Behavior Intervention Monitoring Assessment System (BIMAS™)

8 Development

The Behavior Intervention Monitoring Assessment System (BIMAS™) development project began in 1997 with a series of empirical studies aimed at identifying items that are sensitive for monitoring changes in response to intervention. The project encompassed three years of data collection wherein thousands of ratings from teachers, parents, and clinicians as well as students' self-ratings were gathered from multiple data collection sites. Intensive research and sophisticated statistical analyses informed each step of the instrument's development, as described in this chapter.

First, the rationale and goals for BIMAS development are outlined. The preliminary research forming the basis for the development of the BIMAS is summarized, including a description of: (1) the Intervention Item Selection Rules (IISRs) used for BIMAS item selection, (2) the preliminary structure and item development process, and (3) the development of the BIMAS response schemas. Next, final scale construction during the normative phase of development is described. The normative sample is briefly described, followed by a description of the BIMAS Standard and BIMAS Flex development. Next, the criteria for developing BIMAS Standard Item Descriptors scores are presented. Finally, the creation of the final forms is described.

Rationale and Goals

The BIMAS development process began with the identification of a need for a specific kind of test and a set of test-related goals and purposes. The following section describes the rationale and development goals for the BIMAS.

Rationale

Initial discussions regarding the development of the BIMAS occurred in the context of two major professional trends that highlighted the need for a measure like the BIMAS. First, efforts towards cost containment of health services, including mental health services, led to the creation and rapid spread of managed care companies. Managed care initially slowed health care expenditures, but its creation also brought along a host of negative effects on the provision of behavioral health services (Davis & Meier, 2001); for example, many counselors and counseling agencies, in settings from private practice to college counseling centers, now routinely employ brief therapy. This trend towards brief therapy has put pressure on mental health practitioners to deliver the most cost-effective treatment. Often times, costs (from the managed care company's standpoint) and benefits (from the client's perspective) is a difficult balancing act. As a result, managed care's emphasis on increased accountability has led to a greater focus on more objective measurements of client progress and outcomes that has the potential to improve both science and practice efforts.

In the schools, the 2004 reauthorization of the Individuals with Disabilities Education Improvement Act (IDEA) further emphasized the importance of prevention, early intervention, and accountability as related to special education and intervention services. As a result, educators have been employing Response-to-Intervention (RTI) approaches in the development and monitoring of their programs. As described in chapter 2, *Background*, RTI is a 3-Tier model that integrates assessment and intervention; schools identify students at risk for poor learning outcomes through universal screening and provide evidence-based interventions to maximize student achievement and to reduce behavior problems. The intensity and nature of interventions can be adjusted depending on students' responsiveness to intervention. As a result, the movement of RTI has led to a greater need for measures that can be employed to screen students for behavioral health problems as well as monitor student progress and outcomes resulting from behavioral and psychosocial interventions.

Whether in schools, clinics, or other settings, intervention approaches like RTI require valid instruments that can track a youth's status over time. The main rationale for the development of BIMAS was to produce a measurement system that could be easily administered multiple times within the RTI system for screening, progress monitoring, program evaluation, and data-based decision making.

Development Goals

The following goals were established for the BIMAS development project:

1. Develop a brief, repeatable, psychometrically sound measure for universal screening, progress monitoring, and program evaluation.

2. Provide a change-sensitive measure based on an empirically supported model for selecting change-sensitive items, specifically, Meier's (1997, 1998, 2000, 2004) Intervention Item Selection Rules (IISRs; see *Preliminary Research: Intervention Item Selection Rules* in the next section).
3. Include content to assess both adaptive behaviors as well as the most common behavioral and social-emotional concerns observed in childhood and adolescence.
4. Validate forms for use with multiple informants so that information may be collected across the various contexts and from different perspectives.
5. Create a user-friendly yet sophisticated web-based graphing and reporting system for easy data entry, automated reports, and graphic displays of data for screening, progress monitoring, and program evaluation.

Preliminary Development

Preliminary Research: Intervention Item Selection Rules

In an effort to create a measure that is change-sensitive, a series of studies were conducted (Meier, 1997, 1998, 2000, 2004; Meier, McDougal, & Bardos, 2008; Weinstock, & Meier, 2003), and this body of empirical work formed the basis for the development of the BIMAS. In the Meier et al. (2008) study, for example, parents of 896 elementary school-aged children receiving psychotherapy interventions from community mental health agencies completed a preliminary outcome measure assessing the youth's symptoms and functioning. Pre- and post-intervention data were examined to compare sensitivity to change between item groupings in scales. Results indicated that scales formed with change-sensitive items evidenced larger effect sizes than scales composed of the original item pool, and demonstrate adequate reliability estimates.

