CHAPTER 2

Development of the BDEFS

The motivations behind the development of the BDEFS were several. One was the desire to develop a cost-effective means of conveniently capturing the numerous neuropsychological, behavioral, emotional, and motivational symptoms often attributed to deficits in EF. A second, related motivation arose from accumulating evidence that EF tests were not the most ecologically valid means for clinically evaluating EF, and were fraught with other problems that greatly limited their validity and clinical utility as putative measures of EF (Barkley & Fischer, in press; Barkley & Murphy, 2010b; Dimond, 1980; Lezak, 1995).

Problems with EF Tests

Many tests have been declared as measures of EF in the neuropsychological literature—too many to discuss in any detail here (see the meta-analyses of Frazier et al., 2004, and Hervey et al., 2004, for a lengthy list of those EF tests used just in studies of ADHD; see Lezak, Howieson, Loring, Hannay, & Fischer, 2004, for a more comprehensive review). But the virtually unquestioned premise that EF tests are the gold standard for measuring EF can be challenged on a number of grounds:

- *Many of these tests were typically not originally designed to measure EF* (see Lezak et al., 2004). Even a cursory glance at the history of EF tests shows that many were originally developed to assess other psychological functions, such as attention, memory (both verbal and visual–spatial), sequencing, abstract reasoning, and language. Others were intended to assess response inhibition, planning, and problem solving directly, without regard to these constructs' being involved in EF. And most do not

assess frontal lobe (EF) functions exclusively or differentially (Dodrill, 1997). But once the umbrella term of EF was defined as including these constructs, or once it might be shown that deficits on these tests were apparently associated with injuries to the PFC, the tests were conscripted into being EF tests without regard to whether they actually were sampling the conceptual domain or construct of EF.

- *There is no consensus-based definition of EF that can serve as the standard for determining the construct or even face validity of most EF tests.* EF is quite ambiguously defined (see Chapter 1); researchers disagree on the precise meaning of the term (Castellanos, Sonuga-Barke, Milham, & Tannock, 2006; Willcutt et al., 2005). Instead, reviews of the research seem to focus more on listing the constructs thought to be subsumed under the term and the tests believed to evaluate those constructs, such as response inhibition, resistance to distraction, working memory, planning and problem solving, and set shifting (Boonstra, Oosterlaan, Sergeant, & Buitelaar, 2005; Frazier et al., 2004; Hervey et al., 2004; Willcutt et al., 2005).

- *If EF involves the cross-temporal organization and maintenance of behavior directed toward goals, as many seem to believe, it is not immediately apparent how traditional tests of EF using such small ascertainment windows to sample behavior in the clinic (typically 5–30 minutes per test) are sampling the cross-temporal nature of EF implied by these definitions.* Or if they do so, it is only for exceptionally short temporal durations relative to the hours, days, and weeks over which adults sustain their goal-directed activities. This incredibly brief ascertainment window surely must make it difficult if not impossible for EF tests alone to capture the lengthy cross-temporal structures of normal human action. This problem of grossly limited temporal ascertainment by tests alone would ensure that they are weakly related or unrelated to measures of EF taken in naturalistic settings, as by observations over days or by ratings over weeks and months. In contrast, ratings of EF in daily life ascertain behavior across considerably longer periods of time (weeks to months) and therefore serve better than EF tests may do as indicators of cross-temporal behavioral organization and problem solving directed toward goals in naturalistic settings.

- *EF tests do not evaluate some of the most important features of EF.* EF tests simply do not sample many of the capacities believed to be central to the construct of EF, as commented on by some previous authors from Luria (1966) onward, and even in the earliest descriptions of PFC injuries. As noted by Lezak (1995) (see Chapter 1), these include the constructs of volition and human will; intentionality or purposiveness; self-awareness (of self, context, and others); and even aspects of planning (foresight, objectivity, choice and comparative judgment, hierarchical structuring) and plan execution (self-motivation, self-monitoring, prolonged resistance to interference by goal-irrelevant events, etc.).

- *Moreover, given the near-consensus that EF is self-regulation, it is also not obvious how current EF tests can assess self-modification for long-term self-interestedness, self-sufficiency, and social independence—also mentioned by Lezak (1995) and Eslinger (1996), among others, as inherent in the concept of EF.*

- *If EF evolved for principally social functions, such as reciprocity and social exchange, competition, cooperation, and mutualism more generally (again, see Chapter 1), the absence of such purposes and motives in traditional EF tests would further assure their limited utility in predicting EF in daily life activities and in major domains of adaptive functioning.*

• *EF tasks are not only complex but contaminated; they involve multiple cognitive processes, many of which are not considered part of EF.* Only some of those processes are supposedly reflecting the EF construct that is intended to be sampled (Anderson, 2002; Castellanos et al., 2006). A related concern is that many EF tests are often found to be significantly influenced by overall general cognitive ability or level of intelligence (Mahone et al., 2002; Riccio, Hall, Morgan, Hynd, & Gonzalez, 1994), making their results difficult to interpret as reflecting unadulterated measures of a particular EF construct. This is likely to explain findings that statistically removing IQ from relationships between EF tests and observations and ratings of EF in natural settings often reduces any significant relationships to nonsignificant status (Mahone et al., 2002; Mangeot, Armstrong, Colvin, Yeates, & Taylor, 2002). And it may also account for the fact that some of the strongest relationships noted to date have been those between EF tests and academic achievement scores (Biederman, Petty, Fried, Black, et al., 2008; Gropper & Tannock, 2009; Thorell, 2007) or self-ratings of academic performance (Ready, Stierman, & Paulsen, 2001). Given that both academic achievement and self-rated academic performance are significantly related to IQ, and that there is a shared method (testing) when academic tests are used, this finding is not surprising. This problem of conceptual contamination is found far less often for EF rating scales whose items have been intentionally selected to directly sample the various behaviors specified in EF constructs (Barkley & Murphy, 2010b). Such scales or direct observations also have few or no significant relationships with intelligence (Alderman, Burgess, Knight, & Henman, 2003; Barkley & Murphy, 2010b), and so the issue of contamination by general cognitive ability is far less problematic for rating scales than for EF tests. Hence the conceptual or face validity of rating scales may be superior to that of EF tests, merely as a consequence of their initial construction. As noted above, many EF tests were not initially designed to measure the construct of EF.

