ABSTRACT—A growing literature indicates that temperamental emotionality traits may represent diatheses for mood disorders. This article describes the advantages of laboratory methods for assessing these traits, reviews issues regarding their use, and argues that these techniques provide unique opportunities for advancing understanding of temperamental risk.

KEYWORDS—temperament; depression; laboratory paradigms

Consistent with a rich theoretical tradition (e.g., Kraeplin, 1921; Meehl, 1987), there is growing evidence that temperament traits relevant to emotion represent early emerging diatheses for mood disorders. Depressed individuals and those at risk for mood disorders are distinguished from those at lower risk by their levels of two traits, positive emotionality (PE) and negative emotionality (NE), defined, respectively, by the propensity to respond to stimuli with positive or negative emotions, as well as by the intensity and frequency of those states (Clark & Watson, 2008). Low PE and high NE have been linked to depression in both adults and children (e.g., Brown, Chorpita, & Barlow, 1998; Kendler, Neale, Kessler, Heath, & Eaves, 1993). Three prospective studies also implicate childhood PE and NE in the development of depression. Block, Gjerde, and Block (1991) reported that low PE in childhood predicted depressive symptoms at age 18. Caspi, Moffitt, Newman, and Silva (1996) found that observational ratings of low PE and high NE in preschoolers predicted depressive disorders in young adulthood. In a cohort study, van Os, Jones, Lewis, Wadsworth, and Murray (1997) found that physicians’ ratings of low PE in childhood predicted chronic and recurrent mood disorders in adulthood.

This developmental literature provides support for models that alternatively propose that temperament traits (a) are precursors of depression, (b) are predisposing factors causally implicated in its development, or (c) emerge from causal processes similar to those that produce depression (Klein, Durbin, & Shankman, 2009). Distinguishing among these models represents a significant challenge that requires a thoughtful approach to several developmental issues, both substantive and methodological, as well as the use of longitudinal, etiologically informative research designs. First, a coherent model of how these traits are related to depression must provide a bridge between literatures across the life span, specifying if and how continuity of traits and trait–disorder associations exist. Second, a convincing model must explain the conceptual relations between temperament and depression (mapping areas of overlap and distinction), as well as their causal relations (to one another and to other etiological factors). Third, these models must be tested using multiple strategies to demonstrate generalizability and provide a comprehensive mapping of the underlying processes linking temperament to depression. Valid measurement strategies undergird the ability to address each of these substantive issues. Measurement approaches must be developed that (a) enable comparison of findings across development, (b) tap constructs at a degree of refinement closest to the level at which their causal processes operate, and (c) measure temperament constructs most strongly and uniquely related to depression in order to explore specificity. Below, I outline one measurement strategy that holds considerable promise in this regard—laboratory assessment of temperamental emotionality.

LABORATORY-ASSESSED TEMPERAMENT AND DEPRESSION

A growing literature shows that laboratory assessments of child PE and NE have predictive validity for depression risk.
Behavioral inhibition, a trait with elements of both low PE and high NE, has been linked to maternal history of depression and anxiety (e.g., Rosenbaum et al., 2000). Low PE and high NE in preschoolers have been linked to maternal mood disorders (Durbin, Klein, Hayden, Buckley, & Moerk, 2005), EEG asymmetries (Shankman et al., 2005), the development of depressotypic cognitions at age 7 (Hayden, Klein, Durbin, & Olino, 2006), internalizing symptoms at age 10 (Dougherty, Klein, Durbin, Hayden, & Olino, 2010), and serotonin transporter gene polymorphisms (Hayden et al., 2007). Moreover, lab-assessed PE and NE show stability over intervals ranging from weeks to years (e.g., Durbin, Hayden, Klein, & Olino, 2007; Rothbart, Derryberry, & Hershey, 2000).

These findings complement the impressive body of research on normal development that employed laboratory assessments of temperament. Lab methods have revealed associations between temperament and physiology (Fox, Henderson, Marshall, Nichols, & Ghera, 2005), moral development (Kochanska, Murray, Jacques, Koenig, & Vandengeest, 1996), and the parent–child relationship (Kochanska, Aksan, & Carlson, 2005). Thus, laboratory assessments reveal associations between temperament and both normal and abnormal processes, highlighting their usefulness for developmental psychopathology research. The next stage of research should capitalize on this promise by addressing psychometric issues, developmental considerations, and the components of temperamental emotionality uniquely measured by laboratory tasks.

