

Refining the Architecture of Aggression: A Measurement Model for the Buss–Perry Aggression Questionnaire

Fred B. Bryant and Bruce D. Smith

Loyola University Chicago

Among the most popular measures of aggression is the 29-item, self-report Aggression Questionnaire (AQ; Buss & Perry, 1992; Buss & Warren, 2000). Structural analyses of the AQ have revealed four underlying factors: Physical Aggression, Verbal Aggression, Anger, and Hostility. However, these four factors explain too little common variance (i.e., about 80%) to be an adequate measurement model. In the present study, we used confirmatory factor analysis with a total sample of 1154 respondents to compare four alternative measurement models for the AQ that are currently in use. Replicating earlier work, none of these models fit the data well, and the original four-factor model achieved only mediocre goodness-of-fit in three independent samples ($GFI = .76 - .81$). To develop a more appropriate measurement model, we omitted items with low loadings or multiple loadings based on principal components analysis and excluded items with reverse-scored wording. This yielded a 12-item, four-factor measurement model with acceptable goodness-of-fit ($GFI = .94$). Secondary analysis of two independent data sets confirmed the refined model's generalizability for British (Archer, Holloway, & McLoughlin, 1995; $GFI = .93$) and Canadian (Harris, 1995; $GFI = .94$) samples. The refined model yielded equivalent factor structures for males and females in all three samples. We also replicated the refined four-factor model in two additional American samples, who completed a new short form of the AQ containing only the subset of 12 items in random order. Additional analyses provided evidence supporting the model's construct validity and demonstrated stronger discriminant validity for the refined Hostility factor compared to its predecessor. The new short form of the AQ thus not only contains fewer than half as many items as the original, but also is psychometrically superior. © 2001 Academic Press

The authors thank John Archer and Julie Harris for graciously providing us with data from their published articles. We also gratefully acknowledge the helpful advice of Mary Harris, the invaluable research assistance of Rebecca Guilbault, and the insightful editorial feedback of Craig Colder and an anonymous reviewer. Earlier versions of this article were presented at the American Psychological Association convention, Chicago, IL, August 1997, and at the Joint Meeting of the Classification Society of North America and the Psychometric Society, Urbana, IL, June 1998.

Address correspondence and reprint requests to Fred B. Bryant, Department of Psychology, Loyola University Chicago, 6525 North Sheridan Road, Chicago, IL 60626. E-mail: fbryant@luc.edu.

Much work has emphasized the role of physical aggression, verbal aggression, anger, and hostility as subtraits in a global conceptualization of aggression (Buss, 1961; Buss & Durkee, 1957; Buss & Perry, 1992; Harris, 1995; Zillmann, 1979). Early measurement of aggression used an experimental methodology, required a laboratory, and suffered difficulties in interpreting aggressive intent (Zillmann, 1979). To reduce the time, effort, and resources involved in measuring aggression, Buss and Durkee (1957) developed a 75-item self-report instrument, the Buss–Durkee Hostility Inventory, which aggression researchers have often used.

To improve the psychometric properties of this instrument, Buss and Perry (1992) more recently developed a 29-item self-report questionnaire, the Aggression Questionnaire (AQ; Buss & Perry, 1992; Buss & Warren, 2000). They designed the AQ to measure four dispositional subtraits of aggression, which they defined as follows: “Physical and verbal aggression, which involve hurting or harming others, represent the instrumental or motor component of behavior. Anger, which involves physiological arousal and preparation for aggression, represents the emotional or affective component of behavior. Hostility, which consists of feelings of ill will and injustice, represents the cognitive component of behavior” (Buss & Perry, 1992, p. 457).

In constructing this questionnaire, Buss and Perry (1992) borrowed some items intact from the earlier Hostility Inventory, revised other Buss–Durkee items to improve clarity, and added many new items to generate an initial pool of 52 questions. They then administered this set of 52 questions to three successive samples of 406, 448, and 399 college students and analyzed the structure of responses using exploratory principal components analysis with oblique rotations. Although they had originally generated items for six *a priori* components of aggression (Physical Aggression, Verbal Aggression, Anger, Indirect Aggression, Resentment, and Suspicion), only four correlated factors emerged—Physical Aggression, Verbal Aggression, Anger, and Hostility—on which a core set of 29 items loaded, and this four-factor structure appeared to replicate across all three samples. Buss and Perry (1992) next used confirmatory factor analysis to evaluate the goodness-of-fit of three alternative measurement models for the set of 29 items: (a) a global one-factor model that assumes all items reflect a single general aggression factor; (b) a four-factor model that represents the structure found in the principal components analysis; and (c) a hierarchical factor model that assumes the four correlated, first-order factors reflect a single, second-order “super factor” of aggression. Figure 1 presents a graphical representation of each of these three measurement models.

As a sole measure of each model’s goodness-of-fit, Buss and Perry (1992) computed the ratio of chi-square to degrees of freedom (cf. Hoelter, 1983). Based on the notion that ratios under 2 reflect acceptable fit, Buss and Perry

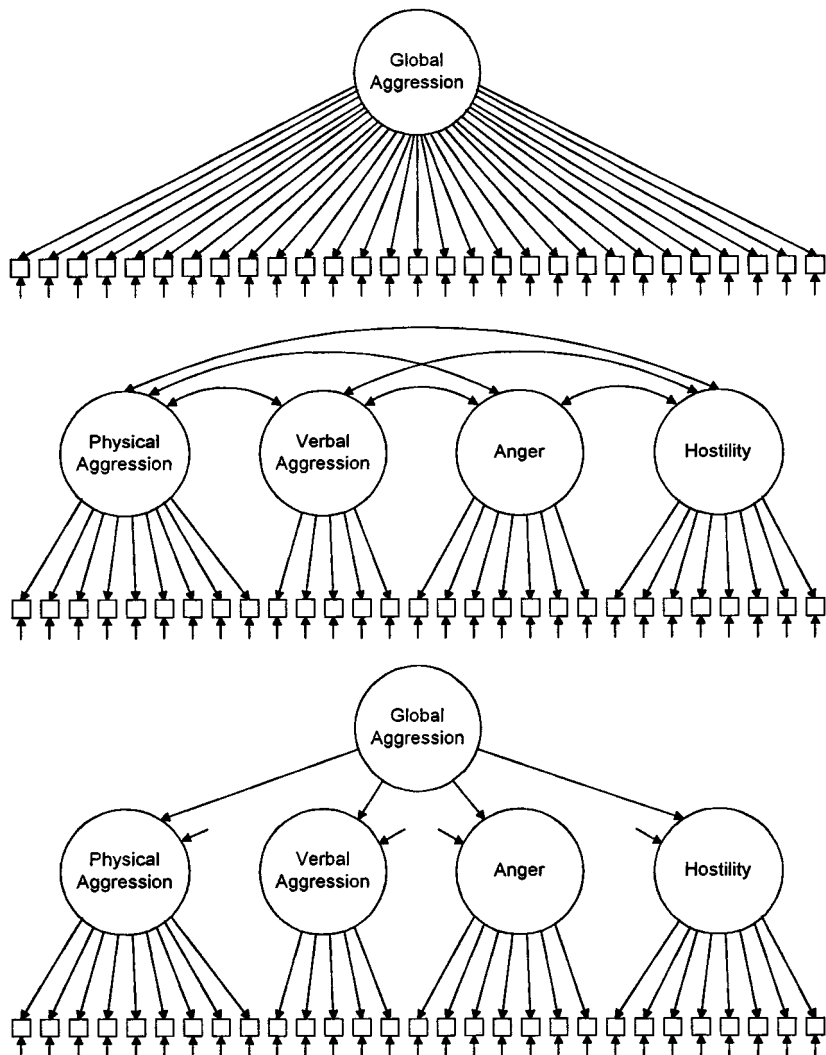


FIG. 1 Three measurement models currently in use for the 29-item Aggression Questionnaire (AQ; Buss & Perry, 1992). Squares represent measured variables (or AQ items), and circles represent latent constructs (or AQ factors). Arrow-headed straight lines connecting latent constructs to measured variables represent item factor-loadings (λ_s). Two-headed, curved lines connecting latent factors represent factor interrelationships (ϕ_s). The small, arrow-headed straight line to each measured variable represents unique variance associated with measurement error, or the joint effect of unmeasured influences and random error (θ_s). The *unidimensional* model (top) assumes that a single, first-order factor (Global Aggression) explains the covariation among the 29 AQ items. The *multidimensional* model (center) assumes that four, interrelated first-order factors (Physical Aggression, Verbal Aggression, Anger, and Hostility) explain the covariation among the 29 AQ items. The *hierarchical* model (bottom) assumes that a single, global second-order factor (Global Aggression) underlies the covariation among the four first-order factors. The small, arrow-headed straight line to each first-order latent factor in the hierarchical model represents specific variance that is unrelated to the second-order latent factor (ψ).