• *Only a minority of patients experiencing frontal lobe injuries, or those with ADHD presumed to have a frontal lobe disorder, score in the impaired range on EF tests.* In contrast, the vast majority do score in this range on ratings of EF in daily life activities or in direct observations of EF performance in natural settings (Alderman et al., 2003; Barkley & Murphy, 2010b; Burgess, Alderman, Evans, Emslie, & Wilson, 1998; Kertesz, Nadkarni, Davidson, & Thomas, 2000; Mitchell & Miller, 2008; Wood & Liossi, 2006). Given the strong historical linkage of EF to PFC functioning (see Chapter 1), it is unlikely that those patients having disorders of the PFC would not likewise show impairment in EF. In regard to ADHD, however, this situation has led some to argue that ADHD is probably not a disorder of EF, given that the majority of cases have no deficits on EF tests (Boonstra et al., 2005; Jonsdottir, Bouma, Sergeant, & Scherder, 2006; Willcutt et al., 2005)—a conclusion much harder to justify logically in patients with demonstrated PFC damage.

• *EF tests have very low or no ecological validity.* That is, they correlate poorly if at all with ratings of EF in daily life activities in natural settings in adults (Alderman et al., 2003; Bogod, Mateer, & MacDonald, 2003; Burgess et al., 1998; Chaytor, Schmitter-Edgecombe, & Burr, 2006; Hummer et al., 2010; Ready et al., 2001; Wood & Liossi, 2006) or in children with frontal lobe lesions, traumatic brain injury (TBI), or other neurological or developmental disorders (Anderson, Anderson, Northam, Jacobs, & Mikiewicz, 2002; Mangeot et al., 2002; Vriezen & Pigott, 2002). This is

also the case both in adults with ADHD and in children with ADHD followed to adulthood (Barkley & Fischer, in press; Barkley & Murphy, 2010a). The results of those studies usually reveal that any single EF test shares just 0–10% of its variance with EF ratings. The relationships are frequently not statistically significant. Even the best combination of EF tests shares approximately 12–20% of the variance with EF ratings or observations as reflected in these studies. Yet these two types of measurement are supposed to be measuring the same construct, EF. If IQ is statistically removed from the results, the few significant relationships found in these studies between EF tests and EF ratings may even become nonsignificant (Mangeot et al., 2002). Something is terribly amiss here if different approaches to measuring the same construct are found to be so poorly related to each other.

Additional evidence for the low ecological validity of EF tests comes from studies in which the performances of frontal-lobe-injured patients on tasks in daily life have been directly observed and correlated with EF test batteries. These studies, too, find little or no relationship between impairment in such performances and EF test results (Alderman et al., 2003; Mitchell & Miller, 2008). Here again, EF tests may account for just 9–15% of the variance or less in ratings of adaptive impairment, primarily in work activities (Ready et al., 2001; Stavro, Ettenhofer, & Nigg, 2007). In contrast, research has noted moderate relationships between EF ratings and measures of daily adaptive functioning in children with TBI and other neurological or developmental disorders (Gilotty, Kenworthy, Sirian, Black, & Wagner, 2002; Mangeot et al., 2002), as well as in adults with ADHD (Biederman, Petty, Fried, Doyle, et al., 2008), those with frontal lobe disorders (Alderman et al., 2003), and college undergraduates (Ready et al., 2001). And EF ratings substantially outpredict EF tests in the variance shared with measures of impairments in various major life activities, such as occupational functioning, educational history, driving, money management, and criminal conduct (Barkley & Fischer, in press; Barkley & Murphy, 2010a). The totality of findings to date concerning the relationship of EF tests to EF ratings and of each to impairment in daily life indicates that EF tests are largely not sampling the same constructs as are EF ratings or direct evaluations of EF in daily life (Alderman et al., 2003; Shallice & Burgess, 1991). It also provides a basis for not accepting EF tests as the primary or sole source for establishing the nature of EF deficits in various disorders.

The evidence to date indicates that if assessing how well people do in using EFs in their daily life activities is important in clinically evaluating EF, then rating scales assessing EF are superior to EF tests in doing so. Yet the very plethora of studies using tests exclusively to evaluate EF with various patient and general population samples indicates that these serious problems with EF tests have gone largely unknown, unappreciated, or ignored.

This significant failure of EF tests to relate well to EF ratings, daily life activities, or impairment in major domains of life could well indicate that the former are not assessing EF. This seems doubtful, given that many of these tests have been shown to index activities in various regions of the PFC that largely underly EF. And it is surely unlikely to be the case that EF ratings are not actually evaluating EF. After all, their item content has been drafted directly from definitions of EF and from lists of putative EF constructs in the literature, as well as from observations

and clinical descriptions of patients with PFC lesions believed to manifest the "dysexecutive syndrome" (Burgess et al., 1998; Gioia et al., 2000; Kertesz et al., 2000). Moreover, as noted above, these ratings are substantially related to impairment in various daily life activities and various domains of adaptive functioning (work, education, driving, social relationships, self-sufficiency, etc.) in which EF would surely be operative.