LABORATORY METHODS AS A COMPLEMENT TO OTHER APPROACHES

Laboratory temperament batteries have been developed for exploring basic issues in emotional development (e.g., Goldsmith, Reilly, Lemery, Longley, & Prescott, 1995). They emerged alongside research aimed at mapping the structure of temperament as measured by parent reports (e.g., Rothbart, Ahadi, Hershey, & Fisher, 2001). There has been long debate over the relative merits of parent report and laboratory methods (see Kagan & Fox, 2006; Rothbart & Bates, 2006). Importantly, parent questionnaires provide a bridge to research on temperament–psychopathology relationships in adults, which has been conducted primarily with self-report questionnaires. As for many domains, one would ideally utilize a multimethod strategy, which produces more reliable measures, allows for a broader mapping of constructs that may manifest differently across contexts, and is amenable to statistical procedures (such as structural equation modeling [SEM]) that quantify shared variance across methods (consistent with the theoretical claim that temperament captures behavioral styles that are pervasive across situations).

The success of a multimethod approach for any particular domain depends on each method's measurement properties, the degree to which the methods converge, and whether factors related to the substantive issue of interest produce divergent findings across methods. For temperament traits relevant to depression, there are known challenges to latent variable approaches combining parent reports and other measures. First, parent temperament questionnaires have low convergence with laboratory measures, nature in observations, and teacher reports (e.g., Seifer, Sameroff, Barrett, & Krafchuk, 1994). Second, risk indices for depression may be particularly prone to weaker cross-method convergence. Agreement across informants (i.e., child, parent, teacher) is, at best, moderate for psychopathology measures, with weaker convergence for internalizing problems than for externalizing problems and decreasing parent–child agreement as child age increases (Achenbach, McConaughy, & Howell, 1987). To the extent that dysphoria-related biases in parent reports exist (Youngstrom, Izzard, & Ackerman, 1999), they may be amplified in studies of depression risk, in which the reporting parents are likely to suffer from depression themselves. Depressed parents, for example, have a lower threshold for reporting depression in their children, resulting in higher rates of both true and false positives (Najman et al., 2000; Richters, 1992). Thus, the most sensible approach is to employ both laboratory and parent-report measures, to explore their shared variance in predicting depression-risk indices (using SEM), and also to test whether one method uniquely contributes to the prediction of the outcome variables of interests.

Laboratory approaches, while more costly than some other methods, offer unique advantages for exploring issues related to depression. Laboratory tasks are more effective at disentangling child and contextual influences that are confounded when measures are taken in settings that exert powerful constraints on behavior (such as the home). Using standardized stimuli to elicit trait-relevant behaviors allows for individual differences in child behavior to be more sharply differentiated from the situations in which traits are expressed. Also, lab approaches can expose children to contexts that readily elicit traits that are expressed at a lower base rate in naturalistic settings (e.g., fear) or that parents may strive to reduce or be less willing to report (e.g., anger). It should be noted, however, that the laboratory itself is a context, one that differs (intentionally) from other contexts in meaningful ways. For example, inhibited children are particularly responsive to the novelty of the laboratory setting; children high in inhibition exhibit reduced levels of approach behavior in the lab, but only during tasks that involve novelty or fear contexts (Laptook et al., 2008). Traits measured in different contexts may be differentially predictive of important outcomes. For example, Kochanska, Aksan, Penney, and Doobay (2007) demonstrated that child PE related to effortful control indices differently depending upon the measurement context: PE measured by structured lab tasks was negatively related to effortful control, but PE measured from naturalistic observations of mother–child interaction was positively related to effortful control. Thus, exploring areas of divergence across multiple methods can be as revealing as exploring their convergent findings.
CONSTRUCT VALIDITY

