10 Assessment of Executive Function Using Rating Scales: Psychometric Considerations

Jack A. Naglieri and Sam Goldstein

Introduction

In any field of scientific study the information we obtain from research is directly related to the quality of the information we obtain from the tools we use. The better the tool, the more accurate and reliable the information that is obtained. Ultimately, the validity of the tools used in science will be proportionate to the quality of the concepts being evaluated. Ultimately, better tools are more effective for researchers and clinicians. The better the tools used in research and clinical practice, the more valid and reliable the decisions will be, the useful the information obtained will be, and ultimately, the better the services that will be provided. In this chapter, the rating scales used for assessment of executive function will be examined.

There are two goals of this chapter. First, to illustrate the relevance reliability and validity have on the decisions made by clinicians and researchers, review of essential psychometric qualities of test reliability and validity will be provided. The practical implication these psychometric issues have for the assessment and the implications for interpretation of results will be emphasized. Special attention will be paid to scale development procedures, particularly methods used to develop derived scores. The second section of this chapter will focus on rating scales used to assess behaviors considered indicative of executive function. The overall aim is to provide an examination of the relevant psychometric issues and the extent to which researchers and clinicians can have confidence in the tools they may use to assess executive function.

Reliability

Good reliability is critical for any test used for clinical practice as well as research purposes. It is essential that clinicians and researchers know the reliability of a test so that the amount of accuracy and the amount of error in the measurement of the construct are known. The higher the reliability, the smaller the error and the smaller range of scores used to build the confidence interval around the estimated true score. The smaller the range, the more precision and confidence practitioners can have in their interpretation of the results.

Bracken (1987) provided suggestions for the evaluation of test reliability (evaluated using some internal reliability estimate). He stated that individual scales from a test (e.g., a subtest or subscale) should have a reliability of .80 or greater and total tests should have an internal consistency of .90 or greater. The level of precision required should be determined in relations to the reason for

J.A. Naglieri (✉)
University of Virginia, Curry School of Education,
White Post Road 6622, Centreville, VA 20121, USA
e-mail: jnaglieri@gmail.com

S. Goldstein
Neurology, Learning and Behavior Center,
University of Utah School of Medicine, South 500
East, Suite 100 230, Salt Lake City, UT 84102, USA
e-mail: info@samgoldstein.com

S. Goldstein and J.A. Naglieri (eds.), *Handbook of Executive Functioning*,
DOI 10.1007/978-1-4614-8106-5_10, © Springer Science+Business Media New York 2014

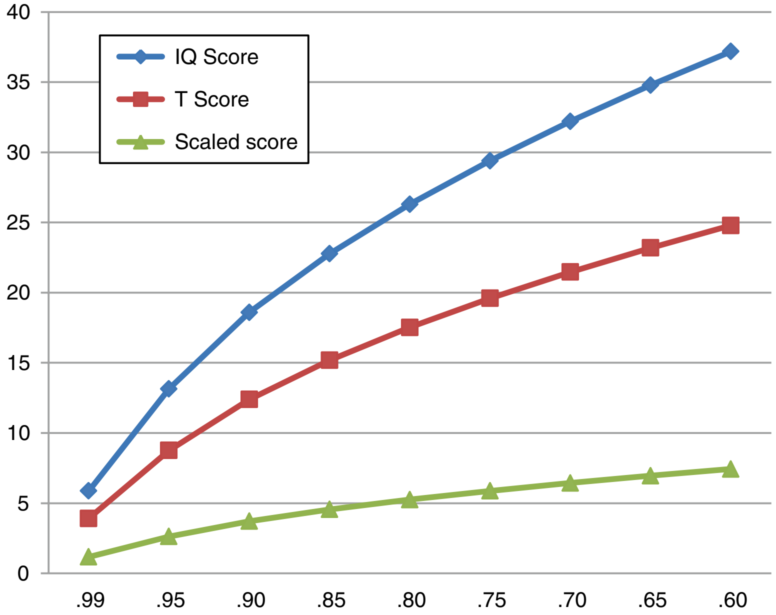

Fig. 10.1 Range of a 95 % confidence intervals as a function of reliability

testing and the importance of the decisions that may be made. If a score is to be used for screening purposes where over identification is preferred to under identification, a .80 reliability standard for a total score may be acceptable. If, however, high-stakes decisions are made, for example about special educational placement, then a higher reliability (e.g., .95) is more appropriate (Nunnally & Bernstein, 1994).