The solution to this paradox of why EF tests and EF ratings are so poorly related lies elsewhere, most likely in the fact that EF is more hierarchically organized than models built entirely on EF tests indicate (Barkley, 2011a; Barkley & Fischer, in press; Barkley & Murphy, 2010a, 2010b). EF tests probably assess the most rudimentary, moment-to-moment, instrumental, and cognitive level of EF. But they are very poor at capturing the higher adaptive, tactical, and strategic levels of EF as they are deployed in daily adaptive functioning, human interactions, and socially cooperative and reciprocal activities that play out over much longer spans of time (days, weeks, months, and years) (Barkley, 2011a). It is at these higher, more complex, and longer-term levels that a rating scale of EF can be useful because of its use of a far longer ascertainment window for capturing summary judgments of behavior over time and because of its ability to capture EF symptoms in their important (largely social) contexts via the reports of the index person and of others who know that person well.

Also, if the purpose of evaluating EF in patients is to render some judgment as to their likelihood of experiencing impairments in major domains of life activities, then EF ratings are far superior to EF tests in doing so. On the other hand, if the purpose of the evaluation is to assess the most proximal neuropsychological EF activities related to moment-to-moment brain activity, as may be important to do in functional brain imaging studies, then EF tests may be preferable. Yet even that point is arguable in view of recent studies linking neuroimaging results to traits assessed via rating scales (Buckholtz et al., 2010). Moreover, such tests can be criticized for stripping out important social and cross-temporal elements of EF, which may reduce the validity of the task in sampling any particular component of EF or its adaptive (evolutionary) purposes.

Disadvantages of Rating Scales

In fairness, ratings of behavior, including those of EF, have their own inherent disadvantages. These problems with behavior rating scales have been discussed in detail previously (Barkley, 1987; Cairns & Green, 1979); the major points are summarized here.

- *Rating scales assume that the respondent and examiner share an understanding with regard to the nature of the item being rated and the meaning of the anchor points or potential answers being provided on the scale for responding to that item.* One source of difficulty, therefore, can be the specificity or clarity of the item to be rated, as well as the answers or anchor points being provided for the rating of the item. How well do the examiner and respondent share an understanding of the items and potential responses? The more precisely the item and its answers can be specified, the greater

is the likelihood that the respondent and examiner will share an understanding of the nature of the behavior or trait to be rated and the answers to be provided.

• *The generality or ambiguity of the item being rated and the imprecision of the anchor points of the answer scaling itself, no matter how clear and specific the wording may be, can lead to problems in testing hypotheses that require far more precise measurement than the scale is likely to provide.* Such a problem arises as a consequence of the nature of a hypothesis being tested or the purpose of the assessment. To the degree that great precision in measurement is needed to fulfill that purpose or to test that hypothesis, rating scales may be a poor choice of assessment method for the construct of interest. On the other hand, there are many questions or hypotheses in psychology or purposes in clinical evaluations that do not require such a level of precision. Indeed, using excessively precise forms of measurement such as tests or direct observations of behavior is often more expensive and time-consuming than is necessary to address the issue, when a rating scale would serve just as well at far less cost and time. The potential problem of precision of measurement with a rating scale is only so when judged relative to the hypothesis being tested or intended purpose of the clinical evaluation, and not some inherent and absolute flaw of rating scales outright.

• *Various factors may influence a rater's capacity to provide an accurate report of the behavior represented in the items on the scale.* Level of intelligence, education, emotional status, range of life experiences, prior experience with similar rating scales, and a myriad of other factors have the potential to bias the rater's reports in ways that may affect the accuracy or validity of the ratings being provided. For instance, an adult's anxiety may result in an overreporting of ADHD symptoms by that adult, relative to the level reported by someone who knows the person well and is using the same scale; the same is true in EF ratings (see Chapter 7; see also Barkley, Knouse, & Murphy, in press). It is conceivable that this represents a true biasing of the rating of this person, such that it is less valid or less accurate than the rating being provided by the collateral adult. Users of rating scales clearly need to be aware of any findings from research concerning the nature of such biasing effects on the type of rating scale under consideration.

Yet it is also possible that level of anxiety actually does result in a greater degree of ADHD symptoms—or, in the present case, ratings of EF deficits in daily life—and therefore that the rater's report is not inaccurate, but simply reflecting this real impact of anxiety on EF deficits. Performance on EF tests can likewise be adversely affected by the presence of other characteristics of the examinee, and, as here, may reflect either a bias away from the true or valid level of performance or a truly adverse effect on EF performance that is not a bias at all. Obviously, far more research is needed on the issue of the extent to which other features of raters influence their ratings of EF, and whether this represents a bias or error in the rating or a true influence on the actual behavior being rated. To the extent that such biases may operate, this will contribute to measurement error, reducing the reliability and especially the validity of the rating. However, evidence to be discussed later (Chapters 6 and 7) and some already presented above indicates that despite such potential biases, rating scales may outperform EF tests in predicting adaptive functioning in daily life activities or impairment in major domains of life. Ratings therefore may

still have considerable value, or, in this case, greater ecological validity, even if they offer imperfect samples of the behaviors being rated. The question here, then, in the selection of a measure is the intended purpose for which the measure is being used. Some purposes will demand the type of data that can be provided by administering an EF test or by directly observing the participant's behavior in selected settings. Other purposes can be easily fulfilled by the type of information to be gleaned from the far more cost-effective rating scale.