(1992) concluded that both the four-factor model ($\chi^2/df = 1.94$) and the higher-order factor model ($\chi^2/df = 1.95$) fit the pooled data reasonably well, whereas the one-factor model ($\chi^2/df = 2.27$) did not. Recalculating the model chi-squares from the known degrees of freedom reveals that both the four-factor model, $\chi^2(371) = 719.7$, and the higher-order model, $\chi^2(373) = 727.4$, fit the data significantly better (both $\Delta\chi^2$ s > 128.3 , $ps < .00001$) than did the one-factor model, $\chi^2(377) = 855.8$. Because Buss and Perry (1992) reported no other goodness-of-fit measures, however, we know nothing about the proportion of variance–covariance information that these models explained in their data or how much better these models fit relative to a worse-case “null” model that assumes there are no common factors. Such measures of absolute and relative fit would be useful in deciding whether either of the multidimensional frameworks represents an acceptable measurement model for the AQ.

Although Buss and Perry (1992) proposed their four-factor solution as a measurement model for the AQ, more recent analyses (Archer, Kilpatrick, & Bramwell, 1995; Harris, 1995; Williams, Boyd, Cascardi, & Poythress, 1996) suggest that this structure explains too little common variance among the 29 items (i.e., about 80%) to serve as a measurement model. Accordingly, the primary goal of the present study was to develop an acceptable measurement model for the AQ and to assess its construct validity.

As a means of improving the measurement precision of the AQ, previous researchers have proposed discarding AQ items that are relatively unreliable indicators of Hostility. Indeed, omitting Buss and Perry's (1992, Table 1, p. 454) sixth and eighth indicators of Hostility has been found to increase the reliability of the Hostility factor in both Canadian (Harris, 1995) and Dutch (Meesters, Muris, Bosma, Schouten, & Beuving, 1996) samples. Yet, this approach is not without its critics (Bernstein & Gesn, 1997). The key theoretical issue here is whether it is better to have a global, somewhat heterogeneous construct of known theoretical utility or to have a more specific and psychometrically purified construct (cf. Bryant, Yarnold, & Grimm, 1996).

Researchers using the AQ have typically adopted one of two dominant strategies for scoring the instrument. The first strategy assumes that all AQ items reflect a single underlying construct reflecting a person's global predisposition toward aggression. With this unidimensional approach, researchers simply sum responses to the 29 items to construct an AQ total score (e.g., Buss & Perry, 1992). The second strategy assumes that aggression consists of four correlated dimensions reflecting a person's predisposition toward aggression in the physical, verbal, emotional (Anger), and cognitive (Hostility) domains. With this multidimensional approach, researchers construct four separate subscale scores by summing or averaging responses to the set of AQ items tapping each domain of aggression (e.g., Felsten & Hill, 1998).

Yet, neither of these two approaches adequately captures the variation in people's responses to the AQ. More specifically, measurement error and other constructs in addition to aggression have an unacceptably large influence on responses to the set of 29 AQ items. Although the theoretical model underlying the AQ is conceptually well grounded in the aggression literature, researchers need a better operational framework for quantifying responses to the instrument. With this aim in mind, we sought to improve the correspondence between the conceptual and operational definitions underlying the AQ and to develop a more reasonable measurement model for the AQ.

We worked with five independent data sets: three primary data sets that we collected for this study and two archival data sets that others had collected earlier. First, we collected a new data set to evaluate and compare the explanatory power of four different measurement models for the AQ that are currently in use in the literature. We also used these new data to develop a better-fitting refined version of Buss and Perry's (1992) four-factor model and to compare the convergent and discriminant validity of this refined model with that of the original. We then obtained two preexisting AQ data sets—a British sample (Archer, Holloway, & McLoughlin, 1995) and a Canadian sample (Harris, 1995)—with which we assessed the cross-sample generalizability of both the original and refined models. Finally, we collected new data from two additional American samples to evaluate the refined model's replicability using a new "short form" of the AQ, containing only the subset of 12 items in random order.

METHOD

Participants and Procedure

Sample 1. The first sample consisted of new data from 307 American undergraduates (173 females, 131 males, and 3 who did not report gender) at a private metropolitan university who voluntarily participated. Average age was 18.94 ($SD = 1.21$). Respondents completed a battery of tests, including the 29-item Aggression Questionnaire (AQ; Buss & Perry, 1992; Buss & Warren, 2000).

Sample 2. The second sample, which we used for cross-validation, was originally collected by Archer, Holloway, and McLoughlin (1995). The sample consisted of 200 British undergraduates (100 females and 100 males). Average age was 25.13 ($SD = 6.17$). Participants completed the 29-item AQ using the same 5-point scale. The data consisted of the raw AQ data analyzed by Archer et al. (1995).

Sample 3. The third sample, also used for cross-validation, was originally collected by Harris (1995). The sample consisted of 306 Canadian undergraduates (151 female and 155 male). Average age was comparable to U.S. college samples (cf. Harris, 1995). Participants completed the 29-item AQ using the same 5-point scale. The data consisted of the covariance matrix analyzed by Harris (1995).

Sample 4. The fourth sample was used to assess the replicability of the refined four-factor model using a new "short form" of the AQ. The sample consisted of 171 American undergraduates (123 females and 48 males) at a private metropolitan university who voluntarily participated. Average age was 18.35 ($SD = 0.84$). Respondents completed a shortened version of the AQ containing only the 12 items comprising the refined four-factor measurement model.

Sample 5. The fifth sample was also used to assess the replicability of the refined four-factor model using the new "short form" of the AQ. The sample consisted of 170 American undergraduates (124 females and 46 males) at a private metropolitan university who voluntarily participated. Average age was 18.64 ($SD = 1.57$). Respondents completed the shortened 12-item version of the AQ.

Aggression Questionnaire

Original AQ. Respondents in Samples 1–3 completed a battery of tests, including the 29-item Aggression Questionnaire (Buss & Perry, 1992; Buss & Warren, 2000). Participants rated how well each AQ item described themselves using the original 5-point scale, ranging from *extremely uncharacteristic of me* (1) to *extremely characteristic of me* (5), as defined by the original instrument. Following Buss and Perry's (1992, p. 453) instructions, each sample received a different random ordering of the 29 AQ items.

Short form of the AQ. Respondents in Samples 4 and 5 completed a new, shortened version of the AQ containing only the 12 items comprising the refined four-factor measurement model. Using Buss and Perry's (1992, p. 454) Table 1, the randomized order of the items was 11, 23, 8, 25, 21, 14, 15, 2, 13, 24, 6, and 20. Participants rated each AQ item using a 6-point scale, ranging from *extremely uncharacteristic of me* (1) to *extremely characteristic of me* (6). Changing from the original 5-point scale to a 6-point scale eliminated the scale's midpoint, thereby forcing respondents to decide whether each statement was characteristic of them. A response scale with an even number of points also better enables researchers to use a median-split on single items to categorize respondents as aggressive versus nonaggressive in specific situations (cf. Sudman & Bradburn, 1982). As a precedent, Velicer, Govia, Cherico, and Corribeau (1985) have modified the response scale of the Buss–Durkee Hostility Scale to make this instrument more reliable.

Criterion Measures

In addition to completing the AQ, a random subset of 180 participants in Sample 1 (70 males and 110 females) also filled out a set of criterion measures for use in evaluating the AQ's construct validity. These criterion measures served as standards for assessing the convergent and discriminant validity of the dimensions comprising both Buss and Perry's original four-factor model and the new, refined measurement model for the AQ.

Physical Aggression. As a criterion measure of physical aggression, we used the Assault subscale from the Buss–Durkee Hostility Inventory (Buss & Durkee, 1957). Buss (1961) reported a 5-week test–retest reliability of .78 for this subscale. As validity evidence, men who have committed domestic violence score higher on the Assault subscale compared to controls (Maiuro, Cahn, Vitaliano, Wagner, & Zegree, 1988). Although the original Buss–Durkee Assault subscale consisted of 10 items, we decided to use only those Assault items that had not been adapted by Buss and Perry in constructing the AQ. Specifically, 5 Assault subscale items were worded almost identically in Buss and Perry's AQ (Buss–Durkee items 9, 17, 25, 65, and 70). We chose to exclude these items from the Assault subscale because their comparable wording might otherwise spuriously inflate the degree of association between this criterion measure and the AQ (cf. Nichols, Licht, & Pearl, 1982).