Details About Reliability

Researchers and practitioners must be aware of the reliability of any score they use. High reliability is essential for all test scores used in research and applied settings. Reliability is important to the practitioner because it reflects the amount of error in the measurement. That is, reliability describes the amount of variability to expect around the true score, assuming that any obtained score comprises the true score plus error (Crocker & Algina, 1986). Because we can never directly determine the true score, we use the reliability coefficient to describe a range of values within which the person's score likely falls with a particular level of probability. The size of the range is determined by the reliability of the measurement with higher reliability resulting in smaller ranges. This provides a way to describe an IQ score as a number and a range. For example, 105 (±5), meaning that there is a 90 % likelihood that the child's true IQ score falls within the range of 100–110 (105 ± 5). The range of scores (called the confidence interval) is computed by first obtaining the standard error of measurement (SEM) from the reliability coefficient and the standard deviation (SD) of the score in the following formula (Crocker & Algina, 1986):

$$\mathrm{SEM} = \mathrm{SD} \times \sqrt{1 - \mathrm{reliabillity}}$$

The SEM, which is the average standard deviation of a person's scores around the true score, is used to compute the confidence interval. To obtain a confidence interval, the SEM is multiplied by a *z* value of, for example 1.64 or 1.96 at the 90 % or 95 % levels, respectively. The resulting value is added to and subtracted from the obtained score to yield the confidence interval. In the example provided above, the confidence interval for an obtained score of 100 is 95 (100 − 5) to 105 (100 + 5). Figure 10.1 provides the range of confidence intervals (95 % level of

Table 10.1 Relationships between obtained scores, estimated true scores, and confidence intervals

Obtained standard score	Estimated true score	True score minus obtained score	Lower confidence band	Upper confidence band	Confidence interval range
55	60	5	52	67	16
70	73	3	65	81	16
85	87	2	79	94	16
100	100	0	92	108	16
115	114	−2	106	121	16
130	127	−3	119	135	16
145	141	−5	133	148	16

Note: This assumes a reliability coefficient of .90 and a 90 % confidence interval

confidence), that is, the values to be added and subtracted from an obtained score to calculate confidence intervals for a typical IQ score (Mn = 100; SD = 15), *T*-score (Mn = 50; SD = 10), and IQ test scaled score (Mn = 10; SD = 3) for measures with a reliability of .60 through .99. The range within which the true score is expected to fall varies as the reliability coefficient changes—the lower the reliability, the wider the range of scores that can be expected to include the true score.

It is important to know, however, that the confidence interval (and SEM) should be centered around the estimated true score rather than the obtained score (Nunnally & Bernstein, 1994). In many published tests (e.g., Wechsler Intelligence Scale for Children Fourth Edition (Wechsler, 2003) and the Cognitive Assessment System (Naglieri & Das, 1997)), the confidence intervals provided in the norms tables are centered on the estimated true score. Table 10.1 illustrates the relationships between obtained and estimated true scores, the lower and upper range of the confidence intervals in relation to the obtained scores, and the actual range of the confidence intervals for a hypothetical test (mean of 100, *SD* of 15) with a reliability of .90 at the 90 % level of confidence.

Examination of the scores in Table 10.1 shows that the confidence interval is equally distributed around a score of 100 (92 and 108 are both 8 points from the obtained score) but the interval becomes less symmetrical as the obtained score deviates from the mean. Ranges for standard scores that are below the mean are *higher* than the obtained score. As shown in Table 10.1, the range for a standard score of 70 is 65–81 (5 points below 70 and 11 points above 60). In contrast scores for standard scores that are above the mean are *lower* than the obtained score. The range for a standard score of 130 is 119–135 (11 points below 130 and 5 points above 130). This asymmetry is the result of centering the range of scores on the estimated true score rather than the obtained score even though the size of the confidence interval is constant (±8 points).

Whether confidence intervals are constructed using obtained or estimated true score methods, measurement error must be considered and communicated when scores from any test are used and particularly when results are explained to consumers. Confidence intervals, especially those that are based on the estimated true score, should be provided for all test scores including rating scales.

The importance of the SEM must be considered when two scores are compared. The lower the reliability (the larger the SEM), the more likely two scores will differ on the basis of chance. In order to account for reliability's influence on the difference between scores, a formula for determining how different two scores need to be can be applied. This formula is based on the SEM of each score and the *z* score associated with a specified level of significance. The difference needed for significance can be computed using the following formula:

$$\text{Difference} = z \times \sqrt{\text{SEM1}^2 + \text{SEM2}^2}$$

The relationship between reliability and the differences needed for significance when comparing two scores is provided in Table 10.2.