• *Rating scales are often criticized for using relatively vague references to frequency of behavior, such as "sometimes," "often," or "very often."* To some extent, such criticisms are quite justified if one is interested in very precise, fine-grained frequency counts of behavior, as might be gleaned from direct behavioral observations or from responses on a reaction time task. Such precise frequency counts may be necessary, in fact, to test certain psychological hypotheses. But at the level of clinical practice and judgment, and for other research purposes where such precision is often unnecessary and unduly costly and cumbersome, the more general judgments of individuals based on their own observations of themselves or others have proven to be sufficiently accurate, convenient, and inexpensive—and, more importantly, reliable, valid, and predictive—to be of great utility. Indeed, research comparing the scores derived from clinical tests, such as continuous performance tests or even driving simulators, has often found these scores to be less predictive of their respective constructs as assessed in natural settings (parent and teacher ratings of inhibition and attention, or department of motor vehicle records or reports of others about one's driving, respectively) than are ratings of an individual's test-taking behavior in that same setting completed by the examiner (Barkley, 1991; Barkley, Murphy, DuPaul, & Bush, 2002; Shelton et al., 1998). To reiterate, each approach to measurement has its place, depending on the purpose of the evaluation.

• *Critics of rating scales often charge that the ratings on a scale are more subjective than the responses of the individual to a test.* This reflects a misunderstanding of the term "subjective," I believe. The answers to rating scale items are not "subjective" and thus cannot be besmirched by this assertion alone. Thoughts, states of mind, and even privately held opinions are subjective as long as they remain in the head/mind and unobservable to others. They cannot be tested against reality for their conformance to it—that is to say, their objectivity (Popper, 1979). But once a thought is expressed publicly in any form (verbal, motor, or emotional behavior), that behavior can be observed by others and recorded, as in the case of a rating scale. The response itself, regardless of the thought it may be linked to, is an item of observable information that can be tested in various ways for its veracity (truth value) and utility. The degree to which the rating actually represents the privately held opinion of the rater may never be known. The degree to which the rating reflects the actual behavior that is to be rated, however, is a different matter and is open to empirical validation. Thus, when people report that they have considerable problems remembering things that they have been told to do, it is not of so much concern that their rating be evaluated for how well it captured their private mental state. More important is the issue of how well the rating corresponds to the level of problems these people actually have in remembering such things. The validity of the rating as an index of the problem can be investigated, whereas the validity of the rating as a reflection

of a private mental state cannot. But it is rare that the latter issue is the one under consideration in the research project or clinical evaluation of a patient. It is the former issue that matters most.

To summarize, any charge of imprecision or subjectivity leveled against interviews and rating scales in an effort to disparage them without reference to further evidence concerning the purpose of the evaluation; the issue under investigation; and the evidence for the interviews' or scales' reliability, validity, and utility is baseless. The validity of relatively broad human estimates as recorded in structured interviews or rating scales concerning the relative frequency of specific behaviors is itself a form of information that can be observed, recorded, and tested, as seen in Chapter 5, for its reliability, validity and utility no differently from a response to any test. That is what matters and should be the basis on which a clinician or scientist chooses a method of measurement—not some inherent bias in favor of tests over rating scales, regardless of the issue at hand.

Whether or not the behavior of a participant is recorded by his or her circling of a response to an item on a rating scale, a button press on a device, or a verbal response recorded on some testing answer sheet is a distinction without a difference. All are forms of observable and recordable behavior, and the proof of their validity is in the evidence available about them. The widespread knee-jerk penchant of neuropsychologists to assume that a test is somehow more objective, precise, and therefore useful than a rating scale is a hypothesis; it is not a fact in the bag. The claim deserves to be subjected to testing in research and is not a fact by mere consensus of opinion or proclamation by an authority alone. So far, this assumption of the test as a gold standard of measurement has proven questionable for pursuing certain issues or purposes, such as the prediction of functioning in naturalistic settings or the determination of whether a particular disorder involves deficits in EF, such as ADHD. Clinicians, industrial/organizational psychologists, and even researchers need not apologize for using rating scales to assess certain domains of behavior and functioning, instead of some form of recording directly observed behavior or testing, if the evidence shows that those scales have merit for the intended purpose. The available evidence shows that for certain purposes rating scales are acceptable forms of measurement—in this case, of EF.

Advantages of Rating Scales

With these caveats and issues surrounding the use of rating scales for evaluating human behavior and psychological traits in mind, let us consider now a list of the advantages of rating scales for clinical and research purposes, again with the caveat that the purpose of the evaluation is the determining factor in the final analysis of the best method to assess EF. To their considerable credit, rating scales have numerous advantages (see Barkley, 1987; Cairns & Green, 1979):

- *Rating scales have the capability of gathering information from informers with many years of experience with the individuals to be rated, or with themselves if they are the subjects of the ratings.*

- *This vast experience occurs across a diversity of settings and circumstances that is nearly impossible to duplicate through any other means*, at least in any cost-effective way.

- *Rating scales permit the collection of data on behaviors that occur extremely infrequently*, such that they are likely to be missed in *in vivo* observations of behavior unless such observations are extended over exceptionally long periods of time.

- *Rating scales can gather information in ways that protect an individual's privacy*, apart from the topic of the rating scale itself. In contrast, direct observations or testing of the individual can prove exceptionally intrusive into the individual's daily life, giving the examiner the opportunity to witness situations and events that have no bearing on the topic or objective of the evaluation.

- *Rating scales are much less expensive, less time-consuming, and hence more efficient and cost-effective* than either tests or direct behavioral observation methods for many clinical and research purposes.

- *Rating scales may have normative data available for establishing the statistical deviance or relative position of the individual within the population concerning a behavior or trait of interest that may not be available for other methods of assessing that behavior or trait.* In the case of EF, such normative data can be less costly to collect than is the case for many tests of EF whose norms, if available at all, are typically not representative of large samples of the general population.

- *Rating scales can incorporate a large number of items that represent a given dimension, behavior, or trait, if so desired, that can be captured in a far less time-consuming format than can equivalent items within a psychological test.*

- *Rating scales can easily be used to collect the judgments of another who knows the index person well for comparison to the person's self-ratings as a means of judging interobserver agreement or the validity of the self-ratings.* Such judgments are far more difficult and cumbersome (if possible at all) for the results of a test, short of having a second examiner readminister the same test battery to the same person under the same or similar circumstances.