Verbal Aggression. As a criterion measure of verbal aggression, we used the Verbal Hostility subscale of the Buss–Durkee Hostility Inventory (Buss & Durkee, 1957). Buss (1961) reported a 5-week test–retest reliability of .72 for this subscale. Supporting construct validity, Verbal Hostility score is a stronger predictor of hostile content in stories told in response to projective stimuli (Buss, Fischer, & Simmons, 1962) and of verbal aggression in role-playing responses to frustrating everyday events (Leibowitz, 1968) compared to the other Buss–Durkee subscales. Although the original Verbal Hostility subscale had 13 items, we omitted 5 of these (Buss–

Durkee items 7, 19, 23, 43, and 63) because of their nearly identical wording with AQ items to avoid artificially inflating the correlation between the AQ and this criterion measure (cf. Nichols et al., 1982).

Anger Arousal. As a criterion measure of anger, we chose the Anger Arousal subscale of the Multidimensional Anger Inventory (MAI; Siegel, 1986). We selected the Anger Arousal subscale of the MAI as an anger criterion measure because it more closely matches Buss and Perry's (1992) conceptual definition of anger as involving "physiological arousal" (p. 457) compared to other anger scales and because it has been found to be more reliable than the anger-range, anger-in, and anger-out MAI subscales (Siegel, 1986). Siegel (1986) reported reliability coefficients for this subscale of .83 for a college sample and .82 for a sample of factory workers. As validity evidence, Siegel (1986) found Anger Arousal scores correlated significantly with the magnitude and duration subscales of the Harburg Anger-In/Anger-Out Inventory (Harburg, Erfurt, Hauenstein, Chape, Schull, & Schork, 1973) and with the magnitude subscale of the Novaco Anger Inventory (Novaco, 1975). Although the original MAI Anger Arousal subscale contained eight items, we used only four of these (MAI items 9, 10, 14, and 26) and omitted items with wording that overlapped the Buss-Perry items to avoid artificially inflating the correlation between the AQ and this criterion measure (cf. Nichols et al., 1982).

Hostility. As a criterion measure of global hostility, we chose the Cook-Medley Hostility Scale (Ho; Cook & Medley, 1954). Based on a subset of 50 true-false items from the Minnesota Multiphasic Personality Inventory, the Ho is intended to assess primarily cynicism and paranoid alienation. Scores on this instrument have a 3-year test-retest correlation of .84 (Shekelle, Gale, Ostfeld, & Paul, 1983). As evidence of prospective validity, the scale appears to be an independent predictor of later coronary disease (Barefoot, Dahlstrom, & Williams, 1983).

Measurement Models Evaluated in This Study

A systematic literature search uncovered four different measurement models for the AQ currently in use: (a) a unidimensional "total score" model that assumes a single, global Aggression factor explains responses to all 29 AQ items; (b) Buss and Perry's (1992) original four-factor model (Physical Aggression, Verbal Aggression, Anger, and Hostility) for the 29-item AQ; (c) Buss and Perry's (1992) hierarchical version of this four-factor model that assumes a single second-order Aggression factor underlies the covariation among the four first-order factors; and (d) a modified four-factor model for 27 AQ items, consisting of Buss and Perry's original Physical Aggression, Verbal Aggression, and Anger factors, along with Harris's (1995) reduced Hostility factor (omitting its sixth and eighth indicators).

We also evaluated the goodness-of-fit of two new factor structures that we developed as potential measurement models for the AQ. The first of these new models consisted of our refined version of Buss and Perry's original four factors for a subset of 12 AQ items. The second new model was a hierarchical version of this refined four-factor model that assumes a single-order Aggression factor underlies the covariation among the four first-order factors.

Overview of Analyses

Our analyses addressed four main questions. (1) Do the existing one-, four-, or hierarchical-factor structures provide an acceptable measurement model for the AQ? To answer this question, we used confirmatory factor analysis (CFA) to impose each of these factor models on three different data sets and to evaluate each model's goodness-of-fit across samples. (2) If available models prove inadequate, then what might be a more appropriate measurement model for the AQ? Here we used principal components analysis to eliminate unreliable AQ items in order to develop an acceptable measurement model. (3) Are the same measurement models

warranted for males and females? To address this question, we used multigroup CFA to test hypotheses about the invariance of the refined AQ measurement model with respect to gender. (4) Is there any evidence for the convergent or discriminant validity of the multiple AQ factors? Here we used CFA to test hypotheses about the relationships among the AQ factors and the criterion measures of aggression.

Analysis Strategy

Stage One. The analysis unfolded in four stages, each using CFA via LISREL 8 (Joreskog & Sorbom, 1996). In Stage One, we began by using the data from Samples 1–3 to examine the fit of the four, existing measurement models for the AQ. To evaluate each model's goodness-of-fit to the data, we used three measures of absolute fit and two measures of relative fit (cf. Hu & Bentler, 1998). These multiple measures of model fit provide complementary information about how well a particular model explains the data and should not be construed as redundant (cf. Bollen, 1989; McDonald & Marsh, 1990).

As measures of each model's *absolute* fit, we used the ratio of chi-square to degrees of freedom (χ^2/df ; Hoelter, 1983), the goodness-of-fit index (GFI; Joreskog & Sorbom, 1996), and the root-mean-square error of approximation (RMSEA; Steiger, 1989). Although the first two of these measures of model fit have limited utility (Hu & Bentler, 1998), we have reported them in order to maximize the comparability of our results with those of prior researchers who reported these fit measures for the same models. In judging absolute fit, smaller ratios of chi-square to degrees of freedom reflect better absolute fit, with ratios near two considered acceptable (Hoelter, 1983). Analogous to R^2 in multiple regression, GFI reflects the proportion of available variation–covariation information in the data that the given model explains, with larger GFI values representing better model fit. Bentler and Bonett (1980) have suggested that formal measurement models have a $GFI \geq .90$. RMSEA reflects the size of the residuals that result when using the model to predict the data, with smaller values indicating better fit. According to Browne and Cudeck (1993), RMSEA of .05 or lower represents “close fit,” RMSEA between .05 and .08 represents “reasonably close fit,” and RMSEA above .10 represents “an unacceptable model.” We also directly compared the absolute fit of nested models by contrasting their goodness-of-fit chi-square values and computing the p value associated with the difference in these nested chi-squares (with accompanying difference in degrees of freedom).

As measures of each model's *relative* fit, we used the comparative fit index (CFI; Bentler, 1990) and the nonnormed fit index (NNFI; Bentler & Bonett, 1980; Tucker & Lewis, 1973). We have chosen to report these particular measures because they have more desirable psychometric properties than other measures of relative fit (Bentler, 1990; Hu & Bentler, 1998; Marsh, Balla, & Hau, 1996; McDonald & Marsh, 1990). Each of these two relative fit measures uses a different formula to contrast the goodness-of-fit chi-square of a given model with that of a “null” model, which assumes sampling error alone explains the covariation among observed measures (i.e., that there is no common variance among the AQ items). For each relative fit index, larger values represent better fit, with values of .90 or higher considered acceptable (Bentler & Bonett, 1980).

Stage Two. Given the poor fit of the unidimensional and multidimensional models, we sought to develop a better-fitting, refined measurement model in Stage Two of the analysis. Stage Two consisted of four phases. In the *first phase* (model refinement), we subjected the data of Sample 1 to a principal components analysis (PCA) with oblique rotation in order to explore the structure underlying the AQ. Extracting four factors, we looked for items with low communalities or multiple loadings across factors to identify questions that are unreliable indicators of aggression or that reflect more than one dimension of aggression. We also decided to omit items that are reverse-scored because we wanted the refined measurement model to use only indicators that reflect the endorsement of aggression rather than the rejection of

nonaggression. This latter approach has been useful in refining measurement models for other constructs, such as affect intensity (Bryant et al., 1996).

In the *second phase* of Stage Two (model evaluation), we used CFA to impose a parsimonious confirmatory version of this new measurement model on the data of Samples 1–3, using the refined subset of AQ items. We also evaluated a hierarchical form of this model in which a single, second-order factor explained the relationships among the four first-order factors. In addition, we used multigroup CFA to evaluate the generalizability of these refined measurement models across Samples 1–3. In these multigroup analyses, we contrasted the goodness-of-fit chi-squares of two nested CFA models: one constraining the magnitudes of the factor loadings to be equal for all three samples and the other omitting this invariance constraint. A statistically significant difference in the chi-square values of these two models ($\Delta\chi^2$) indicates that the given model yields different factor loadings across samples (cf. Bryant & Baxter, 1997; Bryant & Yarnold, 1995). Given a significant overall structural difference, we used multigroup CFA with equality constraints to pinpoint the specific items responsible for cross-sample differences. In accordance with standard practice in the structural equation modeling literature, all multigroup analyses were performed using covariance matrices as input (cf. Joreskog & Sorbom, 1996).