Because lab tasks elicit observable behaviors, it can be tempting to rely on face validity in evaluating them. However, interpreting responses to a lab task as evidence for a particular trait depends upon (a) the degree to which the task elicits individual differences (not merely high mean levels) in trait-relevant behaviors, (b) the salience of the behaviors to the trait, (c) whether the behaviors emerge primarily from that trait or could be multiply determined by several traits, and (d) the extent to which the behaviors predict responses to analogous situations outside the laboratory and to the larger category of stimuli relevant to individual differences in that trait. A more systematic approach to construct validation of laboratory tasks is important for mapping their associations with external correlates, including risk for depression. It is crucial to quantify the degree to which tasks tap traits of interest and near-neighbor constructs; until this is known, it is important to code multiple traits in each task to determine which are, in fact, best measured by the battery. Evidence that laboratory-assessed traits predict measures from other contexts (i.e., naturalistic observations, parent reports) is a vital source of evidence regarding their validity.

PSYCHOMETRICS, PROCEDURES, AND MEASURING TRAITS RELEVANT TO DEPRESSION RISK

Less attention has been paid to the issues of reliability and validity for lab tasks than for other temperament measures. For example, researchers have sometimes used a single task, making it impossible to assess reliability and raising concerns regarding the generalizability of the resulting trait estimates, since they may be equally (or more) influenced by features of the task context than by the trait. Designs using a single task at different assessment points cannot be used to estimate the stability of temperament for these reasons. Using multiple tasks and aggregate measures lessens this concern, but the extent to which all tasks tap the same trait must be examined rather than assumed. In addition to internal consistency reliability, test–retest reliability is also important, particularly since temperament is conceptualized as temporally stable. Because of the time and expense that laboratory studies represent for both researchers and participants, it can be difficult to collect multiple waves to assess stability, although some investigators have reported such data (e.g., Durbin et al., 2007; Rohlsb...). More problematic is that with repeated administration, tasks that involve aspects such as deception or novel stimuli may take on different meanings for participants, making the tasks unsuitable for stability studies. One solution is to use similar yet distinct tasks at each assessment, but stability estimates then reflect not only trait stability and task reliability but also differences in the validity of each task for tapping the trait.

One appealing feature of laboratory tasks is that they can be easily modified, giving researchers flexibility for assessing constructs of interest in ways that are developmentally sensitive for the participant group of interest. A variety of techniques can be used to elicit individual differences in temperament traits, which can be measured using coding of behavior, children’s subjective reports in response to the tasks, and direct performance measures. However, this approach can result in considerable variability across labs in the tasks employed, in the administration of similar tasks, and in coding systems. Thus, laboratory tasks are less comparable across research groups than are other methods, necessitating greater communication across researchers.

A number of suggestions can be made for utilizing laboratory methods to maximize valid assessment of the temperamental processes relevant to depression. As reviewed above, the best replicated temperamental correlates of risk for depression are emotionality traits (PE and NE). Importantly, low PE may be more specific to depression than high NE (Clark, Watson, & Mineka, 1994), and therefore, any study of depression risk should measure both traits. Different considerations are warranted for PE and NE. First, tasks that elicit negative emotion should be followed by breaks between tasks sufficient to allow negative emotions to diminish; they should also be interspersed with positive tasks to avoid carryover effects. Second, reliably measuring PE requires the use of fewer tasks specifically designed to tap positive emotions, since PE measured in negative tasks tends to correlate highly with PE measured in positive tasks (e.g., Durbin et al., 2005), indicating robustness of individual differences in this trait across task type. Third, the level of specificity with which these traits can be measured (as broadband PE and NE composites or more narrow specific emotion variables) may differ for PE and NE. Since there has been less empirical work demonstrating differentiation among positive emotions than there has been demonstrating differentiation among negative emotions, aggregation across negative emotions is more often an issue. Because coding of emotions relies on observable behaviors (facial, vocal, and bodily expressions), states defined by clear morphological differences (as are sadness, anger, and fear) can be more reliably discriminated than can different positive emotions (e.g., exuberance, contentment), which may be more identified by quantitative intensity differences than by qualitative differences.