The process employed in the development of the BIMAS focused on identifying constructs that change as a result of emotional and behavioral interventions. To identify intervention-sensitive items either during test construction or item evaluation, Meier (1997, 1998, 2000, 2004) proposed a set of Intervention Item Selection Rules (IISRs). The central philosophy of the IISRs is that intervention-sensitive items should change in response to an intervention and behave in a theoretically expected manner in other conditions (e.g., remain stable over time when no intervention is present). This approach assumes that (a) test items and tasks differ along a trait-state continuum, and (b) different test construction and item analysis procedures are necessary to select items with a high state loading that reflect the results of interventions. IISRs are designed to test two broadly competing claims regarding change at the item level: such change is the result of an intervention, or the change results from other factors that constitute error in the context of scale development. Meier has developed and studied scales constructed with both traditional and IISRs procedures in a variety of clinical and school settings (e.g., Meier 1998, 2000, 2004). Overall, scales constructed with IISRs procedures demonstrated larger treatment effect sizes than traditional scales, and produced adequate reliability estimates (Meier, McDougal, & Bardos, 2008). Items on change-sensitive measures will share some characteristics with traditional, trait-sensitive tests. Intervention-sensitive items, for example, should also be theoretically based, reliable, and unrelated to systematic error sources. However, intervention-sensitive items should possess additional properties, foremost of which is that they change in response to an intervention.

The nine IISRs are described below. The first two describe assumptions about the nature of intervention-sensitive items, and the remaining seven IISRs describe procedures for collecting and evaluating empirical data.

The IISRs are as follows:

1. Items are initially identified in a literature review such that they are theoretically grounded and related to relevant research.
2. Intervention-based items are aggregated across individuals, but not across items or occasions (as with trait-based tests). Aggregation across individuals decreases random error (Messick, 1989) and increases the possibility of detecting item scores that are responsive to intervention effects (Epstein, 1979, 1980).
3. The range of scores at pre-test is reviewed so that items demonstrating obvious ceiling or floor effects may be removed.
4. Scores on intervention-sensitive items over time must demonstrate change in intervention groups.
5. Scores on intervention-sensitive items over time must exhibit change in the expected direction. This ensures that the items are indeed change sensitive and are able to monitor change in the expected direction.
6. Change observed in an intervention group is investigated relative to a non-treated comparison group.
7. The items must not demonstrate differential effects between these groups prior to intervention.
8. The results of the item change must not be related to systematic error sources such as social desirability.

9. Steps 3 through 8 should be cross validated with repeated studies of new samples from the population of interest.

As indicated in step 9 above, the development of an evaluation instrument based on IISRs requires a recursive process of cross validating items both for screening and progress monitoring purposes with different samples of populations of interest. Towards this end, preliminary versions of the BIMAS were developed using both clinical and school samples in field settings, a combination rarely studied in psychotherapy research (Kazdin, 2000). In fact, Weisz, Huey, and Weersing (1998) found only nine studies (over a period of 50 years) that examined the effects of child treatment in applied settings. This is important because meta-analyses examining the effects of psychotherapy provided in laboratory settings versus field settings indicate that the former produced effect sizes around .75 and the latter around 0 (Weisz, Weiss, & Donenberg, 1992). Thus, these results imply that the psychometric properties of tests developed with non-clients in non-clinical settings may not be generalizable for use with children and adolescents receiving psychosocial interventions in applied settings. Since the BIMAS was developed to be used in school districts or organizational (applied) settings for the purpose of behavioral screening or progress monitoring, field studies utilizing both clinical and school samples add weight to the external validity of the tool.

In an unpublished intervention program evaluation study conducted in 1998, McDougal and Meier utilized an initial item set to evaluate outcomes of children who were receiving school-based mental health services in a number of schools in Syracuse, NY. Over the course of the evaluation, teacher ratings were obtained for 338 children receiving services as well as for 47 children without any behavioral or mental health concerns who served as a control group. Results of the evaluation served as impetus for further research, as children receiving services overall demonstrated improvement in teacher ratings over time while teacher ratings of the control group remained stable. Further, the composite means of children in the clinical group were significantly different ($p > .001$) from the means of non-clinical children, indicating that this item set might have some utility as a screening measure.

Initial published studies have found differences in the psychometric properties of items developed with IISRs and other more traditional item selection rules (Meier, 1998, 2000; Weinstock & Meier, 2003). For example, in an intervention study aimed at changing college students' attitudes toward alcohol use, Meier (1998) contrasted the IISRs approach with traditional item evaluation methods using a 16-item alcohol attitude scale. Application of the more traditional type of intervention item selection criteria versus the IISRs approach resulted in the creation of two sets of items with differing psychometric properties. The intervention-sensitive items demonstrated greater pre-intervention to post-intervention change. In another study, 116 parents completed a social skills scale at intake and follow-up periods as part of treatment for children in a community health center, it was also found that scales composed of items that met IISRs criteria had larger effect sizes (Meier, 2000). Furthermore, in an early study investigating preliminary change-sensitive items to be included on the BIMAS, Weinstock and Meier (2003) subjected both intake and follow-up item scores on a 56-item self-report checklist completed by 615 university counseling center clients to principal component analysis (PCA) as well as an IISRs evaluation. As predicted, items selected using IISRs demonstrated larger effect sizes (when combined into scales) than items selected through PCA. Taken together, the results from these three empirical studies provide evidence that the more traditional use of trait-based methods of item selection and evaluation that focuses on detecting individual differences may be problematic for measuring constructs that should theoretically change in response to intervention.