- *Rating scales can be used to filter out situational variation that has little or no bearing on the purpose of the evaluation* by asking for cross-situational judgments of the frequency, intensity, timing, or other aspects of the topography of the behavior under study. Again, these would be difficult if not impossible to achieve via a test or direct observational method.

- *Rating scales permit quantitative distinctions to be made concerning the qualitative aspects of human behavior* that are often difficult to obtain through other means, such as tests and observations, at least in a cost-effective manner.

- *Rating scales, like tests and observations, can be used for many different purposes.* These include subgrouping various populations along a trait or behavior of interest in epidemiological research; further exploring etiological hypotheses concerning certain behaviors or disorders; determining the prognosis of clinical groups of patients followed over long time intervals; and serving as measures of behavior change that may be secondary either to interventions or to a change in an individual's status, such as evaluating the consequences of various forms of brain injury or the impact of psychiatric disorders or psychological adversities.

For these and other reasons, behavior ratings of EF and its deficits in daily life activities have an important role to play in research and clinical practice in various psychological specialties.

Early Developmental History of the BDEFS

The development of the BDEFS has spanned more than a decade. The scale began originally as an attempt to evaluate EF in adults with ADHD, given the substantial research indicating that the disorder was associated with EF deficits on EF tests, at least at the group level of analysis (Frazier et al., 2004; Hervey et al., 2004). To assess EF deficits in daily life, my colleagues and I developed our own scale.

One other EF rating scale for adults existed at the time: the Dysexecutive Questionnaire (DEX; Burgess et al., 1998; Chaytor et al., 2006; Wood & Liossi, 2006). Though a commendable effort at creating a rating scale for EF deficits, it had several notable limitations that led us to develop our own EF scale (Barkley & Murphy, 2010b). The DEX comprised just 20 items meant to sample the broad range of symptoms believed to be representative of a general dysexecutive (frontal lobe) syndrome. This limited item pool was also not theoretically based but clinically based in its construction, originating in clinical descriptions of patients with PFC injuries. The scale did not yield scores concerning specific EF problem dimensions, but merely provided a single global summary score believed to reflect the dysexecutive syndrome.

A prototype EF scale (Barkley & Murphy, 2010b) was therefore developed for use in two large federally funded research projects on adults with ADHD (Barkley, Murphy, & Fischer, 2008). One of these projects, known as the UMASS Study, examined clinic-referred adults diagnosed with ADHD in comparison to both Clinical and Community control groups. The second study, known as the Milwaukee Study, was a follow-up study of hyperactive children into young adulthood (mean age 27). (See Barkley et al., 2008, for details of both studies.)

The scale's development was largely based on an earlier theory of EF, its five constructs, and their specific adaptive purposes (Barkley, 1997a, 1997b, 2001; see Chapter 1), as well as the larger literature on the nature of EF (see Denckla, 1996; Fuster, 1997; Lyon & Krasnegor, 1996; Stuss & Benson, 1986) and the rich and lengthy history of descriptions of the symptoms of patients with PFC injuries (again, see Chapter 1). The original item pool consisted of 91 items, which were developed to reflect inhibition; nonverbal working memory (self-directed sensing, especially visual imagery, sense of time, and time management); verbal working memory (self-directed private speech, verbal contemplation of one's behavior before acting, etc.); emotional/motivational self-regulation (inhibiting emotion, motivating oneself during boring activities, etc.); and reconstitution (generativity, planning, problem solving, and goal-directed inventiveness). According to this theory, the constructs are interactive and serve the overarching purpose of self-organizing behavior across time to prepare for and attain future goals. Items were also generated from a review of more than 200 charts of adults diagnosed with ADHD at a regional medical center's adult ADHD clinic, given that ADHD is largely a disorder of PFC functioning

(Bush, Valera, & Seidman, 2005; Hutchinson, Mathias, & Banich, 2008; Mackie et al., 2007; Paloyelis, Mehta, Kuntsi, & Asherson, 2007; Valera, Faraone, Murray, & Seidman, 2007), has long been construed as such (Pontius, 1973), and is characterized by many theorists as a disorder involving EF (Barkley, 1997b; Castellanos et al., 2006; Nigg & Casey, 2005; Sagvolden, Johansen, Aase, & Russell, 2005).

The scale items focused on problematic symptoms (deficit measurement) rather than on positive or normative EF functioning. The BDEFS was and is not intended to assess the broad variation of EF in the general population, in order to identify the range of individual differences in normal functioning that may exist in that population. Scales focusing on typical EF in a general population may be quite useful in studying the range of individual differences, as in studies of behavior genetics, normal development across the lifespan, or other purposes in which normal variation in a psychological trait is of interest. The BDEFS, in contrast, was and is intended to be used for clinical purposes, to evaluate the range of EF deficits in clinic-referred or high-risk adults—in other words, to assess symptoms of executive dysfunctioning. The clinician is faced with the question of the extent to which the complaints of a clinical patient concerning problems in EF are indicative of significant difficulties in EF. The question here is how atypical the patient's complaints are, rather than where the person falls within the distribution of normal variation in EF in the general population.

Deficit or symptom measurement is typical in the development of clinical assessment devices such as those used to assess anxiety, depression, bipolar disorders, or other adult psychological disorders—for instance, instruments such as the Symptom Checklist 90—Revised (SCL-90-R; Derogatis, 1986), or other rating scales of EF (e.g., Roth, Isquith, & Gioia, 2005). The distribution of such scores on symptom checklists will be quite different from the more typical bell-shaped curve of the distribution of normal variation in a typical psychological ability, such as intelligence or academic achievement. The score distribution of a deficit (symptom) measurement tool can be expected to have a distribution highly skewed toward the zero point or low score on the scale, with a steep drop in the percentage of the general population that manifests an increasing number of symptoms. Deficit or symptom assessment evaluates the likelihood that a patient's complaints are sufficiently severe or numerous to place the patient significantly outside the distribution of typical complaints for the general population. This was our purpose in constructing the BDEFS: to use items reflecting symptoms or deficits, rather than items reflecting the range of normal abilities.

