The (mis)measurement of body image: ten strategies to improve assessment for applied and research purposes

J. Kevin Thompson∗

Department of Psychology, University of South Florida, 4202 East Fowler Avenue, Tampa, FL 33620-8200, USA

Received 7 July 2003; received in revised form 9 July 2003; accepted 10 July 2003

Abstract

Body image is a multifaceted construct consisting of a variety of measured dimensions. Research in the area has mushroomed in recent years, a phenomenon paralleled by an explosion in the assessments developed to measure some aspect of the construct. Unfortunately, these developments have not always been guided by a clear-cut attention to measurement issues. Assessment errors specific to the field of body image and also germane to basic psychological measurement strategies have characterized much of the work in the area. This article outlines 10 such issues with accompanying examples and suggestions for avoiding such pitfalls. Attention to these seemingly minor methodological details should result in a more consistent conceptualization of body image, utilization of appropriate measures in specific contexts, and engender more valid comparisons of findings across disparate studies.

© 2003 Elsevier B.V. All rights reserved.

Keywords: Body image; Assessment; Methodology; Psychometrics

The field of body image has experienced tremendous growth in the last 50 years, with a particular explosion of interest in the last two decades (Cash, this issue; Pruzinsky & Cash, 2002). As it often happens in science, such a phenomenon both stimulates and is fueled by the development of new measurement indices. In the case of body image, the exponential increase in the number of new and revised measures purporting to assess one or more multiple dimensions of body image has been well documented (Cash, 2002a; Gardner, 2002; Stewart & Williamson, 2004; Thompson, 1996; Thompson & Gardner, 2002; Thompson, Heinberg, Altabe, & Tantleff-Dunn, 1999; Thompson & van den Berg, 2002). In fact, currently there are over 50 measures for the researcher or clinician to consider when designing a scientific investigation or a clinical assessment protocol (for a review of recent advances in measurement, particularly computer-based approaches, see Stewart & Williamson, 2004).

Over my almost 20 years conducting research in the area of body image, I have experienced confusion and concern about the assessment considerations from a variety of vantage points. In reviewing manuscripts for journals, I have often been impressed with the theoretical rationale for a study, but underwhelmed by the attention of the author(s) to measurement issues. A laudable empirical question, sound methodological design, and sophisticated data analysis will not make up for a faulty selection of a measurement tool or misinterpretation of the construct indexed by a particular measure. On the other hand, I have often...
been on the receiving end of negative feedback from a reviewer regarding the reliability or validity of my chosen body image questionnaire or my use of a specific measure with an inappropriate sample. In this article, I detail some of the more common mistakes made by researchers and clinicians in the selection and utilization of measures of body image. The list is by no means exhaustive and is solely my own opinion. I hope that my commentary will generate feedback and debate, along with serving the intended purpose of directing greater attention and care to the issue of body image measurement. With the exception of Caveat 1 below, the list is in no particular order of importance.

1. Be specific in labeling the dimension of body image you wish to investigate and choose a measure that assesses this specific dimension.

This suggestion flows from the difficulty that any individual might have in this field upon first reflecting on the number and variety of measurement options. Thompson et al. (1999, p. 10) defined 14 different terms that have been used to refer to some aspect or dimension of body image, a number that has recently expanded (Stewart & Williamson, 2004). In turn, such concepts and measures have been categorized as reflecting a subjective satisfaction, affective, cognitive, behavioral, or perceptual component of body image (Stewart & Williamson, 2004; Thompson, 1996; Thompson et al., 1999; Thompson & van den Berg, 2002). Cash (2002b) has distinguished perceptual and attitudinal body image, and further distinguishes body image attitudes as comprised of evaluative–affective and cognitive–behavioral investment dimensions.

The basic mistake that occurs is that there is a mislabeling (or non-labeling) of the specific aspect of body image that the measure actually assesses. One source of this error, I believe, is that there is an assumption that most measures address some aspect of subjective satisfaction. Indeed many of the available measures, especially those focused on size or weight, address satisfaction with specific weight-relevant body sites (waist, hips, thighs) or overall size/weight satisfaction. When researchers and clinicians talk about “body image,” they are often thinking about body image evaluation or satisfaction.