Regardless of the level at which emotionality constructs are analyzed, it is important to sample affective experience broadly in order to capture processes relevant to risk for depression and to allow examination of whether associations between temperament and depression risk are specific to narrower emotion constructs (e.g., sadness) or involve broader composites (e.g., NE). One approach is to use tasks that present challenges thought particularly relevant for depression risk, such as those involving disappointment, rejection, or negative performance feedback. Some have proposed that depression may be particularly associated with deficits in anticipatory positive affect (Davidson, Pizzagalli, Nitschke, & Putnam, 2002). Thus, one
might want to include tasks designed to tap this state (e.g., imposing a waiting period between the child’s selection and receipt of a highly valued reward). Another strategy is to tap from all quadrants of the arousal–valence model of emotion (Barrett & Russell, 1999), including both high and low arousal states. With regard to negative emotions, although sadness bears the most direct overlap with depression, the existing literature derived from laboratory emotion-eliciting tasks does not strongly suggest that sadness is uniquely associated with depression risk. It is clear that laboratory measures of fear proneness are also related to depression risk (Rosenbaum et al., 2000), and given that sadness and anger are elicited by similar classes of stimuli (Lewis & Ramsay, 2005), anger may also be important to assess. Practically speaking, it is very difficult to devise tasks that reliably elicit either sadness or anger without also eliciting the other, highlighting both the overlap of negatively valenced emotions and the presence of individual differences in children’s responses to laboratory challenges. Finally, given the conceptual and empirical overlap between some symptoms of depression and emotional reactivity, researchers must consider developmental issues when selecting samples for studying temperamental risk for depression. Laboratory-assessed emotionality may be more clearly interpreted as emerging from individual differences in temperament traits—rather than from frank mood disorders—in younger samples, wherein the prevalence of these disorders is low; for older samples (or those recruited from clinical populations), it may be important to screen participants for the presence of mood disorders before concluding that emotional reactivity in the laboratory is an index of temperament rather than an expression of mood pathology.

**FUTURE DIRECTIONS: CAPITALIZING ON UNIQUE ADVANTAGES OF LABORATORY METHODS TO FURTHER UNDERSTAND TEMPERAMENTAL RISK**

Lab tasks cannot replace parent-report measures, which are economical, ecologically valid, and tap factors that lab measures cannot address (e.g., parent perceptions of the child). Instead, lab measures should be used to capitalize on several particular strengths of this approach: exploring components of temperament that are more obscure to questionnaire measures, assessing narrower components of traits more readily cleaved by lab tasks than by questionnaires, and asking new questions about individual differences in temperamental risk.

There are at least three promising ways in which laboratory tasks can advance understanding of temperamental risk for depression, the last of which I address in detail. First, because parent-report measures may overestimate trait stability, owing to stable characteristics of the parent (Bates, 1994) or to parents’ desire to maintain a consistent view of their child (Kagan, 1998), the use of laboratory batteries with multiple tasks may provide more stringent estimates of change and stability in temperament over time. This would allow a more accurate exploration of (a) trajectories for temperament-relevant behaviors and their link to depression; (b) discontinuities in temperamental risk, which may mark emergence of, or change in, causal factors implicated in depression; and (c) the question of whether the stable components of temperament provide the most predictive validity for depression. Second, a bridge between the adult and child literatures on temperamental risk could be strengthened by the use of lab batteries assessing adult temperament, such as those used in the German Observational Study of Adult Twins (Spinath, Angleitner, Borkenau, Riemann, & Wolf, 2002). Similarity of methods across development allows for richer developmental comparisons across studies and across generations within family studies.

Finally, lab tasks are uniquely powerful for modeling emotional responsivity to context, providing a more refined measure of individual differences in reactivity to stimuli with positive and negative incentive value. This is particularly important for depression, given the proposal that mood disorders represent abnormalities in the contextual regulation of emotion (Davidson, Jackson, & Kalin, 2000; Rottenberg, Gross, & Gotlib, 2005). The context in which traits are expressed in laboratory tasks can allow inferences to be drawn about (a) a person’s level of that trait, (b) the extent to which the trait is expressed maladaptively, and (c) the possible meaning of the context to the person. Traditional laboratory approaches select tasks designed to tap traits of interest and code only the designated traits in each task; when multiple tasks are employed, codes are aggregated across tasks (collapsing across context differences) in a manner analogous to scoring a trait from multiple items. An alternative approach is to directly model the effects of context by coding all emotions of interest in all tasks, regardless of their a priori emotional target. Temperamental reactivity for each emotion can then be modeled as a function of the contextual “potency” of each task, indexing the manner in which emotional reactivity changes across the entire set of tasks with increasing situational press for that emotion.

