. Theoretically, DSM-IV stipulates that symptoms are summed to determine the presence or absence of a depression diagnosis. In so doing, DSM-IV assumes that responses are linked to a single construct of depression severity.\(^{57}\) Statistically, information regarding symptom functioning may be biased if a unidimensional item response model is applied to multidimensional data.\(^{58}\) However, continued innovations in the application of item response models to constructs with multiple dimensions are evolving.\(^{59}\)

An additional assumption of IRT models is that items are locally independent; that is, items must not be correlated for reasons other than measurement of the latent trait.\(^{60}\) For clinical phenomena such as depressive symptoms, this assumption may pose some unique challenges. Indeed, for the depressive symptoms that comprise appetite/weight disturbance, sleep disturbance, and psychomotor disturbance, one could reliably predict the absence of one symptom (e.g. insomnia) from the presence of the other (e.g. hypersomnia), irrespective of depression severity. Thus, when considering the application of IRT models to evaluate psychiatric symptom functioning, it is important for investigators to consider potential model assumption violations and to carefully design models within the constraints of specified IRT assumptions. It should be noted, however, that newer IRT models that can accommodate local dependence are also being developed and tested.\(^{61}\)

**The Two-parameter Logistic Model**

Although there are several item response models one could evaluate, the two-parameter logistic (2PL) model for dichotomous items has been commonly used to answer clinical questions.\(^{62}\) This model estimates a severity parameter to describe the point on the latent trait continuum where a symptom becomes likely to be observed (e.g. >50%) and a discrimination parameter to describe how rapidly the probability of observing the symptom changes across increasing levels of the latent trait continuum (e.g. the slope of the IRC). To answer questions regarding patterns of symptom endorsement, interpretation of the severity parameter may be considered most relevant, as it is reflective of the likelihood that a given symptom will occur at a given severity level. However, the discrimination parameter is important in that it can be used to verify that a given symptom is a good indicator of the underlying latent trait. For example, if a symptom is equally likely to be endorsed at high and low levels of depression severity, it would not appear to index depression in a clinically meaningful way.

**Differential Item Functioning**

In order to evaluate potential group differences in the likelihood of endorsing a given symptom across different levels of the latent trait, IRT methods can be applied to the estimation of DIF. DIF involves comparing a series of analyses that isolate and compare each item parameter (i.e. severity and discrimination) across each identified group.\(^{63,64}\) If the symptoms function similarly across groups, the parameters that describe the symptoms will be estimated similarly in different samples.

There is a diversity of DIF detection methods, and interested readers are referred to Holland and Wainer\(^{65}\) for a review of this methodology. In our research, we have followed Thissen and colleagues,\(^{66}\) who recommend the use of a likelihood-ratio test statistic to provide a significance test for the null hypothesis that the item parameters do not differ between identified groups. Analyses proceed by comparing the goodness of fit between nested models—one in which the parameters of a given symptom are constrained to be equal across groups, and one in which they are not. The difference in the log-likelihoods (\(\Delta LL\)) of the two models provides an omnibus test of whether there is DIF for the discrimination and/or severity estimate for this symptom. If significant, follow-up tests can be conducted to identify whether DIF is present in discrimination or severity estimates by further constraining models and interpreting the resultant common metric.

**Why Apply Item Response Theory to the Study of Bipolar Depression?**

Given the limitations of the literature focused on identification of differences between bipolar and unipolar depression, we believe that IRT carries a number of strengths that might help advance this line of inquiry. The primary strength of this approach over other statistical
Bipolar Disorder

Figure 1: Differential Item Functioning for Suicidal Ideation/Attempt Between Bipolar I and Unipolar Depression (A) and Between Bipolar I and Bipolar II Depression (B)

methodologies is that it allows one to examine the likelihood that a particular symptom will be endorsed at a particular level of depression severity. Thus, differences in symptom endorsement between groups can be evaluated while simultaneously equating for individual levels of depression symptom severity. Given that the majority of research comparing bipolar with unipolar depression has not equated groups on underlying depressive severity, and that several of the differences identified in the literature (e.g. atypical features, suicide risk) may be directly linked to severity, this benefit of IRT overcomes a particularly important limitation of prior research in this area. IRT also relies on large samples that represent the full range of the latent trait. Although this stipulation could be considered a limitation, in that IRT may not be as applicable to smaller clinical samples recruited from specialty settings, it could also be considered a strength in that it allows investigators to extend research focused on differences between bipolar and unipolar depression to large, representative community samples.

Another important consideration when applying IRT methodologies is that in such large samples, results that emerge as statistically significant may not be clinically significant. As such, it is important for investigators to consult published recommendations for effect size interpretation that are especially useful in determining the clinical applicability of findings. In our analyses, we decided a priori what we believed to be a minimally clinically significant effect size. Finally, as mentioned above, there are particular statistical assumptions for IRT that must be met prior to analysis. For the study of depressive symptoms, there is sufficient evidence in the literature that depression can be conceptualized as a unidimensional construct. Application of traditional IRT methodologies to the study of manic symptom expression, however, may be more limited as there is less support for a unidimensional model of mania. In addition, there may be some limits to the study of depression using IRT methods with regard to the assumption of local independence of symptoms, as described above. To address this concern, we have evaluated compound items of appetite/weight disturbance, sleep disturbance, and psychomotor disturbance, in contrast to their component parts, in our research. Although this approach is in keeping with assignment of a diagnosis of a major depressive episode, as per DSM-IV, one trade-off in adopting such an approach is that conclusions regarding depressive subtypes (e.g. atypical or melancholic features) cannot be formed.