However, it is essential to distinguish satisfaction with appearance from an investment or concern with appearance. For example, the Appearance Orientation Subscale of the Multidimensional Body-Self Relations Questionnaire (Brown, Cash, & Mikulka, 1990; Cash, 2000) taps into a cognitive–behavioral investment in one’s own appearance. The scale reflects whether the appearance is psychologically important to the individual and whether thought and behavior centers around appearance. Importantly, this dimension of body image is orthogonal (essentially uncorrelated) with satisfaction. Specifically, the Appearance Evaluation (AE) measure from the same multidimensional inventory (MBSRQ) is unrelated to the Appearance Orientation measure, a finding replicated in other laboratories (Thompson, Altabe, Johnson, & Stormer, 1994). The importance of the distinction between these two subscales and their ability to tap into different aspects of body image was underscored in a study evaluating the body image of African-American and Caucasian males and females (Smith, Thompson, Racyznski, & Hilner, 1999). Both Caucasian males and females had lower levels of satisfaction than their African-American counterparts on Appearance Evaluation, yet African-Americans had higher levels of appearance investment on the Appearance Orientation Scale. Cash, Melnyk, and Heabosky (in press) have further differentiated the body image investment construct, using their Appearance Schemas Inventory-Revised. Their evidence (including these ethnic group comparisons) suggests empirical differences between the “self-evaluative salience” of one’s appearance and its “motivational salience.”

Another example of an error related to the mislabeling of the dimension assessed by a particular measure is the use of general versus specific dissatisfaction measures. For instance, several measures, most notably the EDI-BD, index site-specific dissatisfaction, primarily weight-related sites. However, some measures of subjective dissatisfaction assess a more complex or general dimension of appearance (i.e., an evaluation of general or global appearance; for instance the MBSRQ–AE and several of the “body-esteem” measures mentioned in previous reviews) (Thompson et al., 1999). One of the problems that flows from using only one of these two scales
is an interpretive one. Any analysis of results that does not confine the discussion to the specific type of measure may produce conclusions that are erroneous. For instance, Walden, Thompson, and Wells (1997) evaluated site-specific (breast size) satisfaction, along with several other measures of body image in a study of women before and after explantation of silicone breast implants who were compared with a group of cholecystectomy patients and a non-patient control group. There was a drastic diminution in the satisfaction ratings of breast size, as expected, however, levels of site-specific anxiety for other body sites did not change. Interestingly, overall global evaluation of appearance (as assessed by the MBSRQ–AE) did show reductions, paralleling those of the breast satisfaction. These findings indicate that the negative body image effects of explantation affected the more conceptual evaluation of attractiveness, but not individually rated body site anxiety levels (see Sarwer & Cerrand, this issue).

A third example involves the labeling of a measure reflective of some aspect of body image when the measure actually captures a closely related, but distinct construct. Much of my work over the years with various colleagues has focused on the construct of internalization of societal standards of attractiveness (Heinberg, Thompson, & Stormer, 1995; Keery, Shroff, Thompson, Wertheim, & Smolak, in press; Thompson, van den Berg, Roehrig, Guarda, & Heinberg, in press). This variable has been shown to prospectively predict the onset of body dissatisfaction and is also responsive to treatment programs designed to reduce the magnitude of internalization (e.g., Thompson & Stice, 2001). Yet, the measure is often mislabeled as an indicator of body image, when it should be viewed as a correlate, symptom, or risk factor of some dimension of body disturbance. Much of the research has evaluated dissatisfaction, but conceivably internalization may predict other dimensions of body image, such as appearance schematicity.

Although I may appear to be picking nits about issues only important to methodological elitists, the issues have practical and theoretical significance. In terms of treatment, clearly it is important to understand which, if any, aspect of body image is responsive to interventions because some components may not change (e.g., body image investment or schematicity), whereas, other components (weight-specific dissatisfaction) may improve. In terms of theory, explanatory models often hypothesize directional or reciprocal relations among specific dimensions of body image, such as concern with appearance, level of schematicity, and negative emotions (Cash, 2002b; Williamson, Stewart, White, & York-Crowe, 2002). Clearly, models cannot be tested if boundaries between dimensions of body image are not articulated and adhered to in measurement selection.