In the *third phase* of Stage Two, for comparison purposes we tested a condensed, alternative form of Buss and Perry's (1992) original four-factor that preserved the full range of item content while reducing the number of indicators per factor. Here we used a "partial disaggregation" approach (Bagozzi & Edwards, 1998; Bagozzi & Heatherton, 1994; Hull, Lehn, & Tedlie, 1991) to reconfigure Buss and Perry's (1992) original "total disaggregation" model (which had nine indicators for PA, five indicators for VA, seven indicators for ANG, and eight indicators for HO) in terms of three indicators for each latent variable. This entailed parceling individual items into composite indicators for each factor, so as to modify the "atomistic" original model into a more "molecular" form. As Bagozzi and Heatherton (1994) have noted, CFA models containing more than about five measures per factor are unlikely to fit the data satisfactorily. For this reason, we sought to test the goodness-of-fit of a condensed form of Buss and Perry's original four-factor model and to evaluate its cross-sample generalizability.

In the *fourth phase* of Stage Two (testing gender invariance), we used multigroup CFA to determine whether the refined measurement models provided an equivalent goodness-of-fit to the data of males and females in Samples 1 and 2. This involved contrasting the goodness-of-fit chi-squares of two nested CFA models: one constraining the magnitudes of the factor loadings to be equal for males and females and the other omitting this invariance constraint. A statistically significant difference in the chi-square values of these two models ($\Delta\chi^2$) indicates that the given model yields different factor loadings for men and women (cf. Bryant & Baxter, 1997; Bryant & Yarnold, 1995). Given a significant overall structural difference, we used multigroup CFA with equality constraints to pinpoint the specific items responsible for gender differences. Again, all multigroup analyses were performed on covariance matrices (cf. Joreskog & Sorbom, 1996).

Stage Three. In Stage Three of the analysis, we evaluated and compared the convergent and discriminant validity of the refined and original versions of the four-factor model. Here we used CFA with equality constraints to determine whether each of the four AQ factors showed a stronger relationship with the criterion measure to which it is presumed to correspond than with the other criterion measures.

Stage Four. Having established the cross-sample generalizability of the refined measurement model, we sought to reconfirm it in the final stage of the analysis using a new "short form" of the AQ, which contains only the refined subset of 12 items. Here we used CFA to impose the refined measurement model on the data of Samples 4 and 5, who completed this new, short form of the AQ. We also used multigroup CFA to assess the generalizability of the refined measurement model across Samples 4 and 5 as well as the model's invariance with respect to gender in both of these samples.

RESULTS AND DISCUSSION

Stage One: Assessing Current Measurement Models

We used CFA first to evaluate the goodness-of-fit of the four alternative measurement models currently in the literature, using the AQ data of Samples 1–3. Which model, if any, provides the most reasonable representation of responses to this instrument? Table 1 presents the results of these analyses. Across all three samples, the explanatory power of existing measurement models fell short of accepted standards. First, a single “AQ total score” provided an inadequate measurement model for Samples 1–3, leaving too much common variance unexplained in both an absolute (GFIs = .65 – .70) and relative (fit indices = .59 – .66) sense. For all three samples, this one-factor model also had a relatively high ratio of chi-square to degrees of freedom ($\chi^2/dfs = 3.4 - 4.2$) and an unacceptably large, root-mean-square error of approximation (RMSEAs = .098 – .109). These findings support the AQ’s multidimensionality.

Yet, none of the multidimensional models currently available provides an acceptable measurement model for the AQ. Both Buss and Perry’s (1992) original four-factor model and its hierarchical counterpart fit the data of all three samples better than the unidimensional model, all $\Delta\chi^2(6) > 379.2$, $p < .0001$. However, for all three samples, each of these multifactor structures fails to achieve sufficient goodness-of-fit to be an acceptable measurement model in both absolute (GFIs = .76 – .81) and relative (fit indices = .76 – .82) terms. Furthermore, the chi-square to degrees of freedom ratios for these models ($\chi^2/dfs = 2.4 - 2.8$) show “a poor fit” (Buss & Perry, 1992, p. 454), and their RMSEAs show room for improvement (RMSEAs = .072 – .084). Evidently, current measurement models of the AQ operationalize underlying subtraits of aggression in ways that do not correspond closely enough to the conceptual framework that Buss and Perry (1995) intended. This conclusion is consistent with those of previous researchers (e.g., Archer, Kilpatrick, & Bramwell, 1995; Harris, 1995; Williams et al., 1996) who have noted the inadequacy of existing factor models for the AQ.

Imposed on the data of Samples 1–3, the four-factor model containing Harris’s (1995) condensed Hostility factor fared little better (see Table 1). Although this modified model was a significant improvement in fit over the original four-factor model for all three samples, all $\Delta\chi^2(53)s > 143.6$, $p < .0001$, it nevertheless fell short of accepted standards for a formal measurement in both an absolute (GFIs = .78 – .83) and relative (fit indices = .79 – .83) sense. Taken as a whole, this evidence underscores the need for a better fitting measurement model for the AQ.

Stage Two: Developing a Better Fitting Measurement Model

What might a more appropriate measurement model look like, and how should we best go about developing it? In answering these questions, we

TABLE 1
Goodness-of-Fit Statistics for Various Measurement Models of the AQ Imposed on Samples 1-3

Model	No. items	Sample	Absolute fit measures				Relative fit measures	
			χ^2	df	χ^2/df	GFI	RMSEA	CFI NNFI
One-factor (total score)	29	1	1567.9	377	4.2	.70	.102	.66 .64
		2	1267.8	377	3.4	.65	.109	.62 .59
		3	1469.1	377	3.9	.68	.098	.66 .63
Buss & Perry's four factors:	29	1	1042.8	371	2.8	.81	.077	.81 .79
		2	886.4	371	2.4	.76	.084	.78 .76
		3	950.3	371	2.6	.81	.072	.82 .80
Buss & Perry's hierarchical model:	29	1	1046.4	373	2.8	.81	.077	.81 .79
		2	888.5	373	2.4	.76	.083	.78 .76
		3	969.6	373	2.6	.81	.072	.81 .80
one second-order factor	27	1	881.9	318	2.9	.82	.076	.83 .82
		2	734.2	318	2.3	.78	.081	.81 .79
		3	806.6	318	2.5	.83	.071	.83 .81
PA, VA, ANG, & Harris's HO factor	12	1	105.7	48	2.2	.94	.063	.96 .94
		2	92.4	48	1.9	.93	.068	.95 .93
		3	121.7	48	2.5	.94	.071	.91 .87
Four refined factors:	12	1	108.5	50	2.2	.94	.062	.96 .94
		2	94.4	50	1.9	.93	.067	.95 .93
		3	133.6	50	2.7	.93	.074	.90 .86
Refined hierarchical model:	12	1						
		2						
		3						
one second-order factor		1						
		2						
		3						

Note. PA = Physical Aggression; VA = Verbal Aggression; ANG = Anger; HO = Hostility; GFI = goodness-of-fit index (Joreskog & Sorbom, 1996); RMSEA = root-mean-square error of approximation (Steiger, 1989); CFI = comparative fit index (Bentler, 1990); NNFI = nonnormed fit index (Bentler & Bonett, 1980). Sample 1 = 307 American undergraduates; Sample 2 = 200 British undergraduates (Archer, Holloway, & McLoughlin, 1995); Sample 3 = 306 Canadian undergraduates (Harris, 1995). The above goodness-of-fit statistics are from analyses conducted via LISREL 8 (Joreskog & Sorbom, 1996).

sought first to preserve the solid theoretical foundation underlying Buss and Perry's (1992) original four-factor model, while at the same time sharpening its measurement focus. Accordingly, we set out not only to maintain the conceptual definitions of the original factors, but also to improve the four-factor model's goodness-of-fit by eliminating AQ items that were relatively unreliable indicators of the dimensions they were intended to reflect.

Model refinement. With these goals in mind, we inspected the results from principal components analysis (PCA) of Sample 1 data and decided to exclude AQ items according to three criteria. First, we eliminated items with low loadings ($\lambda < .40$) in order to increase the proportion of variance that factors explained in their constituent indicators. Second, to enhance the conceptual clarity of the model, we excluded items that loaded at least moderately ($\lambda = .40$) on two or more scales, based on CFA modification indices and PCA results. Third, to improve conceptual precision, we omitted items that did not reflect the direct endorsement of aggressive traits. The AQ includes two reverse-scored items—a Physical Aggression item ("I can think of no good reason for ever hitting a person") and an Anger item ("I am an even-tempered person")—which entail the rejection of nonaggressive traits rather than the acceptance of aggressive characteristics. These three exclusion criteria yielded a refined 12-item, four-factor model that reflects the same underlying constructs (Physical Aggression, Verbal Aggression, Anger, and Hostility) as Buss and Perry's (1992) original model, but with an equal number of items for each factor. Table 2 presents the items comprising both the original and refined versions of the four-factor model of the AQ.