Three separate instruments were developed from the original 91-item pool: two rating scales, and an interview to be given by a clinician. First, the 91 items were cast in the form of a rating scale to be completed by each participant. This scale was then recast in a third-person form to be completed by someone who knew the participant well. This was typically a parent or cohabiting partner, but if neither of these was available, a sibling or a close friend was used instead. Each item on these two scales was answered on a 0–3 Likert scale (0 = rarely or not at all, 1 = sometimes, 2 = often, and 3 = very often). The remainder of this chapter focuses mainly on these two rating scales, which served as the prototypes for the scales published in the present volume. Nevertheless, the participants in the UMASS Study also com-

pleted the interview with a clinical psychologist. The interview contained the same items as did the rating scale. The measures differed, however, in that the interview items asked only whether an individual had experienced that symptom as occurring often or more frequently during the past 6 months (see Barkley et al., 2008), rather than being given a choice of four possible ratings, as was the case with the rating scale. The term "often" was chosen to signify a symptom because of our earlier work with an adult ADHD rating scale, which suggested that this response occurs infrequently in the general population of adults to serve as a marker for a problem or symptom (Barkley & Murphy, 2006; Murphy & Barkley, 1996).

Three groups of participants (ages 18–60 years) were used in the initial UMASS Study of adults with ADHD (see Barkley et al., 2008; Barkley & Murphy, 2010b): (1) 146 adults clinically diagnosed with ADHD; (2) a Clinical control group (97 adults evaluated at the same clinic but not diagnosed with ADHD); and (3) a Community control group (109 adult volunteers from the local community). The adults in the first two groups were obtained from consecutive referrals to an adult ADHD clinic in a department of psychiatry at a regional medical school. The Community adults were obtained from advertisements posted throughout the medical school and from periodic newspaper ads in the regional newspaper. (More information on all three of these samples can be found in Barkley et al., 2008.) Slightly more than two-thirds (68%) of those in our group with ADHD were males; this differed from a nearly equal sex distribution (51%) in our Community control group, but was not significantly different from the distribution in the Clinical control group (56%). This finding is in keeping with many studies of children and adults with ADHD, which demonstrate a greater representation of the disorder in males than in females (Barkley, 2006; Kessler et al., 2006). Typically, the male–female ratio in ADHD is approximately 2:1 in adult epidemiological studies—a ratio similar to that found here. The vast majority of participants were White (94%), or of European American ancestry, and the groups did not differ in this respect. Therefore, caution is warranted in extrapolating the initial results described below to adults in other ethnic groups.

We had ratings from others on 129 (88%) of the group with ADHD, 88 (90%) of the Clinical group, and 92 (84%) of the Community group. These differences in the percentage of each group having collateral reports of EF were not significant (χ^2 = 19.38, *df* = 12, *p* = NS). Examination of the pattern of relationships of these other people to the participants revealed no significant differences across the groups.

Identifying Underlying Dimensions of EF in Daily Life Activities

We applied principal-components factor analysis to the 91 original items in the prototype of the BDEFS (hereafter referred to as the P-BDEFS) to explore the underlying dimensions of EF deficits in daily life, using self-reports from the entire UMASS sample of 352 participants. This gave us 3.86 participants per EF item on the scale, which was below the traditionally recommended 5 participants per variable to achieve adequate reliability of the factor structure. Nevertheless, as noted below,

evidence for the actual reliability of this factor structure was obtained from the fact that the same factor structure emerged from the other-report ratings of EF deficits in this study. It also emerged in the Milwaukee Study, the separate study of hyperactive children followed to adulthood, which used the interview version of this rating scale (Barkley & Fischer, in press). Finally, it emerged in the subsequent version of the BDEFS with the normative sample to be discussed in Chapter 3 (N = 1,249).

Although 10 factors were obtained that had eigenvalues greater than 1, only five of these had at least 10 items with their highest loading on a factor; accounted for at least 2% or more of the variance before rotation; and, incidentally, had eigenvalues of 1.8 or higher. We therefore retained just those five factors. We found that 88 items had loadings of at least .400 on any of these five factors. Three items were dropped from the scale because they did not have a loading of ≥ .400 on any of the final five factors in that analysis ("Can't seem to sustain friendships or close relationships as long as other people," "Poor or sloppy handwriting," and "Less able than others to recall events from childhood").

The unrotated factor solution indicated that the first factor accounted for a substantial amount of the variance in these ratings (over 51%). Each of the remaining factors accounted for 2–4.4% of the variance per factor, such that the five unrotated factors explained more than 63% of the variance. It appeared that one large factor accounted for the majority of variance in EF deficits in daily life. A similar finding was evident in the factor analyses of the Behavior Rating Inventory of Executive Functioning (BRIEF)—Adult Version (Roth et al., 2005). An identical finding was noted by Biederman and colleagues (2008), using the P-BDEFS with a different sample of adults who had ADHD and control adults. The highest-loading items on this first factor for the P-BDEFS dealt with sense of time, time management, planning, preparing for deadlines, and other goal-directed behavior. We therefore labeled this factor as Self-Management to Time. Such a finding supports the view of Fuster (1997) and others (see Chapter 1) that EF chiefly comprises the cross-temporal organization of behavior toward goals.