More recently, Meier, McDougal, and Bardos (2008) studied a preliminary version of the BIMAS using parent ratings of 896 youth clients (Kindergarten through Grade 10) receiving psychotherapy at a community clinic. Results indicated that this early version of the BIMAS was a change-sensitive scale which demonstrated adequate reliability, larger effect sizes between pre- and post-intervention group means than most assessment tools created for other purposes (e.g., ones that diagnose a stable trait-like construct), and that it was brief enough to be administered repeatedly for the purpose of progress monitoring.

Preliminary Structure and Item Development

The change-sensitive items found in Meier's (1998, 2000, 2004) research were used as a preliminary pool of items for the BIMAS. The items from these studies could be grouped into two categories: Distress/Problems and Strengths. Items within the Distress/Problems category included feelings of depression, behaving differently than peers, impulsivity, fidgeting, fighting with others, and failing grades. Items within the Strengths category included content related to interpersonal communication, paying attention to others when engaged in a conversation, ability to make friends, and helping with household tasks. More items were developed following a review of the literature on children's behavioral problems (Stiffman, Orme, Evans, Feldman, & Keeney, 1984), reviews of other measures of children's distress and functioning (Meier, 1998), and suggestions from mental health professionals (e.g., school counselors, psychologists, and social workers).

Using clinical judgment and rationale, the initial item pool was examined to (a) categorize items by similar content, (b) determine if item content was consistent for each rater

type (e.g., teacher, parent, or student) or if there was a need to revise or delete certain items on a particular rater form, and (c) examine if item content made sense for different age groups. Additional items were also created as the need arose to meet the goals of (a), (b), and (c).

Next, the initial item pool was categorized on the basis of face validity, into four domains. *Externalizing* behaviors referred to items assessing conduct problems, substance abuse, and deviant behaviors, while *Internalizing* behaviors focused on negative affect related to anxiety and depression. *Cognitive Processing* items focused on themes related to attention, focus, bizarre thoughts as well as behaviors. Finally, *Adaptive Functioning* items consisted of a broad category of content that included academic, social, and interpersonal functioning. With all items, item content was revised or created so that they were observable or behavioral in nature (e.g., "appeared sad" was employed instead of "felt sad").

Another goal for the BIMAS was to create forms, with matching content and roughly equal numbers of items, for use by different raters. A final number of about 30 to 35 items per rater form was desired to keep the BIMAS brief and suitable for a one-page format. Similar to other scales developed for screening and monitoring purposes, comparable forms that can be completed by teachers, parents, clinicians, and youth were created. Given the desire to create parallel forms for these rater types, the appropriateness and wording of item contents for each informant were reviewed. For content related to school (e.g., received failing grades, came prepared to class), clinicians who work with the youth in a community mental health setting may not have access to this information; consequently, these items were dropped from the Clinician form. Some of these school-related items were retained but others were dropped from the Parent form as well. Similarly, the Clinician form contains an item relevant to therapy content (i.e., attended therapy session) that was irrelevant for other raters. The vast majority of items, however, assessed information deemed to be accessible to all four informants.

Item content was also reviewed in relation to age appropriateness. While the item content domains remained the same across rater forms, attention was paid to ensure that item wording was appropriate for the age span across Kindergarten through Grade 12 (Ages 5–18). In addition, a review of item content and previous research indicated that a self-report version for youth below Grade 7 was unlikely to evidence sufficient reliability and validity estimates; therefore, this version of the self-report was dropped.

Response Schemas

Instructions for raters on the BIMAS forms ask the respondent to assess how often a child or adolescent manifested each of the behaviors during the past week. Response choices include 0 = Never, 1 = Rarely, 2 = Sometimes, 3 = Often, and 4 = Very Often. Each of these descriptions is paired with a frequency of occurrence during the past week. Never (the response of 0) is paired with 0 times or never observed; Rarely (response = 1) with 1 or 2 times or to a minimal extent; Sometimes (response = 2) with 3 to 4 times or to a moderate extent; Often (response = 3) is 5 to 6 instances or to a significant extent; and Very Often (response = 4) is 7 or more times or to an extreme extent. In previous research with a preliminary version of the scale (see Meier et al., 2008), the pairing of behavioral frequency ratings with the category descriptors was found to increase the mean and standard deviation of the resulting scale scores compared to data collected with a version using the category descriptors alone (Meier, 2008). By pairing the number of occurrence of a behavior (frequency count), a more objective measure, with the category descriptors (an overall impression of the extent of behavior), the authors believe that respondents would be better able to gauge behaviors using different dimensions, thus increasing the change-sensitivity of the BIMAS.

Creation of the Forms for the Normative Study

Results from the preliminary research and rational analysis of the items guided item selection for the standardization version of the tool. This process resulted in 36 items for the normative version of the BIMAS on the Teacher, Parent, and Self forms and 34 items on the BIMAS Clinician form. The order of the items for each form was randomized using random-number-generating functions (StatSoft, 2001), resulting in new forms for the normative study.