Model evaluation. How well does this refined four-factor model explain responses to the AQ? To address this question, we used CFA to impose the refined measurement model on the data of Samples 1–3. In doing this, we have used Sample 1 to modify the model post hoc and used Samples 2 and 3 to cross-validate the model a priori. This strategy enabled us to assess the degree to which model respecifications based on Sample 1 generalized across independent samples, so as to avoid being misled by the unique characteristics of a single sample (cf. MacCallum, 1986; MacCallum, Roznowski, & Necowitz, 1992).

Table 1 also presents the results of these analyses. As seen in this table, across all three samples, the refined four-factor model explains an acceptable proportion of common variance in both absolute (GFIs = .93 – .94) and relative (fit indices = .87 – .96) terms. Although its ratio of chi-square to degrees of freedom shows cross-sample inconsistency ($\chi^2/dfs = 1.9 - 2.5$), the refined model's RMSEA reflects reasonably close fit across all three samples (RMSEAs = .063 – .071). Considered together, these findings suggest that the modified four-factor model is an appropriate measurement model for the AQ.

TABLE 2
Items Constituting the Original and Refined Measurement Models of the AQ

Factor	Constituent items
Physical Aggression	<ol style="list-style-type: none"> 1. Once in a while I can't control the urge to hit another person. 2. Given enough provocation, I may hit another person. 3. If somebody hits me, I hit back. 4. I get into fights a little more than the average person. 5. If I have to resort to violence to protect my rights, I will. 6. There are people who pushed me so far that we came to blows. 7. I can think of no good reason for ever hitting a person. [reverse-scored] 8. I have threatened people I know. 9. I have become so mad that I have broken things.
Verbal Aggression	<ol style="list-style-type: none"> 10. I tell my friends openly when I disagree with them. 11. I often find myself disagreeing with people. 12. When people annoy me, I may tell them what I think of them. 13. I can't help getting into arguments when people disagree with me. 14. My friends say that I'm somewhat argumentative.
Anger	<ol style="list-style-type: none"> 15. I flare up quickly but get over it quickly. 16. When frustrated, I let my irritation show. 17. I sometimes feel like a powder keg ready to explode. 18. I am an even-tempered person. [reverse-scored] 19. Some of my friends think I'm a hothead. 20. Sometimes I fly off the handle for no good reason. 21. I have trouble controlling my temper.
Hostility	<ol style="list-style-type: none"> 22. I am sometimes eaten up with jealousy. 23. At times I feel I have gotten a raw deal out of life. 24. Other people always seem to get the breaks. 25. I wonder why sometimes I feel so bitter about things. 26. I know that "friends" talk about me behind my back. 27. I am suspicious of overly friendly strangers. 28. I sometimes feel that people are laughing at me behind my back. 29. When people are especially nice, I wonder what they want.

Note: The 29 items constituting the original four-factor model for the Aggression Questionnaire (AQ) are listed in the order presented by Buss and Perry (1992, Table 1, p. 454). Buss and Perry (1992) instructed researchers to randomly order the above items when administering the AQ. The order of the 29 AQ items for Sample 1 was 26, 15, 11, 10, 16, 27, 2, 4, 29, 12, 21, 13, 14, 8, 5, 22, 28, 7, 20, 6, 25, 19, 23, 3, 24, 1, 9, 17, and 18. Items in bold comprise the refined four-factor measurement model. The randomized order of these questions in the 12-item short form of the AQ is 11, 23, 8, 25, 21, 14, 15, 2, 13, 24, 6, and 20. From "The Aggression Questionnaire," by A. H. Buss and M. Perry, 1992, *Journal of Personality and Social Psychology*, **63**, p. 454 (Table 1). Copyright by the American Psychological Association. Adapted with permission.

Results converge on an identical conclusion concerning the hierarchical version of this refined model. Specifically, a model specifying a single, overarching second-order aggression trait that explains the covariation among the four first-order factors fit the data of all three samples reasonably well, both absolutely (GFIs = .93 – .94) and relatively (fit indices = .86 – .96). This hierarchical model also showed chi-square to degrees of freedom ratios ($\chi^2/dfs = 1.9 - 2.7$) and RMSEAs (= .062 – .074) comparable to those of the refined four-factor model across the three samples. Thus, both the refined four-factor model and its hierarchical counterpart represent acceptable measurement models for the AQ.

Testing cross-sample generalizability. Having identified an appropriate measurement structure, we next used multigroup CFA to evaluate the cross-sample generalizability of the four-factor model more systematically. We considered first the question of whether the refined four-factor model produces the same factor loadings across the three samples. In other words, do Physical Aggression, Verbal Aggression, Anger, and Hostility have the same meanings for American (Sample 1), British (Sample 2), and Canadian (Sample 3) undergraduates?

Table 3 presents the loadings for the four-factor model imposed on the data of each sample. An initial omnibus test revealed that the magnitudes of these factor loadings varied across samples, $\Delta\chi^2(16, n = 813) = 29.3, p < .022$. Following up the omnibus test, only Samples 1 and 3 showed significant differences in factor loadings, $\Delta\chi^2(8, n = 613) = 24.3, p < .003$, whereas loadings were equivalent for Samples 1 and 2, $\Delta\chi^2(8, n = 507) = 11.1, p > .19$; and Samples 2 and 3, $\Delta\chi^2(8, n = 506) = 7.7, p > .46$. Additional multigroup CFAs disclosed that only one factor loading actually differed significantly for Samples 1 and 3—the loading for Anger item 1 (“I flare up quickly but get over it quickly”) was stronger for the Canadian sample, $\Delta\chi^2(1, n = 613) = 13.8, p < .0003$ —and all other loadings for the four-factor model were statistically comparable, $\Delta\chi^2(7, n = 613) = 12.5, p > .09$. Supporting this conclusion, a model that constrains (a) all factor loadings except Anger item 1 to be equal across Samples 1–3 and (b) Anger item 1 to load equally for Samples 1 and 2 but not for Sample 3 fit the data of the three samples no worse than a multigroup model with no equality constraints, $\Delta\chi^2(15, n = 813) = 19.5, p > .19$. These results indicate that the loadings of the four-factor model are largely (11/12 = 92%) invariant across the three samples; they also suggest that the meaning of aggression, as defined by the refined measurement model, holds across culture.

The same cannot be said of Buss and Perry’s (1992) original four-factor model, which produced nonequivalent loadings for all three samples, $\Delta\chi^2(50, n = 813) = 554.8, p < .0001$. Indeed, the original four-factor model yielded strong differences in loadings when comparing Samples 1 and 2, $\Delta\chi^2(25, n = 507) = 67.6, p < .0001$; Samples 2 and 3, $\Delta\chi^2(25, n = 506) = 634.4,$

TABLE 3
CFA Factor Loadings for the Refined 12-Item, Four-Factor Measurement Model of the AQ

	PA sample			VA sample			ANG sample			HO sample		
	1	2	3	1	2	3	1	2	3	1	2	3
AQ items												
2. Given enough provocation, I may hit another person.	76	70	58									
6. There are people who pushed me so far that we came to blows.	72	73	65									
8. I have threatened people I know.	80	82	68									
11. I often find myself disagreeing with people.				80	75	70						
13. I can't help getting into arguments when people disagree with me.				82	71	68						
14. My friends say that I'm somewhat argumentative.				58	61	76						
15. I flare up quickly but get over it quickly.							50	62	69			
20. Sometimes I fly off the handle for no good reason							81	83	57			
21. I have trouble countrolling my temper.							71	71	34			
23. At times I feel I have gotten a raw deal out of life.										65	76	45
24. Other people always seem to get the breaks.										77	75	64
25. I wonder why sometimes I feel so bitter about things.										68	68	52

Note. PA = Physical Aggression; VA = Verbal Aggression; ANG = Anger; HO = Hostility. Decimal points are omitted. Item numbers in parentheses refer to the original ordering of items in Buss and Perry's (1992, p. 454) Table 1. Blank loadings were fixed at zero. The loadings of items 2, 14, 20, and 23 were fixed at unstandardized values of 1.0 to scale the latent variables in multigroup confirmatory analyses.