Many investigators might have stopped the analysis with these findings, concluding that EF in daily life activities chiefly comprises this single dimension. An identical finding was noted by Faraone and colleagues (2010). However, another 12% of the variance (prior to rotation) could be accounted for by four additional factors that each accounted for at least 2% or more of the unrotated variance; had at least 10 items loading primarily on that dimension; and, incidentally, had eigenvalues of at least 1.8. The issue concerning the number of items assigned to a factor is important if a dimension is to have any chance of being assessed reliably; reliability (as well as validity) is in part dependent on the number of items for a construct included in the scale. And so we subsequently used a varimax rotation to analyze the scale further, specifying a five-factor solution. (A promax rotation yielded the same results, except that the factor loadings of the items were even higher than those found in the varimax rotation. This was to be expected, given the purposes of each approach to rotation. Varimax seeks to decrease correlations among the factors, whereas promax permits such correlations to exist across factors.) The final five factors for the self-report version of the P-BDEFS, and their variance (after the varimax rotation), were as follows:

• *Factor 1* (*Self-Management to Time*) contained 23 items and accounted for 15.7% of the variance after rotation. As noted above, the highest-loading items dealt with sense of time, time management, planning, preparing for deadlines, and other goal-directed behavior.

• *Factor 2* (*Self-Organization/Problem Solving*) contained 21 items and accounted for 15.2% of the variance. The highest-loading items pertained to organizing one's thoughts, actions, and writing; thinking quickly when encountering unexpected events; and inventing solutions to problems or obstacles encountered while pursuing goals.

• *Factor 3* (*Self-Restraint or Inhibition*) comprised 23 items and accounted for 14.1% of the variance. Its highest-loading items dealt with making impulsive comments, poor inhibition of reactions to events, impulsive decision making, doing things without regard to their consequences, and not thinking about the relevant past or future before acting. A few items also dealt with poor self-awareness and the inability to take other people's perspectives about their own behavior or a situation.

• *Factor 4* (*Self-Motivation*) consisted of 11 items and accounted for 9.8% of the variance, with items mainly dealing with taking short cuts in one's work, not doing all assigned work, being described as lazy by others, not putting in much effort on work, needing more supervision than others while working, getting bored easily, and so forth.

• *Factor 5* (*Self-Activation/Concentration*) contained 10 items and explained 8.6% of the variance. It comprised items dealing with being easily distracted by one's thoughts when doing boring work; staying awake and alert while working; being able to persist at boring activities; sustained concentration in reading, paperwork, meetings, or other activities that were not interesting; being prone to daydreaming when one should be concentrating; and having to reread uninteresting written material in order to comprehend it.

As noted above, a factor analysis with varimax rotation was also conducted on the version of the P-BDEFS completed by others (Barkley & Murphy, 2010b). The same five-factor solution emerged with nearly identical item content. The variance accounted for after rotation by each factor was as follows: Self-Management to Time, 19.9%; Self-Organization/Problem Solving, 16.5%; Self-Restraint, 16.7%; Self-Motivation, 8%; and Self-Activation/Concentration, 3.8%. In only a few instances did an item have its highest loading on a factor different from the one to which it was assigned in the self-ratings. Even then, it had either an equivalent loading or a second-highest loading on the scale to which it was assigned in the self-ratings. It is important to note here that the P-BDEFS was limited by the absence of any items addressing the domain of emotional self-regulation. Therefore, there was no opportunity for such a factor to emerge in these analyses. This limitation was corrected in the final version of the BDEFS, as described in the next section.

We created five subscales from the self-report version of the P-BDEFS reflecting these factors and consisting of just those items having their highest loading on each subscale. For each subscale, we created a total score based on the sum of the indi-

vidual raw item scores. We did the same for the other-report version, using the identical items, so that direct comparisons could be made between the two versions.

Despite the emergence of these five factor-based subscales on the P-BDEFS, the subscales were significantly intercorrelated, ranging from .74 to .88 for the self-ratings and .75 to .88 for the other-ratings. The subscales therefore shared 56–77% of their variance, which, as suggested above and in Chapter 1, may indicate the possible existence of a single underlying metaconstruct of deficits in EF (Barkley & Murphy, 2010b).

For the self-ratings, three of the five subscales showed a low but significant correlation with participant age: Self-Management to Time ($r = -.11$, $p = .05$); Self-Motivation ($r = -.21$, $p < .001$); and Self-Activation ($r = -.11$, $p = .04$). This was true for these same three subscales rated by others: Self-Management to Time ($r = -.16$, $p = .01$); Self-Motivation ($r = -.27$, $p < .001$); and Self-Activation ($r = -.16$, $p = .009$) (Barkley & Murphy, 2010b). Thus older individuals within the age range of this study had fewer deficits in EF in daily life.

Only one of the five self-rating subscales was modestly but significantly correlated with participant IQ: Self-Organization/Problem Solving ($r = -.15$, $p = .007$). This pattern was also the case for the other-ratings ($r = -.17$, $p = .008$) (Barkley & Murphy, 2010b). These results suggest that, unlike EF tests, ratings of EF in daily life activities are largely not contaminated by general intelligence. Even the single subscale that was related to intelligence shared just 2.9% of its variance with the IQ measure used in our study.

Our results indicated that the adults with ADHD rated themselves as having significantly more severe EF deficits on all five subscales, compared to both the Clinical and Community control groups. The adults in the Clinical group also rated themselves as having more severe EF deficits than the Community group. This was also the case for the other-ratings.

We also used the Community control group to determine a threshold for being clinically deficient on each of the five subscales. We specified this deficiency as being +1.5 standard deviations above the mean of that group on the P-BDEFS. For self-reports, we found that 89–98% of the group with ADHD and 83–93% of the Clinical group fell into this clinically significant range (7th percentile of the Community group), compared to just 7–11% of the Community group, on the five subscales of the P-BDEFS. The ADHD group had a significantly greater percentage thus impaired on the Self-Motivation scale than the Clinical group, but these differences were not significant on the other subscales. When other-ratings were used, the figures were 84–99% of the ADHD group and 68–89% of the Clinical group, compared to 9–14% of the Community group. The ADHD group had a higher percentage rated as impaired by others on four of the five subscales, the exception being Self-Organization/Problem Solving.