Table 10.2 Differences required for significance when comparing standard scores with a mean of 100 and SD of 15 ($p=.05$)

Reliability	.99	.95	.90	.85	.80	.75	.70	.65	.60
.99	4	7	10	12	13	15	16	18	19
.95	7	9	11	13	15	16	17	19	20
.90	10	11	13	15	16	17	19	20	21
.85	12	13	15	16	17	19	20	21	22
.80	13	15	16	17	19	20	21	22	23
.75	15	16	17	19	20	21	22	23	24
.70	16	17	19	20	21	22	23	24	25
.65	18	19	20	21	22	23	24	25	25
.60	19	20	21	22	23	24	25	25	26
.55	20	21	22	23	24	25	25	26	27
.50	21	22	23	24	25	25	26	27	28

To use this table, find the row that corresponds to the reliability of one score and the column that corresponds to the second. Read into the table for the difference required for significant. The significance level is based on the assumption that one pair is compared. The values in Table 10.2 can be used to compare more than one pair of scores; however, doing so changes the actual level of significance in proportion to the number of comparisons made. For example, using a .05 level of significance 6 times makes the experiment-wise error rate actually .265, not .05, because six pair wise increases error (the chance of a type I error is obtained using the formula 1—$(1-.05)\times6$). One way to control for inflation in the level of significance is by using the Bonferroni correction method. This procedure controls for the number of comparisons by setting the experiment-wise error rate on the basis of making all six comparisons simultaneously (e.g., $.05/6=.008$).

The differences needed for significance when comparing two scores with reliability coefficients that range from .55 to .99 are shown in Fig. 10.2 for scores that have an *SD* of 15 (a typical IQ test), 10 (a *T*-score used by many rating scales), and 3 (an IQ test subtest scaled score) calculated using the formula above. These findings demonstrate that in research and most importantly in clinical settings, test scores with high reliability are desired.

Researchers and clinicians assessing behaviors associated with executive function should use test scores possessing a reliability coefficient of .80 or higher and any total or composite score should have reliability of at least .90. If a rating scale does not meet these requirements, then their inclusion in research studies and particularly in clinical settings should be questioned. Clinicians are advised not to use measures that do not meet reliability standards because there will be too much error in the obtained scores to allow for reliable interpretation and especially comparison with other scores. This is particularly important when the decisions clinicians are making could have substantial and long-lasting impact on a child, adolescent, or adult.

Validity

While having a measure with good reliability is essential, reliable measurement of a construct that has limited validity has little use to the clinician and researcher. Validity is described as the degree to which empirical evidence supports an interpretation of scores representing a construct of interest. For example, a rating scale for evaluation of behaviors associated with executive function should include questions that accurately reflect the concept. Authors striving to produce a measure of executive function are especially burdened with the responsibility to define the concept carefully and the observable behaviors associated with it. When the behaviors and characteristics

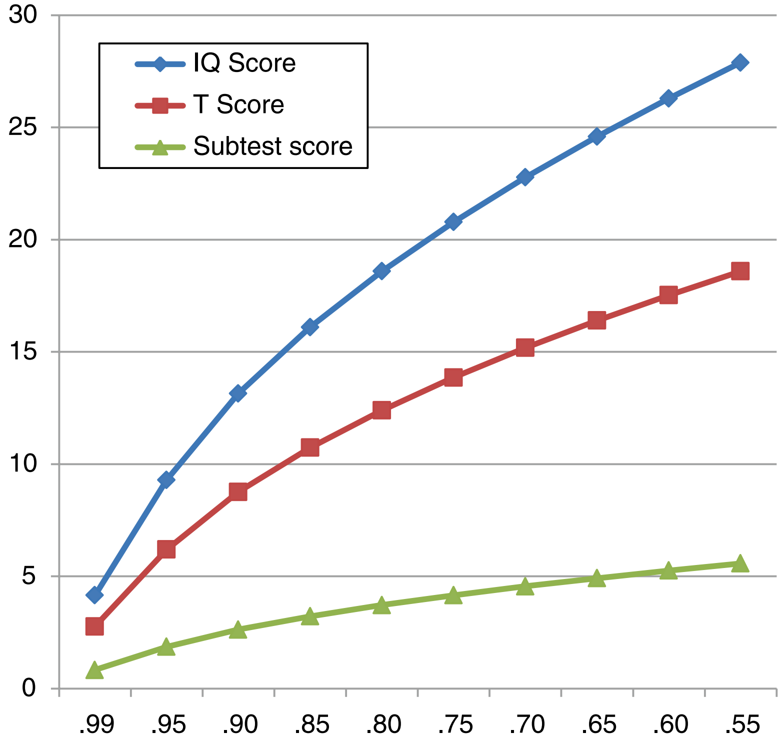

Fig. 10.2 Differences required for significance at $p=.05$ when comparing two IQ, or *T*, or subtest scaled scores at various levels of test reliability