$p < .0001$; and Samples 1 and 3, $\Delta\chi^2(25, n = 613) = 126.9, p < .0001$. Clearly, the refined measurement model is superior to the original in terms of both its goodness-of-fit and its cross-cultural generalizability.

Comparing the refined hierarchical model across samples, the second-order Aggression factor had the same relationships with the four first-order factors in Samples 1 and 2, $\chi^2(3, n = 507) = 0.6, p > .89$; but the Aggression "super factor" had less to do with Anger in Canadian Sample 3 compared to American Sample 1, $\chi^2(1, n = 613) = 25.5, p < .0001$; and British Sample 2, $\chi^2(1, n = 506) = 24.6, p < .0001$ (see Table 4). These results are consistent with the earlier finding that the four-factor model fits the data of Sample 3 better than the hierarchical model.

Comparing the original hierarchical model across samples, the second-order Aggression factor had the same relationships with the four first-order factors in Samples 1 and 2, $\chi^2(3, n = 507) = 1.9, p > .59$; but the Aggression "super factor" had less to do with Anger in Canadian Sample 3 compared to American Sample 1, $\chi^2(1, n = 613) = 22.6, p < .00005$; and British Sample 2, $\chi^2(1, n = 506) = 28.8, p < .00001$ (see Table 4). Unlike the refined hierarchical model, however, second-order Aggression also had more to do with Hostility in Canadian Sample 3 compared to the American Sample 1, $\chi^2(1, n = 613) = 34.0, p < .00001$; and British Sample 2, $\chi^2(1, n = 506) = 16.6, p < .0009$ (see Table 4). Thus, the refined second-order CFA model showed stronger cross-cultural generalizability than did the original second-order CFA model.

We also addressed the question of whether the four factors comprising the refined model interrelate in the same ways for the American, British, and Canadian samples. Table 5 presents the reliabilities and factor intercorrelations for this model imposed on the data of Samples 1–3. Using the model with partially invariant loadings as a baseline (cf. Byrne, Shavelson, & Muthén, 1989), the refined four-factor model produced different factor correlations for the three samples, $\Delta\chi^2(12, n = 813) = 65.3, p < .0001$. Although Samples 1 and 2 had equivalent factor intercorrelations, $\Delta\chi^2(6, n = 507) = 3.1, p > .79$; Sample 3 had different factor intercorrelations compared to Sample 1, $\Delta\chi^2(6, n = 613) = 53.1, p < .0001$; and Sample 2, $\Delta\chi^2(6, n = 506) = 46.3, p < .0001$. Thus, the four refined AQ factors interrelated differently among the Canadian sample than among the American and British samples.

Although the correlation between Physical and Verbal Aggression was the same in all three samples, $\Delta\chi^2(2) = 4.2, p > .12$, Physical Aggression correlated more strongly with Anger in the Canadian sample than in the other two samples, both $\Delta\chi^2(1)s > 5.1, ps < .025$. For the Canadian sample, Verbal Aggression also correlated more strongly with Anger, both $\Delta\chi^2(1)s > 14.0, ps < .002$; but Hostility correlated less strongly with Physical Aggression, both $\Delta\chi^2(1)s > 9.2, ps < .0025$, with Verbal Aggression, both

TABLE 4

Second-Order Factor Loadings and Residual Variances of the First-Order Factors for the Original and Refined Hierarchical CFA Model Imposed on the AQ Data of Samples 1-3

AQ factors	Model	Sample 1				Sample 2				Sample 3			
		PA	VA	ANG	HO	PA	VA	ANG	HO	PA	VA	ANG	HO
Factor loading (γ)	Original	1.00*	.99	1.01	.54	1.00*	1.18	1.33	.63	1.00*	.93	.42	1.44
	Refined	1.00*	.80	1.21	.77	1.00*	.75	1.23	.81	1.00*	.96	.47	1.01
Standard error of factor loading	Original	—	.10	.11	.07	—	.16	.17	.11	—	.12	.08	.17
	Refined	—	.12	.16	.12	—	.13	.20	.16	—	.16	.11	.18
Standardized factor loading	Original	.76	.80	.99	.64	.76	.86	.95	.61	.75	.65	.63	.99
	Refined	.71	.83	.90	.66	.77	.84	.86	.54	.71	.62	.70	.99
Squared multiple correlation	Original	.57	.64	.99	.41	.58	.73	.90	.37	.56	.43	.40	.99
	Refined	.50	.68	.81	.43	.59	.70	.74	.29	.51	.38	.48	.99
Residual variance (Ψ)	Original	.55	.40	.01	.31	.35	.25	.10	.33	.33	.48	.11	.01
	Refined	.38	.11	.13	.30	.30	.10	.22	.70	.32	.51	.08	.01
Standard error of residual variance	Original	.08	.07	.04	.06	.07	.08	.06	.07	.06	.08	.03	.06
	Refined	.07	.11	.05	.06	.08	.04	.08	.14	.08	.09	.03	.04
Standardized residual variance	Original	.43	.36	.01	.59	.42	.27	.10	.63	.44	.57	.60	.01
	Refined	.50	.32	.19	.57	.41	.30	.26	.71	.49	.62	.52	.01

Note: PA = Physical Aggression; VA = Verbal Aggression; ANG = Anger; HO = Hostility. Sample 1 = 307 American undergraduates; Sample 2 = 200 British undergraduates (Archer, Holloway, & McLoughlin, 1995); Sample 3 = 306 Canadian undergraduates (Harris, 1995). Tabled are results from second-order CFA models imposed separately on the data of each sample. Second-order factor loadings (γ s) are unstandardized regression coefficients representing the amount of change in first-order factor scores resulting from a one-unit change in the second-order latent variable. Standardized second-order factor loadings are standardized regression coefficients representing the change in standard deviations in first-order factor scores resulting from a 1-standard-deviation change in the second-order factor. Squared multiple correlations represent the proportion of total variance in each first-order factor that is associated with the second-order factor, analogous to R^2 in multiple regression. Residual variances (ψ s) represent the amount of variance in first-order factors that is unexplained by the second-order factor. Standardized residual variances represent the proportion of total variance in each first-order factor that is unrelated to the second-order factor and that is specific to that first-order factor. Standard errors of second-order factor loadings (and of residual variances) represent the standard deviation of changes in loadings (and in residual variances) that would be expected to occur from sample to sample.

* Fixed at value of 1.0 to define the metric of the second-order latent variable in the unstandardized solution.

TABLE 5
Reliabilities and CFA Factor Intercorrelations for the Original and Refined Four-Factor Models

AQ factors	Model	Sample 1				Sample 2				Sample 3			
		PA	VA	ANG	HO	PA	VA	ANG	HO	PA	VA	ANG	HO
PA	Original	86 ^a				84 ^a				84 ^b			
	Refined	80 ^a				80 ^a				—			
VA	Original	60	74 ^a			64	68 ^a			62	75 ^b		
	Refined	57	77 ^a			63	73 ^a			58	—		
ANG	Original	74	81	82 ^a		72	82	83 ^a		37	37	80 ^b	
	Refined	63	76	71 ^a		66	74	73 ^a		39	35	—	
HO	Original	54	47	63	76 ^a	52	48	57	81 ^a	75	64	67	83 ^b
	Refined	53	53	57	73 ^a	47	43	43	77 ^a	80	66	85	—

Note: PA = Physical Aggression; VA = Verbal Aggression; ANG = Anger; HO = Hostility. Sample 1 = 307 American undergraduates; Sample 2 = 200 British undergraduates (Archer, Holloway, & McLoughlin, 1995); Sample 3 = 306 Canadian undergraduates (Harris, 1995). Tabled below the reliability coefficients are factor correlations from confirmatory factor analyses in which latent factors and measured variables have been standardized separately within each sample. Decimal points are omitted.

^a Cronbach's alpha (an index of internal consistency) for unit-weighted original and refined factor scores.

^b Reliability coefficients originally reported by Harris (1995, Table 2, p. 993). We could not compute reliability coefficients for the refined factors in Sample 3 because only the covariance matrix (but not the necessary raw data) was available for reanalysis of this data set.

$\Delta\chi^2(1)s > 3.8$, $ps < .05$, and with Anger, both $\Delta\chi^2(1)s > 7.5$, $ps < .006$, compared to the American and British samples. Although interesting from a cross-cultural perspective, these differences in factor interrelationships do not alter the conclusion that the 12 AQ indicators measure the latent factors in comparable ways for all three samples (cf. Kline, 1998).

