

# The "Goodness" of Attachment Assessment: There Is A "Gold Standard" But It Isn't As Simple As That

**Everett Waters**

*SUNY Stony Brook*

Attachment research depends on good assessment. There are lots of senses in which a measure can be considered "good". They are all important.

**Reliability** (Are the scores it yields representative of the subject's typical behavior? Are they reproducible?") is essential. If a behavior sample is too brief, the target behavior is too rare, or the behavior scored too subject to situational factors, you can expect to obtain scores that vary widely around a subject's typical ("true") behavior.

Reliability is also affected by the range of individual differences in your sample.

Low reliability reduces correlation coefficients and diminishes statistical power in group comparisons. In addition, if your study involves several measures and some are more reliable than others, the patterns of results can be seriously distorted.

For example, suppose the correlations among variables A, B, & C are *in fact* equal. If your measures of A, B, & C are not equally reliable, the data will suggest that the more reliably measured variables are more highly correlated. This is a fact about your measures not about the world. Differential reliability can also reverse patterns of results. It can also produce very misleading results in multiple re-

gression, factor analysis, and causal modeling. It typically receives far less attention than it deserves.

Note that reliability is not the same as rater agreement. Raters can agree 100% on subjects' behavior and yet the behavior be too small a sample to be a representative indication of how individual subjects differ from one another.

Reliability does not guarantee either stability or validity but without it you won't be able to detect either one even if it is there. The methods of test construction and reliability assessment developed for use in IQ and personality trait assessment are important for attachment assessment as well - be it observational, interview, laboratory tasks, or self report.

Fortunately, reliability is not a property of the construct being measured. There are no unreliable traits or behaviors - only unreliable measures. The reliability of any measurement can always be increased to any level required by increasing the amount of observation, aggregating multiple observations.

**Validity** ("Does it measure what it is supposed to measure? And not what it isn't supposed to?") can only be defined in relation to a theoretical construct. Does the measure act

like the theory says it should. If not, the problem could lie in the measure, the theory, or both. The more clearly (and unavoidably) a theory predicts particular results, the easier it is to evaluate validity. No theory is perfect from the start. What you are looking for is enough of a theory to guide developing a "good enough measure"; this is used to obtain empirical results that can help refine the theory; the refined theory leads to better measurement, and so on. Paul Meehl called this process "bootstrapping" - pulling yourself up by your own bootstraps.

### **The validity of attachment measures: How would you know?**

Strange Situation classifications can be quite stable and have a wide range of correlates in early care and later competence and adjustment. But stability and wide range of correlates in later competence and adjustment are not sufficient to prove that a procedure is measuring attachment security. Even correlations with maternal care is not definitive. No theory predicts that maternal care affects only attachment security. And no theory predicts that only attachment security influences later competence and adjustment. So there are always alternative interpretations of measures that are stable, related to early care, and to later competence and adjustment. But only Bowlby's theory links them to secure base behavior at home. This is why we consider the link to home behavior to be the "gold standard" against which any measure of infant attachment security should be tested. It is the only way to know whether the Strange Situation is valid in older age groups, infants who experienced significant amounts of day care, in other cultures, etc. This should always be established before interpreting the Strange Situation in such samples. In many cases it has not been.

Of course it might be possible to show that the "attachment security" construct is broader than just secure base behavior. In all likelihood both Bowlby and Ainsworth thought so. But Bowlby strategically chose to tie it closely to the secure base phenomenon because doing so

allowed him to develop his control system motivational model and escape Freud's scientifically indefensible (and largely discredited) drive reduction motivation model. See Waters & Cummings (2000) (On-line articles section of this site) for an extended discussion of the central role of the secure base concept in attachment theory.

The same line of reasoning can be applied to the problem of validating the Adult Attachment Interview. Lots of theoretical frameworks, including theories of general adjustment, anxiety, and stress and coping, might predict that such an interview would be related to marriage, parenting, and adjustment. But only Bowlby's theory would predict that it is related also to the components of marriage that we call secure base use and secure base support. See Waters, E., Merrick, S., Treboux, D., Crowell, J., & Albersheim, L. (2000). *Child Development*, 71, 684-689 for evidence that AAI classifications are related to ones ability to use mother as a secure base in infancy. See also Crowell, Treboux, Gao, Fyffe, Pan, & Waters (*Dev. Psych*, 2002, 38, 679-693) for evidence that the AAI is also related to the ability to use (and serve as ) a secure base in marriage. These are important evidence linking adult attachment representations assessed via the AAI to the Bowlby-Ainsworth construct.

### **Using valid measures in "dangerous" tests**

The process of validating attachment measures is necessarily closely intertwined with the process of validating key postulates of attachment theory. Nonetheless, test validation hardly exhausts the possibilities for using a measure to evaluate and extend a theory. In fact, some of the most interesting and important work only begins once we know we can trust our measures.

It makes sense to tolerate a certain openness to any theory, especially early on. Theorists often need a taste of empirical data before they begin to see the most productive lines of analysis. But if the theory doesn't soon be-

come specific enough that formulate "dangerous" empirical tests, (i.e., predictions that better be true or the theory has something wrong with it), if it never becomes more than the theory that "all good things go together", then you have to wonder whether the theory really says anything.

The prediction that attachment security is related to specific aspects of maternal care, can be stable during infancy, changes if patterns of care change, and is related to patterns of attachment in adulthood are examples of such tests. So are the prediction that adult attachment representations are related to the ability to use and provide secure base support in adult relationships.

There simply isn't any way Bowlby's attachment theory could accommodate failures in such tests without needing major revisions - perhaps revisions that would render it a very different theory. The majority of empirical studies in attachment research, useful though they may be as descriptive information, are not "dangerous" in this sense. Many are so loosely connected with the theory that you can tell a story about almost any outcome, positive or negative. A few are so predictably, necessarily, trivially true that there is no risk in them at all - an correspondingly not much information.

Nothing specific to attachment theory here. This is an issue in any area of science. It's just that some commentators have suggested that it is unusually common in the behavioral sciences. See for example the classic, pointed, and often amusing analyses in C. Wright Mills' "The sociological imagination", Paul Meehl's "Why I don't attend case conferences", and Jan Smedslund's "What kind of propositions are set forth in developmental research?" The bottom line in these often amusing critiques is not that you can't do anything important in the social and behavioral sciences - only that "true" and "important" are not the same thing. Most empirical studies are merely true and soon forgotten. Well formulated and skillfully conducted, dangerous tests are always important.

Lots of theories afford predictions about competence, relationships, and emotion. The secure base construct is one of the most distinctive features of Bowlby's attachment theory. If you want to formulate dangerous tests, this is a good place to start. But from a measurement perspective it is a difficult construct to work with. It is played out over time and space and can't be equated with the frequency or intensity of discrete behaviors.

To paraphrase Albert Einstein:

Not everything that can be measured is important.

Not everything that is important is (easily) measured.

E.W.  
12/02

APA reference format for material retrieved from the Internet:

Waters, E. (2002). The "Goodness" of Attachment Assessment: There Is A "Gold Standard" But It Isn't As Simple As That. Retrieved (*current date*) from [http://www.psychology.sunysb.edu/attachment/measures/content/attachment\\_validity.html](http://www.psychology.sunysb.edu/attachment/measures/content/attachment_validity.html)