associated with the disorder are thoroughly operationalized, then the way the concept has been measured by the observations of, for example, a parent or teacher can be tested and the task of establishing validity begun. Evidence of validity will be dependent upon the extent to which the items have adequate reliability.

The concept of executive function has been defined in multiple ways. Additionally, there are many different methods researchers and practitioners have utilized in attempting to measure this concept. Given that conceptualizations and methods vary and are evolving, we have particular responsibility to provide validity evidence of the effectiveness of any method we choose (rating scales, tests, interviews, etc.). Examining the validity of the concept of executive function is much hard than establishing its reliability because of the many different ways the concept has been defined in the literature (see Chap. 1 in this volume). Thus, the author of any measure of executive function defines the concept by the questions that are included. This can provide a broad or truncated view of how executive function could be measured.

Establishing concurrent validity is especially difficult for a rating scale of executive function because of the variability in the way the concept has been conceptualized and measured. That is, the author has to decide what marker test should new tests be compared to? The variability in conceptual and measurement approaches used by different authors will have direct influence on any research findings that may be found as well as the psychometric quality of the tests and methods used. Researchers and clinicians should be mindful that until there is sufficient maturity in the concept and tools used to measure executive function, any and all method should be examined carefully and high psychometric quality demanded.

Development of Scales to Assess Executive Function

There is a need for a number of well-standardized measures of executive function with demonstrated reliability and validity. At this time, there are only a few published behavior rating scales for clinical use possessing varying degrees of reliability and validity (a detailed analysis of these will be provided later in this chapter). Given the relatively small number of options, there is a need for practitioners and researchers to have an understanding of the psychometric qualities of these tools. It is particularly important to pay close attention to the development methods used by the authors of any scale intended to be used to evaluate executive function. Development of a rating scale should follow a series of steps to ensure the highest quality and validity. The development of such a scale is a task that demands well-known procedures amply described by Crocker and Algina (1986) and Nunnally and Bernstein (1994) that are summarized in the following section.

Initial development of a rating scale for executive function should begin with a clear definition of the concept or concepts and the behaviors that can be used to assess them. The items used to evaluate these behaviors must be written with sufficient clarity that they can be answered reliability over time and across raters. Items should be included that represent the author's carefully defined view of executive function.

The first test development step is to prepare an initial pool of items. The main goal at this initial stage is to evaluate the clarity of the directions and items and manage other logistical issues. For example, it is important to determine raters' reactions to the size of the fonts, clarity of the directions, colors used on the form, and position of the items on the paper.

Once initial testing is completed, a larger study of the items can be conducted. This research effort helps determine if there is confidence that the items have been adequately operationalized and the following information is obtained:

- Means and SDs and difficulty of each item should be obtained.
- The contribution each item has to the reliability of the scale(s) on which it is placed should be evaluated.
- Items designed to measure the same construct should correlate with other items designed to measure that same construct higher than items designed to measure different constructs. If this is not found, the item may be eliminated.
- The internal reliability of those items organized to measure each construct should be computed, as should the reliability of a composite score.
- The factor structure of the set of items may be examined to test the extent to which items or scales form groups, or factors, whose validity can be examined.

The number of times preliminary research studies are conducted depends upon the results of the statistical analysis which in turn is dependent upon the quality of the (a) original concepts, (b) initial pool of items, and (c) the sampling used to study the instrument. The results of these efforts should be used to develop an experimental version of an instrument that is ready to be used in a larger national standardization study. This would include sufficient data to establish quality norms and also to conduct a research program to examine the reliability and validity of the final scale.

Standardization and norms development requires that a sample represents the population of the country in which the scale will be used. Standardization samples are designed to be representative of the normal population so that those that differ from normality can be identified and the extent to which they differ from the norm (50th percentile) can be calibrated as a standard score. Dispersion from the mean should also be calculated. Development of norms is an art as much as a science. There are several ways in which this task can be accurately accomplished (see Crocker & Algina, 1986; Guilford & Fruchter, 1978; Nunnally & Bernstein, 1994; Thorndike, 1982).