Comparing the original and refined factors. For purposes of comparison, Table 5 also displays the correlations among Buss and Perry's (1992) original four factors in these same three samples. As evident in this table, the pattern of correlations among the refined factors is strikingly similar to the pattern of correlations among the original Buss–Perry factors. Indeed, the refined factors have only slightly lower internal consistency reliabilities than their original counterparts. This is important because it suggests that refining the factors improved the model's overall goodness-of-fit, but did not substantially reduce the reliabilities of the individual factors.¹

Testing a "Partially Disaggregated" form of Buss and Perry's original model. Before abandoning Buss and Perry's (1992) original four-factor model, we reconfigured it into a "partially disaggregated" measurement model that preserved the full range of "totally disaggregated" item content while reducing the number of indicators per factor (cf. Bagozzi & Edwards, 1998; Bagozzi & Heatherton, 1994; Hull et al., 1991). To do this, we modified the nine single-item indicators for PA into three composite measures, using as indicators (a) the mean of the three PA items from the refined measurement model; (b) the mean of original AQ items 1, 3, and 4; and (c) the mean of original AQ items 5, 7, and 9. We modified the five single-item indicators for VA into three measures, using as indicators (a) the mean of the three VA items from the refined measurement model, (b) original AQ item 10, and (c) original AQ item 12. We modified the seven single-item indicators for ANG into three composite measures, using as indicators (a) the mean of the three ANG items from the refined measurement model, (b) the mean of original AQ items 16 and 17, and (c) the mean of original AQ items 18 and 19. Finally, we modified the eight single-item indicators for HO into three composite measures, using as indicators (a) the mean of the three HO items from the refined measurement model; (b) the mean of original AQ items 22, 26, and 27; and (c) the mean of original AQ items 28 and 29.

We then imposed this "partially disaggregated" four-factor model on the

¹ As further evidence concerning the degree of conceptual overlap between the refined and original Buss–Perry factors, we correlated unit-weighted factor scores for the former and the latter within Samples 1 and 2. (We were unable to correlate factor scores in Sample 3 because only the covariance matrix, and not the necessary raw data, was available for reanalysis.) For Samples 1 and 2, respectively, these intercorrelations were uniformly high: (a) Physical Aggression (.91 and .86), (b) Verbal Aggression (.90 in both samples), (c) Anger (.90 and .85), and (d) Hostility (.83 and .85). These findings suggest that the refined AQ factors basically measure the same latent constructs as the original factors.

data of Samples 1 and 2 to evaluate its goodness-of-fit and cross-sample generalizability. (We could not construct partially disaggregated indicators for Sample 3 because only the covariance matrix was available for reanalysis of this sample.) Although the partial disaggregation model provided a satisfactory fit to the data of Sample 1, $\chi^2(48, n = 307) = 157.6$, $\chi^2/df = 3.3$, GFI = .93, RMSEA = .083, CFI = .93, NNFI = .91; it did not adequately fit the data of Sample 2, $\chi^2(48, n = 200) = 197.4$, $\chi^2/df = 4.1$, GFI = .87, RMSEA = .120, CFI = .88, NNFI = .83. Multigroup analyses further demonstrated that neither the factor loadings, $\Delta\chi^2(8, n = 507) = 32.2$, $p < .0001$, nor the factor variances–covariances, $\Delta\chi^2(10, n = 507) = 115.8$, $p < .00001$, of this partially disaggregated model were invariant across the two samples. Thus, even in a “partially disaggregated” form, Buss and Perry’s (1992) original four-factor model does not provide an acceptable measurement model for the AQ.

Testing the Gender Invariance of the Refined Model. Does aggression mean the same thing to men and women? Is the refined measurement model for the AQ equally applicable to the data of males and females? This is a critical question for researchers interested in using the AQ to compare levels of aggression in men and women. With respect to this question, Buss and Perry (1992) reported that men’s and women’s loadings differed in separate four-factor PCA solutions, though they did not directly test the gender invariance of the four-factor model.

The refined four-factor model generated gender-equivalent loadings for the British Sample 2, $\Delta\chi^2(8, n = 200) = 6.0$, $p > .64$, but not for the American Sample 1, $\Delta\chi^2(8, n = 304) = 21.4$, $p < .007$. (Because only the pooled covariance matrix was available for reanalysis of Harris’s data, we could not test for gender invariance in Sample 3.) Follow-up multigroup CFAs revealed that only one factor loading actually differed significantly for males and females in Sample 1—the loading for Physical Aggression item 3 (“There are people who have pushed me so hard that we came to blows”) was stronger for men than women, $\Delta\chi^2(8, n = 304) = 12.3$, $p < .0005$ —and all other loadings for the four-factor model were gender-equivalent, $\Delta\chi^2(7, n = 304) = 2.9$, $p > .89$. These findings suggest that the loadings of the refined four-factor model are largely invariant with respect to gender. Likewise for the hierarchical model, the overarching second-order Aggression factor showed comparable relationships with the four first-order factors for males and females in both Sample 1, $\Delta\chi^2(4, n = 304) = 1.5$, $p > .82$, and Sample 2, $\Delta\chi^2(4, n = 200) = 5.9$, $p > .20$. Thus, the refined factors appear to have substantially the same meaning for men and women.

Stage Three: Assessing Construct Validity

What evidence of convergent or discriminant validity is there for the refined AQ factors? And how does this validity evidence compare to that for

Buss and Perry's original AQ factors? To address these questions, we used CFA with equality constraints in Sample 1 data to test hypotheses about relationships between (a) the four AQ factors in both their refined and original forms; and (b) the four criterion measures of physical assault, verbal hostility, anger arousal, and global hostility. We used CFA rather than traditional correlational methods because it allowed us to estimate the relationship between aggression subtraits and criterion measures, partialing out measurement error, and also provided a way to systematically test hypotheses about the strength of associations across measures.

Because measurement error attenuates relationships, different measures may demonstrate different interrelationships due to differences in reliability. For this reason, it is important to control for differential reliability when assessing the strength of relationships between observed measures. Traditional correlational methods assume that all analyzed variables are measured perfectly and therefore do not allow researchers to adjust for differential reliability (Kline, 1998; Maruyama, 1998). Using CFA, in contrast, enabled us to control for differences in the reliabilities of the AQ factors and the criterion measures, which might otherwise influence the strength of the observed associations (Bagozzi, 1993; Judd, Jessor, & Donovan, 1986).

Another advantage of CFA over the traditional correlational approach is that it allowed us to use equality constraints to systematically test hypotheses about the strength of the relationships among the latent constructs. Typically, researchers have simply eyeballed differences in correlation coefficients to determine the degree to which measures show convergent or discriminant validity. Using CFA, in contrast, enabled us to test the statistical significance of differences in the magnitude of validity coefficients by contrasting the goodness-of-fit chi-square values (and degrees of freedom) of two nested models: one that constrained the correlations in question to have equal value and one that contained no equality constraint. A significant difference in these two nested chi-square values signifies that the correlations differ in magnitude (cf. Bryant & Baxter, 1997; Bryant & Yarnold, 1995).

To obtain the multiple indicators required for CFA while also minimizing the number of measured variables in the model, we used the "partial disaggregation" approach to parcel each of the four criterion measures into two composite indicators. For the Buss–Durkee Physical Assault scale, we summed responses to the odd-numbered items (1, 3, and 5) to create one indicator and summed responses to the even-numbered items (2 and 4) to create a second indicator. For the Buss–Durkee Verbal Hostility scale, we summed responses to items 1–4 to create one indicator and summed responses to items 5–8 to create a second indicator. For the MAI Anger Arousal scale, we summed responses to items 1 and 2 to create one indicator and summed responses to items 3 and 4 to create a second indicator. Total scores on split-halves of the Cook–Medley Hostility Scale served as indica-

TABLE 6
Correlations of the Refined and Original AQ Factors with the Criterion Measures

Criterion measures	Refined AQ factors				Original AQ factors			
	PA	VA	ANG	HO	PA	VA	ANG	HO
Buss–Durkee Physical Assault	85	62	65	43	90	62	70	44
Buss–Durkee Verbal Hostility	57	64	72	45	63	66	74	45
MAI Anger Arousal	64	66	91	77	64	64	93	88
Cook–Medley Hostility	46	60	64	89	49	58	65	84

Note. PA = Physical Aggression; VA = Verbal Aggression; ANG = Anger; HO = Hostility. These data are from a random subset of 180 American undergraduates (70 males and 110 females) from Sample 1. Tabled are standardized ϕ coefficients (analogous to Pearson correlation coefficients) from confirmatory factor analyses. These coefficients reflect the degree of association between latent constructs that have been adjusted for differences in measurement reliability. Cronbach's α s for unit-weighted factor scores on the criterion measures were as follows: Physical Assault (.64), Verbal Hostility (.63), Anger Arousal (.82), and Cook–Medley Hostility (.86). All ϕ s are statistically significant at $p < .00001$, two-tailed.

tors for the criterion measure of Hostility. This yielded eight composite indicators of four criterion factors. CFA revealed that the intended four-factor model for the criterion measures (i.e., correlated factors of Physical Assault, Verbal Hostility, Anger Arousal, and Global Hostility) was an acceptable measurement model for the eight composite indicators, $\chi^2(14, n = 180) = 19.2, p > .16$, GFI = .97, RMSEA = .0454, CFI = .99, NNFI = .98.