2. Use multiple measures of body image.

As noted above, one problem with a disparity between the body image construct you have selected and the dimension you are actually measuring, is that misleading interpretations are made, adding to the confusion in the literature. One strategy for dealing with this problem is to include multiple measures of body image, taking care to have a broad range of pertinent dimensions represented. Such a plan is especially indicated for exploratory studies, whether evaluating groups for base rates (i.e., cross-cultural work and gender comparisons) or gauging the specific domain (dimension) effects of a treatment. Such a strategy of using multiple measures will ensure that misleading interpretations regarding outcome are minimized, which might occur if a single scale that captures a very specific dimension of body image was utilized. I am not encouraging the strategy of throwing “everything body image” at the wall to see what sticks. I am recommending that one should thoughtfully consider the inclusion of multiple but pertinent facets of body image.

For example, the full-version of the MBSRQ contains subscales regarding health and fitness body image dimensions. Just because they are there does not make them relevant to the study and analyses of results. One issue that frequently arises when multiple measures are used is whether the overlap (multicollinearity) is too high to permit certain analyses (regressions) or interpretation. In fact, often the degree of overlap is not known for some measures, so the investigator has to make decisions based on the specific findings of his or her study. One guideline is to evaluate the shared variance between or among measures (i.e., the square of the simple Pearson correlation coefficient). A guideline that I have used is that if $R^2 \geq 0.50$, there is substantial overlap between the measures, and
3. Select measures with established reliability and validity.

Do not “make up” measures if you do not have to. As noted at the outset of this article, there are a veritable plethora of extant measures that have been published. Of course, this does not necessarily mean that all of them are reliable and valid, but it does mean that there is an excellent starting point for reviewing assessments with an eye toward selection of one with established psychometrics. Reliability might include either test–retest or internal consistency (ideally, both although internal consistency is not relevant for single-item measures) and estimates should reach an approximate minimum of 0.70. Validity is more complicated, but essentially there should be some evidence that the measure (a) concurrently correlates at a reasonable level with measures that it should theoretically correlate with, such as other relevant measures of body image, eating disturbance, or self-esteem, (b) prospectively predicts these measures, (c) doesn’t correlate with measures it should not (e.g., socially desirable responding), and/or (d) discriminates between groups that it should (e.g., specific clinical groups versus controls). Furthermore, if you wish to assess changes in the construct due to an experimental manipulation or a treatment or prevention program, there should be evidence of the measure’s responsiveness.

All too often, researchers trumpet findings using a new measure without any accompanying psychometric evaluation. Often, the measure has received extensive development work and has been used in several published studies, yet psychometric data, once collected, indicate an unacceptably low reliability level (Cafri, Roehrig, & Thompson, in press; Cafri & Thompson, in press).

4. Take care to use your measure with appropriate samples.

This suggestion follows from Caveat 3 above and expands on the importance of psychometric evaluation. Not only should the measure have received analysis suggesting that it is reliable and valid in some sample, it is important that the measure have been tested on a sample similar to the one you plan to study. To take an extreme example, just because a measure was found to have good psychometrics on a sample of Caucasian college females does not mean it is appropriate to use the measure with a community sample of obese African-American men. Ideally, you would select a measure that has received extensive background validation with many and diverse samples (for instance, the MBSRQ). In reality, the measure you want to use may have received limited validation work. This does not necessarily mean that you should immediately discount this measure, just be sure that you have considered the available options (Gardner, 2002; Thompson & Gardiner, 2002; Thompson & van den Berg, 2002). Cash (personal communication) often receives inquiries from persons who want to use his adult and late-adolescent body image assessments with children. His answer is “no, consider the available options.”

5. Assess reliability and validity of the measure in your sample.

Streiner (2003), in his discussion of the four myths of internal consistency, noted that the belief that alpha was a “fixed property of a scale” was erroneous and advocated “that it is not sufficient to rely on published reports of reliability if the scale is to be used with another group of people” (p. 101). Internal consistency estimates merely require that you enter individual item response values and compute reliabilities via the statistical package you are using. Test–retest estimates do not have to be based on samples in hundreds; a sample of 30–50 usually provides a sufficient sample size for evaluating test–retest reliability. Validity is more involved, and may involve the inclusion of some body image or relevant psychological functioning measures (e.g., self-esteem, depression, eating disturbance, etc.) that you did not plan to include in your study. Most of these can be assessed with available brief scales (e.g., the 10-item Rosenberg Self-Esteem Scale; Rosenberg, 1965). The benefits far outweigh the costs in terms of what you can say (and publish!) when you assess reliability and validity in your sample. As discussed in
6. Thoughtfully adapt measures for your own purposes.