The second task of national standardization efforts includes analysis of data for establishing reliability (internal, test retest, inter-rater, intrarater) and validity (e.g., construct and content).

Of these two, validity is more difficult to establish and should be examined using a number of different methodologies and to assess the extent to which the scores the scale yields is valid for the purposes for which it is intended. The many different types of validity studies needed to fully evaluate any scale make it impossible to establish validity by a single study. According to the Standards for Educational and Psychologist Testing (AERA, APA, NCME, 1999), evidence for validity "integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses" (p. 17). There are 24 standards that relate to validity issues that should be addressed by authors and test development companies. Some of the more salient issues include the need to provide evidence that supports:

- The interpretations based on the scores the instrument yields
- The utility of the measure across a wide variety of demographic groups or its limitations based on race, ethnicity, language, culture, and so forth
- The appropriate relationships between the scores of the instrument with one or more relevant criterion variables
- The expectation that the scores provided differentiate between groups as intended
- The alignment of the factorial structure of the items or subtests with the scale configuration provided by the authors

Documentation in test manuals of scale development often focuses on construction, standardization, reliability, and validity. Reporting this information is important, but authors also have the responsibility to inform users about how the scores should be interpreted (AERA, APA, NCME, 1999). This includes how test scores should be compared. It is especially important to provide the values needed for significance when the various scores a rating scale provides are compared, for example, across raters. This information is critically important if clinicians are to be expected to interpret the scores from any instrument in a manner that is psychometrically defensible.

Professionals have a responsibility to choose scales that have been developed using the *highest* standards available because important decisions will be made on the basis of the information these measures provide. This includes ample documentation of methods used to develop the measure as well as ample evidence of validity and explicit instructions for interpretation of the scores that are obtained. Because of the impact score interpretation has on those individuals who seek help from professionals in clinical practice, in addition to being reliable, tools used to evaluate any condition must have been standardized and the scores based on norms developed from a large sample that represents the country in which the scale is used.

Obtaining information about the psychometric characteristics of psychological and educational tests is a time consuming and sometimes confusing task. Some test manual information is clear and concise, and at other times it is hard to ascertain enough details to fully evaluate the results being presented. Comparisons across instruments are complicated by this inconsistency and the logistical task of collecting the information. In the remainder of this chapter, a systematic examination of the scales used to assess the behaviors associated with executive function will be provided. The goal is to be informative of the specific details associated with important issues such as reliability, validity, standardization samples, and norming procedures. The information provided is intended to include essential topics such as description of the scale and standardization characteristics provided by the authors in their respective test manuals. Following this summary, a commentary of the relative advantages of the scales is provided.

Descriptions of Rating Scales Used to Assess Executive Function

BRIEF Parent and Teacher Reports

The Behavior Rating Inventory of Executive Function (BRIEF; Gioia, Isquith, Guy, & Kenworthy, 2000) was designed to assess the behavioral manifestations of executive functions in children aged 5–18 years as rated by parents or teachers. The 86-item rating scale evaluates two general domains—Behavioral Regulation (Inhibit, Shift, Emotional Control) and Metacognitive Problem-Solving (Initiation, Task Organization/Planning, Environmental Organization, Self-Monitoring, Working Memory) across the eight interrelated sub-domains.

The normative group for the BRIEF-Parent and BRIEF-Teacher ratings was based on data obtained from 25 schools in the State of Maryland (12 elementary, 9 middle, and 4 high schools). The sample description in the manual is very limited, mainly focused on sex (approximately equal percentages of males and females) and race/ethnicity. Table 10.3 shows that the distribution of the normative sample by race/ethnicity is quite different from that in the US population. The sample was dominated by Whites and considerably underrepresented by Hispanics. Even after statistically weighting the samples that were obtained by race/ethnicity, the values for Hispanics are still considerably lower than the US population values based on the 2011 Census.

BRIEF-Self-Report

The Behavior Rating Inventory of Executive Function—Self-Report (BRIEF-SR; Guy, Isquith, & Gioia, 2004) was designed to assess the behavioral manifestations of executive functions from self-reports of individuals aged 11–18 years. The 80-item rating scale evaluates two general domains—Behavioral Shift (Inhibit, Shift, Emotional Control, Monitor) and Cognitive Shift (Working Memory, Plan Organize, Organization of Materials, Task Completion) and eight sub-domains. Items are scored 1 (Never), 2 (Sometimes), and 3 (Often). Raw scores are converted to *T*-scores (mean of 50 and standard deviation of 15) and scaled so that scores above 70 are termed clinically significant. That is, the higher the score, the more difficulty with executive function is indicated. Two validity scales are also included an Inconsistency and Negativity Scale.