Final Scale Construction

Development of the final scales involved collection of data for normative and clinical samples, factor analyses to determine the factor structure of the forms, and the development of the Behavioral Concern scales, Adaptive scales, and Flex items.

Data Collection

Data for the normative versions of the BIMAS forms were collected from the general population and from selected clinical groups. The extensive data collection project resulted in normative samples that included ratings from 1,400 teachers and 1,400 parents (on youth aged 5 to 18 years), and 700 youth (aged 12 to 18 years). The normative samples were representative of the general U.S. population in terms of age, gender, race/ethnicity, geographic location, and parental education level (parent version only) in accordance with the 2000 U.S. Census (see chapter 9, *Standardization*, for a full description of the normative samples). Additionally, 538

teacher ratings, 467 parent ratings, and 350 youth self-ratings were collected for youth with a clinical diagnosis (diagnoses included Disruptive Behavior Disorders, Attention-Deficit/Hyperactivity Disorder, Anxiety Disorders, Depression, Pervasive Developmental Disorders, Learning Disorders, and Developmental Delays; see *Clinical Samples* in chapter 11, *Validity*, for a description of the clinical groups). More detailed descriptions of the normative sample, the standardization process, and the psychometric properties of the BIMAS are provided in chapters 9 to 11, *Standardization*, *Reliability*, and *Validity*.

Development of the BIMAS Standard

Data from the normative and clinical samples were analyzed to confirm the item content of the BIMAS Standard. Item-level analyses conducted on the 36 BIMAS Standard items revealed that two items (i.e., "behaved differently than his/her peers" and "expressed strange or bizarre thoughts") showed extreme floor effects (i.e., they had very little variability and were almost always rated as 0 or 1), and were therefore cut from the BIMAS Standard. As a result, the final number of items on the Teacher, Parent, and Self forms is 34 and that for the Clinician form is 31 (four items tapping academic functioning were replaced by one item on the attendance of therapy appointments). In order to verify the proposed scale structure of the BIMAS Standard (see *Preliminary Structure and Item Development*, earlier in this chapter), these final items were subjected to confirmatory factor analyses (CFA using Maximum Likelihood, generalized least squares estimation). Results indicated adequate model fit (see *Content Validity and BIMAS Scale Structure* in chapter 11, *Validity*, for more information on the confirmatory factor analysis procedure and results). Decisions regarding the final scale structure were made based on the authors' clinical experience, the research literature, as well as results from the CFA.

Development of the Behavioral Concern Scales

The statistical and rational analysis of the items resulted in three main areas pertaining to behavioral distress/problems: externalizing, internalizing, and cognitive/information processing. Further review of the item content (all negatively worded) contributed to the renaming of the BIMAS Behavioral Concern scales. The Conduct scale (9 items) taps issues related to anger management, bullying behaviors, substance abuse, and deviance (e.g., "appeared angry"). The items within the Negative Affect scale (7 items) assess anxiety and/or depressive types of symptoms (e.g., "acted sad/withdrawn"). The Cognitive/Attention scale (7 items) looks at issues related to attention, focus, organization, planning, and memory (e.g., "had trouble paying attention"). Refer to appendix A for a full list of items on each BIMAS Behavioral Concern scale.

Development of the Adaptive Scales

The remaining 11 items on the BIMAS forms had been previously identified as adaptive functioning types of items (see the *Preliminary Structure and Item Development* section earlier in this chapter). These are mostly positively-worded items that tap adaptive skills pertaining to social and academic functioning (note that on the Clinician form, two items that may contribute to therapeutic gains replace the academic functioning items). Having strength-based scales alongside the Behavioral Concern scales may assist in the development of an intervention plan by utilizing the youth's strengths to overcome potential areas of difficulty. As a result, rather than reverse-scoring all the positively worded items to calculate a scale score that would indicate a behavioral concern or a lack of adaptive skills, all positively worded items are scored as they are worded—higher scores mean that the behaviors in question are more frequently observed. (To this end, two negatively worded items "received failing grades at school" and "was absent from school" were determined to require reverse-scoring.) The Adaptive Scales include: the Social scale (6 items), which taps communication, friendship maintenance, and interpersonal skills (e.g., "shared what he/she was thinking about") and the Academic Functioning scale (5 items), which assesses a youth's academic performance, attendance, and attitude in learning (e.g., "worked up to his/her academic potential"). As mentioned before, four of the items on the Academic Functioning scale, with the exception of "followed directions," were items that pertain more to an academic setting rather than a clinical setting. Instead, the "followed directions" item as well as the item "attended his/her scheduled therapy appointments" composed two standalone Clinician Adaptive items on BIMAS–Clinician. Refer to appendix A for a full list of items on the BIMAS Adaptive scales.