You have thoroughly searched for the right assessment and you have come up short. What about adapting or modifying an existing scale? My bottom line is that if a slight change in the wording of a scale or the inclusion of additional items is required by the research question, then it is bad science not to tinker with existing scales. Many individuals feel that once a questionnaire is published, it must be used in the way it was originally constructed, otherwise comparisons to previous data are suspect. It is true that such comparisons will be limited, which is a concern (see Caveat 10). However, having the right scale to test the research question or to document the patient’s improvement on some focal aspect of body image is more important than normative comparisons. If such an alteration is done, it might be advisable to include a second or third scale (see Caveat 2) to provide for some normative comparisons.

Is it permissible to modify a scale published by others? It is a good idea to ask for permission to use an author’s measure, even if you propose to use it exactly as it was developed. When the question is one of modifying the measure, it is not acceptable to simply reference the original paper and instrument. Instead, contact the author, describe the need for modification, and ask permission. For example, I have been asked on many occasions for permission to modify our teasing scale (e.g., Thompson, Cattarin, Fowler, & Fisher, 1995) and I have always given permission because our teasing items focused on a specific form of teasing (being overweight) from a generic source (people). Clearly, if someone were interested in boys who were teased about being underweight from gym teachers, it would be impossible to conduct the study without modifying our measure (see Boroughs & Thompson, 2002; Lundgren, Anderson, & Thompson, in press; Sands & Wardle, 2003; Smolak, Levine, & Thompson, 2001, for examples of measure adaptations) The same advice regarding permissions applies if one wishes to translate an instrument from one language to another.

7. Determine if you need a state or trait measure.

The great majority of body image measures assess a trait dimension. However, there are measures available that index a more immediate, state-like variable (Cash, 2002a). The distinction between the two is very important, yet a common mistake is to use an unmodified trait measure to determine the short-term effects of an experimental manipulation (e.g., viewing fashion magazines). In 1990, we developed parallel measures of body image anxiety that assessed both a state and trait component (Reed, Thompson, Brannick, & Sacco, 1990), and over the years have used Visual Analogue Scales to determine the short-term, experimental effects of media exposure on body image (Cattarin, Thompson, Thomas, Williams, 2000; Heinberg & Thompson, 1995; Tantleff-Dunn & Thompson, 1998). Recently, Cash, Fleming, Alindogan, Steadman, and Whitehead (2002) developed the Body Image States Scale which taps six facets of state body image experiences.

State measures should clearly assess the level of the variable at a specific point in time. Importantly, the instruction should be for the individual to rate the items to capture how they “feel at this moment” or “right now.” I particularly like the specificity of the Body Image States Scale, which asks participants to rate “how you feel right now at this very moment” (Cash et al., 2002). There may be instances wherein no state measure is available for the dimension of body image under investigation. In such a case, carefully adapting the trait measure by altering the instructional protocol might be indicated.

State measures are ideal for use in a clinical setting. For instance, the individual might self-monitor or self-record daily levels of body dissatisfaction, along with a description of environmental and contextual influences. The Body Image Diary (Cash, 1997; Cash & Hrabosky, 2004) is an excellent tool for this purpose, providing an ongoing method for evaluating the state level of body image experiences in a range of selected contexts and providing information regarding influences that can provide essential information for ongoing clinical work. State measures might also be quite useful for immediate pre-post testing of the effects of a particular clinical intervention, such as mirror desensitization or a cognitive restructuring treatment session.
8. Use care in determining the instructional protocol.

The instructional protocol is important for determining if a state or trait dimension of body image is assessed, as noted in Caveat 7. It is also important to consider that vague instructions may yield imprecise information regarding the exact magnitude or level of the body image variable under investigation. For instance, research indicates a significant effect on body size ratings when an affective versus a cognitive instructional protocol is utilized (Thompson, 1991; Thompson & Dolce, 1989). Specifically, affective instructions asked respondents to respond based on how they felt, and cognitive instructions asked them to respond based on a rational or intellectual view of their body. In both studies, affective instructions produced ratings indicative of a greater degree of body image disturbance. Thus, how you ask the question matters. These results further raise the interesting question of whether the distinction between one’s affect and cognitive view of appearance may be related to certain specific risk factors, be a risk factor itself, or have clinical implications in terms of therapeutic strategy.