The BRIEF-SR norms were based on 1,000 11–18 year olds who completed the 80-item rating scale. The authors report in the manual that the sample was obtained through public and private school recruitment in the states of Maryland, Ohio, Vermont, New Hampshire, Florida, and Washington. These states do comprise the four regions of the country; however, the percentages of cases from each of these locations were not reported. Table 10.3 also illustrates that the distribution of the normative sample for African Americans was slightly underrepresented, Hispanics were underrepresented by about 40 %, and Whites were overrepresented by about 20 % in relation to the US population based on the 2011 Census. All five parental education levels deviated from the US population figures ranging from about 20 % for those without college experience to 65 % for those with bachelor’s degree. These characteristics suggest that the sample characteristics are quite dissimilar to the US population based on the 2011 Census.

BDEFS-CA

The Barkley Deficits in Executive Function Scale—Children and Adolescents (BDEFS-CA, Barkley, 2012) was designed to assess the behaviors associated with executive functions as rated by parents of their children aged 6–17 years. The xx-item rating scale provides scores for five scales: Self-Management to Time, Self-Organization/Problem Solving, Self-Restraint, Self-Motivation, and Self-Regulation. Items are scored on a five-point Likert scale. Raw scores are converted to percentile scores for each subscale and an EF Summary Score. The scores are scaled such that the higher the score, the more the deficit in executive function.

The normative sample for the BDEFS-CA was obtained from parent raters distributed across the four regions of the USA with fairly equal proportions to the overall population. As shown in Table 10.3, the distribution of the normative sample by race/ethnicity, however, is substantially different from that in the US population. The sample was dominated by Whites and

Table 10.3 Number of items, age range, normative sample size, and percentages of normative sample by region, race/ethnicity, and educational level for the BRIEF, BDEFS, D-REF, and CEFI

	BRIEF-Parent	BRIEF-Teacher	BRIEF-Self-report	BDEFS-CA (parent)	D-REF parent	D-REF teacher	D-REF self report	CEFI-Parent	CEFI-Teacher	CEFI-Self report	US Pop %
Scale description											
No. of items	86	86	80	70	36	36	36	100	100	100	
Age range	5–18	5–18	11–18	6–17	5–18	5–18	11–18	5–18	5–18	12–18	
Standardization											
Sample size	1,416	720	1,000	1,922	500	342	220	1,400	1,400	700	
Region											
Northeast	0	0	–	18	16.1	12.2	5.4	16.0	16.2	16.0	17.0
Midwest	0	0	–	28	15.6	19.3	13.9	22.1	22.0	22.0	21.7
South	100	100	–	31	58.6	57.2	77.8	37.9	38.0	38.0	37.2
West	0	0	–	23	9.8	11.3	2.9	24.1	24.0	24.0	24.1
Race/ethnic											
Asian	3.8	6.1	(In other)	–	–	–	–	4.0	3.8	4.0	4.2
Black	11.9	13.5	14.7	7.7	16.5	19.8	5.4	14.0	14.0	14.0	13.9
Hispanic	3.1	4.2	12.5	12.4	19.2	15.8	13.9	22.0	22.0	22.0	21.2
White	80.5	72.1	67.3	73.0	58.0	56.4	77.8	56.0	56.5	56.0	56.5
Other	.5	.4	5.5	–	6.2	8.1	2.9	4.0	3.7	4.0	4.2
Parental education level											
<High school	–	–	12.1	4.1	9.2	10.0	7.9	14.1	–	13.9	14.7
High school grad	–	–	33.6	28.1	26.0	28.9	26.6	28.0	–	28.0	28.5
Some college	–	–	12.4	29.8				30.0	–	30.0	28.9
Bachelor's degree	–	–	29.0	22.6				18.0	–	18.1	17.6
Graduate degree	–	–	12.9	15.4	64.9	61.1	65.5	10.1	–	10.0	10.3

Notes: US population percentages based on 2009 Census. Percentages by race/ethnicity for BRIEF-Teacher reported in the manual do not sum to 100 %. D-REF values were averaged across age groups considerably underrepresented by Hispanics and Blacks. Importantly, the parental education levels are not consistent with the US population figures. There are too few cases with parental education levels less than high school and too many from the top two educational attainment categories. This disparity on these two important demographic variables indicates that the characteristics of the DBEFS-CA normative sample are quite dissimilar to the US population based on the 2011 Census.