Adult ADHD was therefore associated with significant deficits in EF in daily life as assessed by the P-BDEFS, and the majority of those adults were in the clinically deficient range in EF in daily life (Barkley & Murphy, 2010b). These results indicated that the P-BDEFS could readily distinguish clinic-referred patients having a psychiatric diagnosis from nonclinical controls; in fact, the two distributions of scores (clinic vs. nonclinic) showed only limited overlap. Our research also found that adults with ADHD had the most severe deficits in the five dimensions of EF

assessed by the scale relative to the Clinical controls, and that both of these groups manifested substantially more EF deficits than did the Community control group.

Using these factors and their item loadings, we computed the same scores for using the interview form of the P-BDEFS mentioned earlier. Again, these items were answered as yes (the item occurred often or more frequently) or no (the item did not occur often). The interview form therefore yielded the equivalent of an EF symptom count. For the present manual, the factor scores from the interview were correlated with those noted above for the P-BDEFS self-report rating scale. The results showed that the scores for each of the factors were highly correlated between these two measures: Self-Management to Time ($r = .921$), Self-Organization/Problem Solving (.885), Self-Restraint (.873), Self-Motivation (.867), Self-Activation/Concentration (.904), and the Total EF Summary Score (sum of all factor scores) (.925). All correlations were significant at $p < .001$. Given the high magnitude of these relationships, it is likely that any of the findings discussed here and in Chapters 6 and 7 (on reliability and validity of the BDEFS rating scale) would apply just as much to the interview form.

This became evident when the interview was used in a subsequent study with different samples: In the Milwaukee Study, the separate study of hyperactive children followed to adulthood, we found the same factor structure for the interview version of the P-BDEFS (Barkley & Fischer, in press). Children whose ADHD had persisted to age 27 years were found to have significantly more EF deficits in daily life than were such children whose ADHD had not persisted. And the latter group had more severe EF deficits than did the Community control group followed to adulthood. Thus, again, we found that persistence of ADHD into adulthood was associated with more severe EF deficits in daily life activities, and that by adulthood more severe ADHD was linked to more severe EF deficits. Such findings have also been reported in research on children with ADHD by investigators using several different rating scales of EF (Gioia, Isquith, Kenworthy, & Barton, 2002; Thorell, Eninger, Brocki, & Bohlin, 2010).

The P-BDEFS thus appeared to have utility for the clinical assessment of EF in daily life activities. The initial scale assessed five dimensions of EF deficits in daily life activities and was found to be useful in identifying such deficits in adults diagnosed with ADHD and other clinically referred adults relative to a Community group, as well as in children with ADHD followed to adulthood. This led us to proceed with further development of the scale, the collection of norms for the present BDEFS, and publication of that information in this manual.

Further Development of the BDEFS

For the BDEFS rating scales published in this manual, the items from the P-BDEFS that were found to have factor loadings of at least .500 or greater on one of the five factors in our earlier study (Barkley & Murphy, 2010b) were retained. Therefore, a few items on each of the P-BDEFS subscales were abandoned, in an attempt to reduce the overall scale's length and hence the time needed to complete it without loss of information on the five subscales. A further examination of the scale items, as well as the resulting factor structure, suggested that one domain of EF appeared

to be substantially underrepresented in the P-BDEFS and so did not have an opportunity to emerge in the foregoing analyses. This component was the self-regulation of emotion. The few items on the prototype scale that dealt with emotion were primarily representing the impulsive expression of impatience, frustration, and anger, and loaded on the Self-Restraint (Inhibition) factor, as might be expected. But there were no items pertaining to self-management of emotion on the P-BDEFS. This is an important component of EF that is often neglected in developing EF test batteries, but appears to be commonly observed among the deficits in clinical patients with PFC injuries (see Chapter 1).

Emotional self-control is believed to comprise a two-stage process: (1) the inhibition of strong emotional reactions to events; and (2) the subsequent engagement of self-regulatory actions and strategies that include (a) self-soothing, (b) refocusing attention away from the provocative event, (c) reducing and moderating the initial emotion, and (d) organizing the eventual emotional expression so that it is more consistent with and supportive of individual goals and long-term welfare (Barkley, 2010; Gottman & Katz, 1989; Hinshaw, 2003; Martel, 2009; Masters, 1991; Melnick & Hinshaw, 2000). It is argued that deficits in both of these components of emotional self-control lead to impulsive emotional expression and the subsequent deficient self-regulation of those emotions in conformity with personal or social purposes and goals.

I used the model of emotional self-regulation developed by Gross (1998; Gross & John, 2003; Gross & Thompson, 2007) to generate an additional 10 items to reflect these problems with poor self-regulation of emotional states, as well as two additional items related to self-motivation. This was done in an effort to strengthen (lengthen) the item content of the Self-Motivation subscale in the P-BDEFS. Gross and Thompson (2007) have argued that there are five sets of emotion regulation strategies: situation selection, situation modification, attention deployment (such as distraction or gaze aversion), cognitive change (such as conscious reappraisal), and response modulation (such as suppressing expressions). These five sets of strategies can be further subdivided by whether they are antecedent-focused or response-focused. Efforts were made to have at least one item in the BDEFS evaluate each of these strategies.

The final version of the BDEFS on which the national norms would be collected therefore contained 100 items. Following further analyses to be reported later in this manual, the final scale was reduced to 89 items. And from that scale, a shorter version was also developed, comprising 20 items having the highest loadings on their respective factors. With this background information, the survey methods used to collect normative data on the scale and the demographic characteristics of that sample are presented in the next chapter.