The above example illustrates only one specific instance of instructional effects, but the general principle that flows from the example is that importance should be placed on the instructions and this information should be contained in written reports of research. Additionally, from a clinical perspective, if multiple assessments are being conducted to gauge the treatment’s effectiveness, the consistency of instructional format is imperative.

9. Consider diversity and analyze data by pertinent participant characteristics.

Caveats 9 and 10 address issues that deal with data analysis and interpretation as opposed to assessment selection and utilization. Often, in research articles, data are pooled for analyses across individuals of diverse characteristics and interpretation of findings are based on such decisions. Perhaps the most egregious example over the years was the decision to pool data from males and females. Now, we know that the muscularity versus thinness dimension of body image is an essential gender issue to consider in analyses (Caffè, Strauss, & Thompson, 2002; McCabe & Ricciardelli, 2004). Not only should data be evaluated separately, care should also be taken when using the same measures for both genders (see Caveat 4). Another example of premature pooling is compositing scores from individuals of different ethnic backgrounds, ages, and weight classifications. I am not saying that there is no basis for combining data, only that analyzing first by participant characteristics may yield important findings regarding base rate differences or even associations among body image measures and other variables. If these preliminary analyses suggest relative equivalence across dispositional characteristics, then collapsed analyses are indicated. Of course, this presumes that these subgroup comparisons can be made with sufficient statistical power. Also, transforming raw scores from each scale to Z-scores prior to compositing is always indicated.

10. Evaluate your data in terms of norms and clinical versus statistical significance.

An examination of a single patient’s score in terms of norms for a given measure is an accepted clinical practice. Actuarial data are available for many widely used psychological tests, and lack of reference to such material when interpreting clinical assessment protocols is virtually unthinkable. Certainly, if a practitioner is working with a patient who has a body image issue, the use of a well-standardized and normed instrument will lend itself to science-based clinical practice.

Unfortunately, many researchers do not avail themselves of normative data when interpreting their findings. Mean levels are reported, analyses between groups or relationships among variables are offered, yet the basic data are not viewed in light of extant levels in other samples. Such comparisons could be very informative in terms of how the current levels compare to other data, especially if the samples are otherwise quite different. In effect, if the samples are quite similar in characteristics, a significant variation in the findings may need to be explored and explained. For instance, if the new data indicate a very high or low level on the variable, when compared to the normative data, there may be something specific about the sample or data collection method that needs to be explored to explain the findings. An examination
of current levels of prior data can reveal something as mundane but critical as a scoring error (i.e., falling to reverse-key appropriate items, which internal consistency analyses will also detect). On the other hand, replicability of levels may be used to buttress the soundness of the methods of the study and add to the database of normative information for the specific instrument.

In treatment outcome research, clinical versus statistical significance is extremely important. An evaluation of results only in terms of statistically significant changes on the dependent variables does not address whether the magnitude of the changes is reflective of a meaningful improvement. One way of gauging clinically relevant improvement is by comparing the post-treatment scores to the normative levels (averages) of a non-patient sample. If scores move from the elevated range in comparison to the normative group to a level comparable to the non-patient group, a meaningful change in symptom level can be said to occur (see Jacobson & Truax, 1991, for a detailed discussion of the determination of the clinical significance of changes.)

Conclusions

Selection of body image measures for use in clinical and research settings is a complicated endeavor because of the multidimensionality of the construct and the availability of an extraordinary number of assessment tools that purport to index one or more of these dimensions. Care should be taken in the selection of the measure for the intended purpose. The key question is this: What dimension of body image fits your research or clinical need? In many instances, the dimensions and measures are appropriately multiple. Finally, it is imperative to attend to the reliability and validity of the chosen measures with specific samples, with attention to the diversity of these samples.

We may reach a time when specific practice guidelines for measurement selection in the field are determined (Cash & Pruzinsky, 2002); in the meantime, the thoughtful selection or development of the right measures for the right questions and populations is crucial. As scientists and clinicians, the conclusions we reach are only as good as the measurements from which we derive our results.

References