D-REF

The Delis Rating of Executive Functions (D-REF; Delis, 2012) is a set of rating forms designed to assess executive functioning in individuals ages 5–18. The scale has three forms: Parent, Teacher, and Self, each comprises 36 items. The D-REF is designed to evaluate a child or adolescent's behavioral, emotional, and executive functioning in four specific areas of executive functioning: Attention/Working Memory, Activity Level/Impulse Control, Compliance/Anger Management, and Abstract Thinking/Problem-Solving. Raw scores are converted to *T*-scores (*Mn* = 50; *SD* = 10) for each of the four index scores (low scores suggest better executive function).

The normative samples for the parent, teacher, and self scales of the D-REF were distributed across the four regions of the USA with varying correspondence to the overall population. For example, the parent, teacher, and self-rating samples underrepresented cases in the West and overrepresented cases from the South considerably. The cases from the Northeast were also very underrepresented (see Table 10.3). The inclusion of cases by race/ethnicity was also problematic. For example, Blacks were very underrepresented in the self-rating sample. Hispanic groups underrepresented in the parent, teacher, and self-rating samples, and Whites were very overrepresented in the self-rating sample. The sample by parental education was underrepresented for those with less than high school education and overrepresented for those with greater than a high school education. These differences in demographic variables indicate that the characteristics of the D-REF normative samples are substantially inconsistent with the characteristics of the US population based on the 2011 Census.

Comprehensive Executive Function Inventory

The Comprehensive Executive Function Inventory (CEFI, Naglieri & Goldstein, 2013) is a rating scale designed to evaluate observable behaviors that are related to executive function. The CEFI is completed by parents (or similar caregiver) or teachers (or similar professional) who rate behaviors of children ages 5–18 years. There is also a self-report version for 12–18 year olds. The 100 items of the CEFI items are organized on the basis of their content into nine scales (Attention, Emotion Regulation, Flexibility, Inhibitory Control, Initiation, Organization, Planning, Self-Monitoring, and Working Memory). A total (Full Scale) is also included. In addition, three scales that evaluate the quality of the ratings are provided: one that examines the consistency of the ratings (Consistency Index), one that is designed to assess the likelihood that the rater's scores are overly negative, and one that suggests an overly positive view of the person being evaluated (Negative and Positive Impression Scales, respectively). Each of these scales is scaled to have a normative mean of 100 and SD of 15 where higher scores indicate better executive function.

The norms for the CEFI were based on the standardization sample including cases from all 50 states in the USA. The normative samples included 1,400 for the Parent Form, 1,400 for the Teacher Form, and 700 for the Self-Report Form. The stratifications by region, race/ethnicity, and parental education were within one percentage point of the values for the US population. The report of demographic variables indicates that the characteristics of the CEFI normative samples are very consistent to the US population based on the 2011 Census.

Normative Sample Disparities

In this chapter we have discussed various psychometric characteristics of rating scales and the samples upon which their norms were based. It is clear from this summary that some of the demographic characteristics of the samples upon which the derived scores were based vary considerably. It is reasonable to ask, does this matter? In order to examine the impact a variable such as parental education can have on the resulting normative scores, norms for parent ratings from the CEFI (Naglieri & Goldstein, 2013) were prepared for four different groups by parent educational levels (PEL). This study began with the calculation of means and standard deviations for the standardization data ($N = 1{,}400$) for PEL levels reported in the manual. The mean raw scores were 252.1 (no high school diploma), 269.2 (high school diploma), 280.3 (some college), and 285.6 (bachelor's degree or higher). Using these raw scores, standard scores were computed using the formula ((raw score– *mean* raw score)/raw score *SD*)*15+100. The scores were calibrated so that high standard scores indicated better executive functioning.

The resulting values presented in Table 10.4 illustrate how much differences in CEFI total scale scores vary across parental education. The raw scores associated with a standard score of 100 vary from 250 to 285 (35 points) across the four PEL levels. The difference between the score of 100 based on the total sample and the lowest PEL level is 6 standard score points which is nearly half a standard deviation. Of particular importance are the differences that are found at the standard score of 85, which indicates a very poor score on this scale of executive function. The same raw score of 210 yields a score of 85 when based on the total sample, but a standard score of 92 (which falls in the average range) when the reference group is those with less than a high school education and a score of 81 when the highest education level is used as a reference group. The 11 point difference between the 81 and 92 represents the range of scores that can be expected due to the influence of PEL on CEFI scores.