What Have We Learned (Thus Far) About Differences Between Bipolar and Unipolar Depression Using Item Response Theory?

We have recently applied IRT methodologies to evaluate differences in DSM-IV depression symptom endorsement across three groups of individuals who participated in the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC). Those reporting depressive symptoms and a history of mania (i.e. bipolar I depression), those reporting depressive symptoms and a history of hypomania (i.e. bipolar II depression), and those reporting depressive symptoms without a history of hypomania or mania (i.e. unipolar depression). To date, the NESARC represents the largest epidemiological survey of psychiatric conditions in the US, and therefore provided a rich opportunity to evaluate DIF across these groups. Methods for obtaining and assessing the sample have been detailed elsewhere.

Consistent with our review above, we made several decisions prior to our analyses. First, we accounted for the assumption of local independence by evaluating the compound items of appetite/weight, sleep, and psychomotor disturbance, rather than their component parts. Second, we decided a priori what we believed to be a minimally clinically significant effect size for interpretation, following guidelines provided by Steinberg and Thissen. Third, we employed the Benjamini-Hochberg procedure to adjust p-values in order to account for risk of type I error due to multiple comparisons.
Results from analyses revealed that, by and large, groups were more similar than different in their endorsement of DSM-IV depression symptoms. Of the few differences that did emerge, individuals with bipolar depression were more likely than those with unipolar depression to endorse psychomotor disturbance, and fatigue was more likely to be endorsed in unipolar depression but only at moderate (versus low or high) levels of depression severity. There was no significant DIF across groups on any of the remaining symptoms, with the exception of suicidal ideation/attempt. Results from DIF analyses revealed that individuals with bipolar depression were more likely to endorse suicidal ideation/attempt than both the bipolar II or unipolar groups\textsuperscript{52,53} (see Figure 7). However, there were no differences between bipolar II and unipolar depression in the likelihood of endorsing suicidal ideation/attempt. These findings run counter to arguments that suicide risk is highest in bipolar II depression—a discrepancy that may be accounted for by the fact that IRT adjusts for overall depression severity, whereas most prior research comparing suicide risk in mood disorders has not.\textsuperscript{54} Nevertheless, it is notable that, at an average level of depression severity (i.e. latent trait = 0), the probability of endorsing suicidal ideation was quite high in all groups, ranging from 30 to 76\%, which suggests that this area of inquiry warrants continued clinical and empirical attention.

Conclusions and Future Directions

Given the great heterogeneity of depression symptom expression, especially across diverse patient samples and settings, it has been difficult to interpret and synthesize findings from the many studies that evaluate differences in symptom endorsement between bipolar and unipolar depression. By equating for depression severity and anchoring evaluation of DIF to the latent trait of depression as opposed to a specific scale or set of questions, we argue that IRT provides a number of advantages that can be applied toward advancing our understanding of the unique versus shared aspects of depression in bipolar disorder and MDD. To date, our study of bipolar disorder remains in its early stages, and there are a number of areas in which IRT may further advance an understanding of bipolar disorder phenomenology and assessment. We recommend below just a few of the myriad potential directions for future research.

With the advent of newer IRT methods that can accommodate local dependence of items,\textsuperscript{56} there are several areas of inquiry that deserve increased attention. Given the continued interest in depression feature specifiers, it is recommended that future research focus on additional symptoms of depression that have not yet been evaluated using IRT (e.g. mood reactivity, hypersomnia versus insomnia). In addition, given our data concerning the differential likelihood of endorsing suicidal ideation/attempt across mood disorder diagnoses, it would be informative to expand this area of inquiry and to apply these newer IRT techniques to evaluation of group differences in the endorsement of this item’s components (i.e. passive thoughts of death, suicidal ideation with plan, suicide attempt). Such a fine-grained analysis would further clarify the nature of suicide risk across bipolar and unipolar depression.

Also of interest is the evolution of newer IRT methods that can be applied to multidimensional constructs, such as mania. The expansion of IRT to the study of mania might be especially useful in the context of understanding symptom expression in pure versus mixed mania, for example, and researchers are just beginning to evaluate the phenomenology of manic symptom presentations using these methods.\textsuperscript{21} Finally, hearkening back to the origins of IRT, investigators have recently begun to apply IRT methodologies to the development and refinement of assessment measures for bipolar disorder.\textsuperscript{22,23} Given the difficulty in assigning a proper bipolar disorder diagnosis, and the limitations of existing measures,\textsuperscript{24,25} this area holds great promise. In particular, applications that rely on computer adaptive testing\textsuperscript{25} may be particularly useful in providing accurate diagnosis, thus hastening proper treatment and improved outcomes for those suffering from major affective disorders.
Bipolar Disorder

52. Rejas J, Azorin J, Suicide attempts: differences between unipolar and bipolar patients and among groups with different lethality risk, J Affect Disord, 2004;82:437–42.
69. Rejas J, Azorin J, Suicide attempts: differences between unipolar and bipolar patients and among groups with different lethality risk, J Affect Disord, 2004;82:437–42.