Development of the BIMAS Flex Items

The BIMAS Flex items were designed to target specific behavioral concerns or skills as measured by the standardized BIMAS Standard items. While the BIMAS Standard was designed to provide information on the individual's progress across standardized scales, the Flex items were designed to provide information on specific intervention targets, such as "interacted well with friends" or "worked out problems with peers by him/herself," serving as the basis for Individualized Education Program (IEP) or treatment plan goals. On average, each BIMAS Standard item across the five BIMAS scales has a pool of 10–15 corresponding positively or negatively worded Flex items for assessors

to choose from. Altogether, a total of 655 Flex items were developed for each of the BIMAS Standard Teacher, Parent, and Self-Report forms; and for the Clinician form, a total of 586 Flex items were developed.

Because the Flex items were conceived as a tool to help assessors develop and monitor a set of behavioral goals for a youth, several goal setting guides in the literature, such as the SMART Goals Setting Guide proposed by Nikitina (2006), guided the development of the Flex items. This model suggests that each intervention goal should be:

1. **Specific.** Goals need to be straightforward, specific, clear and easy, and emphasize what needs to happen by answering "who, what, why, and how."
2. **Measurable.** Establish concrete criteria for measuring progress so you can see the change occur. "If you can't measure it, you can't manage it."
3. **Attainable.** Goals need to challenge the individual, and the individual must believe that he/she will be able to meet the challenge.
4. **Realistic.** Goals that are unrealistic and set too far out of reach will not motivate an individual to commit to intervention and will set the individual up for failure. Change does not usually occur overnight, so devise a plan or way of getting to the goal that is realistic.
5. **Timely.** Set a timeframe for the goal that gives the individual a clear target to work towards.

Appendix B presents the Flex items for each BIMAS Standard item.

Determining the Scoring Criteria of Item Descriptors for BIMAS Standard Items

Both the BIMAS Standard and BIMAS Flex provide Item Descriptors to guide the interpretation of item-level results: *Concern, Mild Concern,* and *No Concern* on Behavioral Concern Scales; *Concern, Mild Concern, Fair,* and *Positive* on Adaptive Scales, since only the BIMAS Standard is norm-referenced, this section regarding the development of the Item Descriptors pertains only to BIMAS Standard Item Descriptors[1].

Each item on the BIMAS Standard has been standardized on a U.S. national sample closely matching the most recent U.S. Census in terms of demographics in much the same way as the scale-level norms were developed. This standardization process allows assessors to compare a youth to his/her normative group on an item-to-item basis. At the item level, the approach used to designate the risk levels of a BIMAS Behavioral Concern scale item score is similar to the technique used by Naglieri, McNeish, and Bardos (1991), Naglieri, LeBuffe and Pfeiffer (1994), and LeBuffe and Naglieri (2003); all suggested that an individual item score that falls in the top 15% of the normative group distribution (e.g., exceeds the mean normative item score plus one standard deviation) can be considered problematic. To this end, means and standard deviations were calculated for each normative age group (5–6, 7–9, 10–11, 12–13, 14–16, and 17–18 on BIMAS–T & BIMAS–P; 12–13, 14–16, and 17–18 on BIMAS–SR) for each of the 34 BIMAS Standard items for both combined gender (default scoring option) and gender-specific groups. After that, cumulative percentiles were calculated for the various age groups for each Standard item using both combined and gender-specific groups. An Item Descriptor was assigned for each possible item score (0 to 4) on each of the Standard items, separately for each rater form. This descriptor was based on the mean, standard deviation, cumulative frequency, and percentile (see Tables 8.1 and 8.2). In some cases, clinical judgment on item content in relation to gender, age, and observer- vs. self-report differences also assisted in the final Item Descriptors. The following rational rules were considered:

1. *Mild Concern/Concern* should never be assigned for a rating of 1 = Rarely unless the behavior in question is a serious one (e.g., "thoughts of hurting self").
2. If the response frequency for an item score is more than 25% in the normative sample (e.g., more than 25% of respondents answered a rating of 2 = Sometimes), then that item score should not be flagged as a *Mild Concern/Concern.*
3. For borderline instances, consistency should be maintained as much as possible across normative age groups and across raters (i.e., BIMAS–T, BIMAS–P, and BIMAS–SR).

Behavioral Concern Scales

As mentioned above, a combination of rationale and data, based on individual item means, standard deviation, cumulative frequency, and percentile across age groups, contributed to the development of the Item Descriptors. Table 8.1 illustrates the data-driven guidelines for the Item Descriptors on the Behavioral Concern scales. The Item Descriptor (e.g., *Concern/Mild Concern/No Concern*) obtained by a rater's item score (0 = Never, 1 = Rarely, 2 = Sometimes, 3 = Often, or 4 = Very Often) for any given item depends on the item content, type of rater (Teacher/Parent/Self) as well as the youth's age group (and gender if gender-specific norms were used as the scoring option instead of the default combined-gender norms).

[1] The same Item Descriptors are used on BIMAS Flex reports where assessors set up their own item-level scoring criteria based on what item response is considered to be "concern" or "typical" for an individual youth. Assessors are advised to consult the information on how the item-level scoring criteria were developed for the norm-based BIMAS Standard form to help inform the setup of Flex item scoring criteria in accordance to their own school-wide expectations or the individual youth's behavioral status.