Table 10.4 Calibration of standard scores ($Mn = 100$; $SD = 15$) across parental education levels for CEFI parent ratings

	Standard scores				
Raw score	Less than high school	High school graduate	Some college	College graduate	Total sample
180	**85**	80	76	74	79
185	86	81	77	75	80
190	87	82	79	76	81
195	88	83	80	77	82
200	90	**85**	81	79	83
205	91	86	82	80	84
210	92	87	83	81	**85**
215	93	88	**85**	82	86
220	94	89	86	84	88
225	95	90	87	**85**	89
230	96	91	88	86	90
235	97	92	89	87	91
240	98	93	90	89	92
245	99	95	92	90	93
250	**100**	96	93	91	94
255	101	97	94	92	95
260	102	98	95	94	97
265	103	99	96	95	98
270	104	**100**	98	96	99
275	105	101	99	97	**100**
280	106	102	**100**	99	101
285	107	103	101	**100**	102
290	108	105	102	101	103
295	109	106	103	102	105
300	110	107	105	104	106
305	111	108	106	105	107
310	112	109	107	106	108
315	113	110	108	107	109
320	114	111	109	109	110
325	115	112	111	110	111
330	116	114	112	111	112

Note: Standard scores of 100 (at the normative mean) and 85 (one standard deviation below the mean) are in bold text

The variability of standard scores obtained across levels of parental education illustrates the importance of having a normative sample that closely represents the US population. Of course, the results presented here represent only one variable. In those standardization samples that are not representative of the US population on more than one variable, the potential impact on the resulting scores, and the decisions made by practitioners when evaluating executive function, cannot be ignored. For this reason, it is advised that only measures that have been normed on a nationally representative sample that closely corresponds to the US population should be used in professional practice as well as research to ensure accurate calibration of an individual's executive function.

Conclusions

The information summarized in this chapter provides clinicians and researchers with information about the psychometric characteristics of rating scales used to assess behaviors associated with the concept of executive function. Special attention was paid to the quality of the standardization samples used to create norms. The information provided here illustrates very different approaches to test development and the quality of the standardization samples used to create the norms. For example, some of the scales are short (the D-REF has 36 items) while others such as the CEFI contain many items (100). Some authors provide only percentile scores (BDEFS-CA) which make use of the scores in any mathematical formula difficult, others (e.g., BRIEF) provide *T*-scores scaled so that high scores indicate more deficits in executive function, and others (CEFI) use the familiar mean of 100 and standard deviation of 15 where high scores indicate better executive function. Although these rating scales of behaviors related to executive function all strive to evaluate essentially the same concept, the characteristics of the samples upon which their derived scores are based reflect a fundamental difference in test development. That is, some normative samples are more closely matched to the US population characteristics than others. The closer the samples are to the US population, the more confidence users can have with the obtained scores and the greater likelihood that accurate and valid information can be obtained.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association, American Psychological Association, National Council on Measurement in Education.

Barkley, R. A. (2012). *Barkley deficits in executive function scale-children and adolescents*. New York: Guilford Press.

Bracken, B. A. (1987). Limitations of preschool instruments and standards for minimal levels of technical adequacy. *Journal of Psychoeducational Assessment, 5*, 313–326.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Hold, Rinehart and Winston.

Delis, D. C. (2012). *Delis rating of executive functions*. Bloomington, MN: Pearson.

Gioia, G. A., Isquith, P. K., Guy, S. C., & Kenworthy, L. (2000). *Behavior rating inventory of executive function*. Lutz, FL: Psychological Assessment Resources, Inc.

Guilford, J. P., & Fruchter, B. (1978). *Fundamental statistics in psychology and education*. New York: McGraw Hill.

Guy, S. C., Isquith, P. K., & Gioia, G. A. (2004). *Behavior rating inventory of executive function—Self-report version*. Lutz, FL: Psychological Assessment Resources, Inc.

Naglieri, J. A., & Das, J. P. (1997). *Cognitive assessment system interpretive handbook*. Austin, TX: ProEd.

Naglieri, J. A., & Goldstein, S. (2013). *Comprehensive executive functioning index*. Toronto: Multi Health Systems.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. New York: McGraw-Hill.

Thorndike, R. L. (1982). *Applied psychometrics*. Boston: Houghton-Mifflin.

Wechsler, D. (2003). *Wechsler intelligence scale for children fourth edition*. San Antonio: Pearson.