This can be accomplished using multilevel modeling (MLM), in which the emotion variable of interest is the dependent variable, nested within children across tasks, and each task is assigned a potency value reflecting the normative ability of that task to elicit the trait (entered at Level 1 of the MLM). Potency values can be determined using several strategies, rank ordering tasks by (a) theory regarding the salience of particular stimuli for the emotion or (b) the amount of that emotion elicited in the sample of interest or in previously collected samples from one’s own or other laboratories. Tasks are assigned separate potency values for each emotion so that each task contributes to estimating each trait; for example, a stranger task would have a high potency value for fear but a low value for happiness. MLM analyses produce estimates of the overall level of the child’s expression of an emotion (the intercept) and the slope of the emotion variable across increasingly potent tasks (an index of temperamental reactivity to contextual press).
Slope and intercept parameters can be fixed (estimated for the study population) or allowed to vary randomly across individuals (random coefficients); the latter quantifies the amount of individual difference variance for each parameter. When random coefficients are significant, parameters can be modeled as dependent variables in a higher order (Level 2) equation to explore whether predictors (e.g., other markers of risk, such as familial loading for depression) can account for individual differences in children's intercept or slope. Level 2 predictors of slopes with a positive sign indicate that, compared with children low on the predictor, children high on the predictor increase expression of the emotion more strongly as task potency for that trait increases (e.g., insecurely attached children increase expressions of sadness more steeply in response to increasingly sad provoking stimuli). Predictors with a negative sign indicate one of two possibilities for children high on the predictor. The first is that they respond less sensitively to increasing potency, requiring a more potent task to reach the same degree of emotional expression as that of a child low on the predictor (e.g., children of depressed parents respond more weakly to increasingly positive tasks than do children of nondepressed parents). The second possibility is that they show similar degrees of the trait across tasks varying in potency (e.g., showing context-inappropriate anger to a task with low anger potency, such as playing with a bubble toy, and the same degree of anger for tasks with higher potency, such as having a preferred toy locked in a box).

The MLM approach provides one means of capturing individual differences in emotionality relative to their context. In addition to modeling reactivity of emotional systems to contexts most salient to that emotion, one could explore specificity of task-emotion reactions and their relevance for depression by testing whether, for example, tasks that are increasingly fear eliciting also pull for increases in other negative emotions, such as sadness. Stranger tasks, for instance, often elicit sadness in addition to (or in the absence of) fearful responses. By extension, more atypical inappropriate reactions can also be modeled, such as responding to increasingly positive tasks with increasing anger. Similarly, modeling degree of positive emotions across tasks that vary in negativity may allow one to tap potential buffering effects of PE on NE, to index individual differences in any dampening effect of negative stimuli on positive emotions, and to assess whether these processes are associated with depression risk.

The MLM approach to modeling emotional reactivity has a number of advantages. First, it builds upon existing use of MLM to account for the nested nature of repeated observations of children across lab tasks. Kiel and Buss (2006) used MLM to examine whether accuracy of mothers' predictions of their child's emotional reactions to lab tasks was moderated by task context (fear-eliciting vs. non-fear-eliciting tasks). Second, it takes into account children's behavior across a wide range of evocative contexts relevant to emotion and allows one to incorporate all data on emotions collected across an entire laboratory battery. Third, it yields conceptually meaningful indices of temperamental emotionality that are more refined than simple averages across tasks (which ignore contextual differences) or ones that model an emotion only in a smaller number of tasks in which it is predicted to occur. However, there are also costs to the approach. Coding multiple emotions across all tasks in a battery increases the time and resources devoted to coding. It requires the use of tasks varying in their potency for particular emotions, necessitating access to normative data for the tasks employed, and possibly the inclusion of additional tasks.

A number of personality theorists have argued for a "taxonomy of situations" (Funder, 2001) and a greater consideration of interactions between traits and the situations in which they are evidenced. Understanding how situational variants are related to the expression of traits and how connections between situations and behaviors vary across individuals is important to understanding temperament and mental health. In summary, owing to their unique advantages, laboratory tasks allow developmental psychopathologists to ask and answer more refined questions about the nature of temperamental risk for depression.

REFERENCES