Table 8.1. Development Guidelines for Item Descriptors: Behavioral Concern Scales

Item Descriptor	Development Guideline
Concern	> *M* + 1 *SD* or ≥ 85th percentile
Mild Concern	= *M* + 1 *SD* or between 75th–84th percentile
No Concern	< *M* + 1 *SD* or < 75th percentile

Note. Development guidelines are in reference to the normative sample's distribution

In general, on any of the Behavioral Concern scales (i.e., Conduct, Negative Affect, or Cognitive/Attention) a *Concern* item score indicates that the behavior was rated as occurring much more frequently than observed amongst most youth in a comparable age group. A *Mild Concern* item score indicates that the youth's behavior was rated as occurring slightly more frequently than in the normative group. On the other hand, a *No Concern* item score means that the behavior occurs at a frequency comparable to the youth's normative age group.

Adaptive Scales

Table 8.2 illustrates the data-based guidelines for the item-level norms on the Adaptive scales. Since higher scores on the BIMAS Adaptive scales indicate fewer concerns (because they consist of mostly positively worded items), there is an extra Item Descriptor added, *Positive*, to reflect adaptive behaviors which are considered positive or performing beyond the expected level of functioning. Thus, there are four Item Descriptors associated with the Adaptive Scales: *Concern, Mild Concern, Fair,* and *Positive*. Again, item content, type of rater (Teacher/Parent/Self) as well as the age group (and gender if gender-specific norms was selected as the scoring option) determines the specific Item Descriptor that can be associated with a rater's score (0 to 4) on any item.

A *Concern* item score on any of the Adaptive scales (i.e., Social or Academic Functioning) indicates that the adaptive behavior in question was rated as occurring *much less* frequently than observed amongst most youth in a comparable age group by the same rater. A *Mild Concern* score indicates that the youth's adaptive behavior was rated as occurring slightly *less* frequently than in the normative group. *A Fair* item score means that the youth is displaying the behavior at a frequency comparable to the normative group within the youth's age group. Lastly, a *Positive* item score may indicate that the youth displays the adaptive behavior more frequently than observed amongst most youth in a comparable age group by the same rater.

Table 8.2. Development Guidelines for Item Descriptors: Adaptive Scales

Item Descriptor	Development Guideline
Concern	< *M* – 1 *SD* or ≤ 10th percentile
Mild Concern	= *M* – 1 *SD* or between 11th–20th percentile
Fair	> *M* – 1 *SD* & < *M* + .67 *SD* or between 21st–74th percentile
Positive	≥ *M* + .67 *SD* or ≥ 75th percentile

Note. Development guidelines are in reference to the normative sample's distribution

Creation of the BIMAS Standard Final Forms

After a review of the normative data, the item counts were finalized: 34 items for the BIMAS–T Standard (5–18 Years), BIMAS–P Standard (5–18 Years), BIMAS–SR Standard (12–18 Years) and 31 items for the BIMAS–C Standard (5–18 Years). Appendix A presents the final list of items by scale for each BIMAS form.

9 Standardization

This chapter describes the process and methods used to develop the norms for the Behavior Intervention Monitoring Assessment System (BIMAS™) Standard teacher, parent, and self-report versions[1]. As a norm-referenced test, the BIMAS required a nationally standardized sample, which is essential to establish its psychometric qualities. This chapter is organized in the following sections: data collection, description of the normative samples, followed by a discussion on the norming procedures and the derivation of the standardized scores.

Data Collection

Data collection took place between March 2007 and May 2009. During the standardization and research phase of instrument development, thousands of BIMAS forms were completed; data from these forms were included in the standardization, reliability, and validity research studies (see chapter 10, *Reliability*, for a description of the reliability samples, and chapter 11, *Validity*, for a description of the validity samples). Twenty-five site coordinators throughout the U.S. assisted with the data collection. Site coordinators were recruited by contacting data collection sites known to the publisher, through author contacts, and through several mailing and email campaigns. Each site coordinator was instructed to follow standardized procedures that included obtaining informed consent, having raters follow specific written instructions, and debriefing the raters as needed upon completion of the assessment(s). All participants were aware that forms were being completed as part of the process of BIMAS test development. Site coordinators were compensated for taking part in the data collection process. Once the rating forms were returned to the publisher, they were subjected to visual inspections followed by some initial statistical analyses. Rating forms that had more than 10% of responses missing were excluded from the dataset.

For all assessments, in addition to ratings on the BIMAS items, demographic information about the rated youth was collected, including age, gender, parental education level (PEL; on the parent forms only), race/ethnicity, and geographic region. Race/ethnicity categories were: Asian/Pacific Islander, Black/African American, Hispanic, Native/Aboriginal, White, Multiracial, and Other. For ease of presentation, these groups are referred to herein in the following manner: Asian, African American, Hispanic, White, and Other (including Native/Aboriginal and Multiracial due to the small sample sizes of these groups). On the parent reports, data were collected on PEL for both parents, and the higher of the two was used to classify the PEL of the child.

Additionally, 538 teacher ratings, 467 parent ratings, and 350 youth self-ratings were collected from youth with a clinical diagnosis. In order to ensure the accuracy of all diagnoses, for every child classified as a clinical case, the data collection site coordinator completed a Clinical Diagnostic Information Form. Clinical cases were accepted only if: (a) a single primary diagnosis was indicated, (b) a qualified professional (e.g., psychiatrist, psychologist) had made the diagnosis, (c) the proper criteria were assessed using either the Diagnostic and Statistical Manual of Mental Disorders-Fourth Edition: Text Revision (DSM-IV-TR; APA, 2000) or the International Statistical Classification of Diseases and Health Related Problems 10th revision (ICD-10; WHO, 2004), and (d) appropriate methods (e.g., record review, rating scales, observation, interview) were used to make the diagnosis. These cases were used for the establishment of the validity of the BIMAS. Refer to *The BIMAS Standard as a Screening Tool* section in chapter 11, *Validity*, for a detailed description of the various clinical samples and their performance on the BIMAS.

Normative Sample Description

The BIMAS normative samples include 1,400 ratings from teachers on the BIMAS–Teacher (BIMAS–T), 1,400 ratings from parents on the BIMAS–Parent (BIMAS–P), and 700 ratings from adolescents on the BIMAS–Self-Report (BIMAS–SR).

[1] Neither the BIMAS Flex nor the BIMAS Standard Clinician versions are discussed in this chapter, as these versions are not norm-referenced. Since only the norm-referenced Standard form is discussed in this chapter, "BIMAS" is used to denote the BIMAS Standard throughout the chapter.

Teacher Normative Sample (BIMAS–T)

Teachers completed the BIMAS–T for a normative sample of 1,400 youth. All of the teachers had known the students they were rating for at least 1 month (specifically, the duration of teacher to student familiarity, in months, was: 1–3 = 7.5%, 4–6 = 23.6%, 7–11 = 44.4%, ≥ 12 = 23.7%; this data was missing for 0.7% of teachers), thereby meeting the minimum acquaintance requirement for completing the BIMAS.

The normative sample included ratings of 50 males and 50 females at each age (from 5 through 18 years). The sample characteristics were compared to the U.S. population (based on the 2000 U.S. Census report) on race/ethnicity and geographic region. The collected data were very similar to the U.S. Census in terms of race/ethnicity; however, some discrepancies existed between the actual collected data and Census targets for geographic region. To address these discrepancies, the sample was weighted through statistical procedures so that the weighted sample closely matched the U.S. Census statistics both in terms of race/ethnicity (see Table 9.1) and geographic region distribution (see Table 9.2).

Table 9.1. Race/Ethnicity Distribution: BIMAS–T Standard Normative Sample

Race/Ethnicity of the Rated Youth	Normative Sample (Weighted)		U.S. Census
	N	%	%
Asian	55	4.0	3.8
African American	217	16.0	15.7
Hispanic	203	14.9	15.1
White	836	61.4	61.9
Other	50	3.7	3.5
Total	1,361	100.0	100.0

Table 9.2. Geographic Region Distribution: BIMAS–T Standard Normative Sample

U.S. Region of the Rated Youth	Normative Sample (Weighted)		U.S. Census
	N	%	%
Northeast	251	18.4	18.1
Midwest	299	22.0	21. 9
West	325	23.9	23.3
South	486	35.7	36.7
Total	1,361	100.0	100.0

Parent Normative Sample (BIMAS–P)

The BIMAS–P rating form was completed for a normative sample of 1,400 children and adolescents. The majority (n = 1,116; 79.7%) of the BIMAS–P normative sample comprised assessments completed by the youth's biological mother, while the remaining assessments were completed by the youth's biological father (n = 164; 11.7%) or by other significant adults (including non-biological parents and other relatives; n = 120; 8.5%). The 1,400 rated youth included 50 males and 50 females at each age (for ages 5 through 18 years). The sample characteristics were compared to the U.S. population (based on the 2000 U.S. Census report) on race/ethnicity, PEL, and geographic region. While race/ethnicity very closely matched the Census targets, a similar statistical weighting procedure described in the Teacher normative sample section was applied to the BIMAS–P sample to correct for discrepancies in PEL and region. The resulting weighted sample therefore closely matched the U.S. Census statistics in terms of race/ethnicity (see Table 9.3), PEL (see Table 9.4), and geographic region distribution (see Table 9.5).

Table 9.3. Race/Ethnicity Distribution: BIMAS–P Standard Normative Sample

Race/Ethnicity of the Rated Youth	Normative Sample (Weighted)		U.S. Census
	N	%	%
Asian	30	2.2	3.8
African American	214	15.3	15.7
Hispanic	207	14.8	15.1
White	873	62.4	61.9
Other	75	5.4	3.5
Total	1,400	100.0	100.0

Table 9.4. Parental Education Level Distribution: BIMAS–P Standard Normative Sample

Parental Education Level of the Rated Youth	Normative Sample (Weighted)		U.S. Census
	N	%	%
High school diploma or lower	646	46.2	46.6
Apprenticeship/ Vocational training/ 2-year college/ some university	385	27.5	27.2
4-year college/ university or higher	369	26.4	26.2
Total	1,400	100.0	100.0

Table 9.5. Geographic Region Distribution: BIMAS–P Standard Normative Sample

U.S. Region of the Rated Youth	Normative Sample (Weighted)		U.S. Census
	N	%	%
Northeast	272	19.4	18.1
Midwest	265	18.9	21.9
West	333	23.8	23.3
South	530	37.9	36.7
Total	1,400	100.0	100.0

Self-Report Normative Sample (BIMAS–SR)

The BIMAS–SR normative sample consisted of 700 youth aged 12 to 18 years old (350 males, 350 females, 100 youth in each age group by year). Table 9.6 describes the sample's racial/ethnic distribution, which very closely approximated the U.S. Census. A similar weighting statistical procedure was applied to the BIMAS–SR sample so that the normative sample regional representation would be a close match to U.S. Census data (see Table 9.7).

Table 9.6. Race/Ethnicity Distribution: BIMAS–SR Standard Normative Sample

Race/Ethnicity	Normative Sample (Weighted)		U.S. Census
	N	%	%
Asian	28	4.0	3.8
African American	110	15.6	15.7
Hispanic	107	15.2	15.1
White	433	61.6	61.9
Other	25	3.5	3.5
Total	703	100.0	100.0

Table 9.7. Geographic Region Distribution: BIMAS–SR Standard Normative Sample

U.S. Region	Normative Sample (Weighted)		U.S. Census
	N	%	%
Northeast	128	18.3	18.1
Midwest	159	22.6	21.9
West	157	22.4	23.3
South	259	36.8	36.7
Total	703	100.0	100.0

Norming Procedures and Derivation of Standardized Scores

Following the gathering of teacher, parent, and self-report normative data, the raw score means, standard deviations, and frequency distribution statistics for each of the BIMAS scales were analyzed on the weighted normative samples. The follow-up analyses included an examination of the sample's performance across the youths' chronological age and gender[2]. Multivariate Analyses of Variance (MANOVAs) were employed to examine the relationships between gender and age with the BIMAS raw scale scores (see appendix G for results). Results indicated significant main effects for age with small to moderate effect sizes. On the teacher and parent reports, in general, scores on the Behavioral Concern scales decreased slightly from age 5 to age 12, and then increased again until age 18, while the reverse pattern was true for the Adaptive scales. On the self-report, scores on the Behavioral Concern scales tend to increase slightly from age 12 to age 15, and decreased again until age 18; the reverse pattern was observed on the Adaptive scales. Results also indicated some significant (though very small) gender effects. In general, where significant differences were found, boys were rated slightly higher than females on the Behavioral Concern scales, while females were rated slightly higher than males on the Adaptive scales. The main effects were qualified by significant Age × Gender interactions for the majority of the teacher and parent scales. Decomposition of these interactions generally indicated that the gender effects were significant only at certain ages. These results informed the construction of norms broken into six age groups for the BIMAS–T and BIMAS–P (i.e., ages 5–6, 7–9, 10–11, 12–13, 14–16, and 17–18), and three age groups for the BIMAS–SR (i.e., 12–13, 14–16, and 17–18).

The gender differences that were found are consistent with the differences that are reported in the literature (e.g., higher externalizing type of behaviors for males than females). To reflect these differences, the default setting in the BIMAS scoring and reporting software uses norms that are based on a sample comprising males and females (i.e., combined gender norms). Although the combined sample norms reflect the actual differences in the general population, some users might prefer to use gender-specific norms for certain settings and purposes. Therefore, gender-specific norms were also calculated and are available as a scoring option.

[2] Analyses were also conducted to examine the sample's performance across the youth's race/ethnicity. Results revealed that there were few meaningful differences between scores on the BIMAS across the races/ethnicities (see appendix F for details).

In the process of developing the BIMAS norms, a cumulative frequency distribution of raw scores was developed for each of the five BIMAS scales across the various age groups. As expected in the case of behavior rating scales, these distributions were skewed. Thorndike (1982) states that in general, "the tendency to psychopathology may not be normal in shape." Percentiles were computed for BIMAS raw scores to retain the actual shape of the original distribution. Data points that diverged significantly from a smooth curve partly reflect true differences and partly reflect sampling variability (Zachary & Gorsuch, 1985). To mitigate the effect of sampling variability, smoothed percentiles were also obtained using regression analysis. At each age, the predicted percentile score from the regression was used in conjunction with the original (unsmoothed) percentiles score to produce the final set of percentile scores. Specifically, the final "smoothed" percentiles were derived from the original, unsmoothed percentiles (given a 70% weighting) and a regression generated value (given a 30% weighting). Use of this smoothed normative value allows for irregular but real differences between age groups to have an effect, while reducing the impact of random fluctuation. The final smoothed percentile scores were then converted to standard *T*-scores with a mean of 50 and standard deviation of 10. For a discussion of non-linear *T*-score transformations, and smoothing procedures, the reader is referred to educational and psychological measurement textbooks (Urbina, 2004; Crocker & Algina, 1986).