Refined AQ Factors. We next analyzed the 12 items constituting the refined four-factor model for the AQ together with the eight criterion measures and examined the factor intercorrelations. Table 6 presents the correlations between the four refined AQ factors and the four criterion constructs for this CFA model. To assess discriminant validity, we first examined these correlations separately within each factor, conducting an initial omnibus test of the homogeneity of correlations across criterion measures for each column in the table. The hypothesis of equality in correlations across criteria (i.e., no discriminant validity) was rejected for the AQ factors of Physical Aggression, $\Delta\chi^2(3, n = 180) = 22.5, p < .0001$; Anger, $\Delta\chi^2(3, n = 180) = 23.0, p < .0001$; and Hostility, $\Delta\chi^2(3, n = 180) = 50.0, p < .0001$; but not for Verbal Aggression, $\Delta\chi^2(3, n = 180) = 1.1, p > .77$.

Supporting the construct validity of the *Physical Aggression* factor, follow-up CFA tests using equality constraints revealed that levels of this AQ factor were more strongly correlated with the Physical Assault criterion than with the criteria of Verbal Hostility, $\Delta\chi^2(1, n = 180) = 10.0, p < .002$; Anger Arousal, $\Delta\chi^2(1, n = 180) = 6.6, p < .01$; and Global Hostility, $\Delta\chi^2(1, n = 180) = 19.2, p < .0001$. Supporting the construct validity of the *Anger*

factor, levels of this AQ factor were more strongly correlated with the Anger Arousal criterion than with the criteria of Verbal Hostility, $\Delta\chi^2(1, n = 180) = 9.2, p < .025$; Anger Arousal, $\Delta\chi^2(1, n = 180) = 6.8, p < .009$; and Global Hostility, $\Delta\chi^2(1, n = 180) = 20.7, p < .0001$. And supporting the construct validity of the *Hostility* factor, levels of this AQ factor were more strongly correlated with the Global Hostility criterion than with the criteria of Physical Aggression, $\Delta\chi^2(1, n = 180) = 13.3, p < .003$; Verbal Aggression, $\Delta\chi^2(1, n = 180) = 14.0, p < .0002$; and Anger Arousal, $\Delta\chi^2(1, n = 180) = 6.0, p < .015$. Thus, three of the four refined AQ factors showed evidence of convergent and discriminant validity.²

Original AQ Factors. We next analyzed the 29 items constituting Buss and Perry's (1992) original four-factor measurement model together with the eight criterion measures and examined the factor intercorrelations. Table 6 also presents the correlations between the four original AQ factors and the four criterion constructs for this CFA model. As with the refined measurement model, we first examined these correlations separately within factors, conducting an initial omnibus test of the homogeneity of correlations for each column in the table. Paralleling results for the refined AQ factors, the hypothesis of equality in correlations across criteria (i.e., no discriminant validity) was rejected for the original AQ factors of Physical Aggression, $\Delta\chi^2(3, n = 180) = 27.4, p < .0001$; Anger, $\Delta\chi^2(3, n = 180) = 31.1, p < .0001$; and Hostility, $\Delta\chi^2(3, n = 180) = 50.6, p < .0001$; but not for Verbal Aggression, $\Delta\chi^2(3, n = 180) = 1.5, p > .68$. Thus, in both refined and original forms, Verbal Aggression lacked discriminant validity.

Supporting the construct validity of the original *Physical Aggression* factor, follow-up CFA tests using equality constraints revealed that levels of

² We also assessed the discriminant validity of unit-weighted factor scores for the refined AQ model in terms of their predictive utility in distinguishing males and females. Buss and Perry (1992, p. 455) originally reported that males had higher scores than females on Physical Aggression (PA), Verbal Aggression (VA), and Hostility (HO), but not on Anger (ANG). Multivariate analyses of variance revealed a significant multivariate main effect of gender for Sample 1, $F(4, 299) = 21.1, p < .0001$; Sample 2, $F(4, 195) = 4.3, p < .002$; Sample 4, $F(4, 166) = 6.5, p < .0001$; and Sample 5, $F(4, 165) = 8.9, p < .0001$. (We could not test for gender differences in factor means for Sample 3 because only the covariance matrix was available for reanalysis of these data.) Univariate ANOVAs revealed that males scored higher than females on (a) PA in American Samples 1, 4, and 5, all $ps < .0001$, effect sizes $d = .85, .54$, and 1.0 , respectively, but not in British Sample 2, $p > .60, d = .07$; (b) VA in Samples 1 and 4, $ps < .02, ds = .45$ and $.42$, respectively, but not in Samples 2 and 5, $ps > .17, ds = .11$ and $.23$, respectively; (c) ANG in Sample 5, $p < .02, d = .58$, but not in Samples 1, 2, and 4, $ps > .37, ds = .15, .12$, and $.03$, respectively; and (d) HO in Samples 1 and 5, $ps < .02, ds = .35$ and $.46$, respectively, but not in Samples 2 and 4, $ps > .07, ds = .25$ and $.22$, respectively. Though somewhat inconsistent across samples, these results generally converge with those of previous researchers and provide at least partial support for the discriminant validity of the refined AQ factors.

this AQ factor were more strongly correlated with the Physical Assault criterion than with the criteria of Verbal Hostility, $\Delta\chi^2(1, n = 180) = 11.7, p < .0007$; Anger Arousal, $\Delta\chi^2(1, n = 180) = 11.9, p < .0006$; and Global Hostility, $\Delta\chi^2(1, n = 180) = 26.1, p < .0001$. Supporting the construct validity of the *Anger* factor, levels of this AQ factor were more strongly correlated with the Anger Arousal criterion than with the criteria of Verbal Hostility, $\Delta\chi^2(1, n = 180) = 8.7, p < .004$; Anger Arousal, $\Delta\chi^2(1, n = 180) = 8.3, p < .004$; and Global Hostility, $\Delta\chi^2(1, n = 180) = 30.2, p < .0001$. And partially supporting the construct validity of the *Hostility* factor, levels of this AQ factor were more strongly correlated with the Global Hostility criterion than with the criteria of Physical Assault, $\Delta\chi^2(1, n = 180) = 20.0, p < .0001$, and Verbal Hostility, $\Delta\chi^2(1, n = 180) = 24.7, p < .0001$. However, unlike the refined Hostility factor, Buss and Perry's original Hostility factor correlated equally with both the Global Hostility and Anger Arousal criteria, $\Delta\chi^2(1, n = 180) = 0.7, p > .41$. This result is consistent with earlier findings of a high degree of overlap between these two dimensions (Harris, 1997). Thus, only two of the four original AQ factors showed evidence of convergent and discriminant validity.

Considered together, what do these results tell us about the refined measurement model? First, the construct validity of the refined AQ factors appears to be as good or better than the construct validity of the original Buss–Perry factors. For both models, the Physical Aggression and Anger factors showed the strongest evidence of convergent and discriminant validity, whereas the Verbal Aggression factor showed the weakest. The Hostility factor, in contrast, had stronger discriminant validity in its refined form than in its original form. Taken as a whole, these findings support the construct validity of the refined measurement model for the AQ, and they suggest that the process of modifying the original factors has essentially preserved their conceptual content.

Stage Four: Comparing the Original and Short Forms of the AQ

Our final research objective was to readminister the refined subset of 12 AQ items as a shortened version of the AQ and to evaluate the degree to which the four-factor measurement model replicated in this new form (cf. Yarnold, Bryant, & Grimm, 1987). Accordingly, we analyzed the responses of Samples 4 and 5 to this new, short form of the AQ. Results revealed that the refined four-factor model replicated in both samples. Specifically, the intended four-factor model provided an acceptable measurement model for the 12 items in both Sample 4, $\chi^2(48, n = 171) = 116.94, \text{GFI} = .90, \text{RMSEA} = .092, \text{CFI} = .92, \text{NNFI} = .90$; and Sample 5, $\chi^2(1, n = 170) = 119.14, \text{GFI} = .90, \text{RMSEA} = .094, \text{CFI} = .91, \text{NNFI} = .88$. The refined hierarchical also provided an acceptable measurement model for the 12 items in both Sample 4, $\chi^2(50, n = 171) = 118.1, \text{GFI} = .89, \text{RMSEA} = .090$,

CFI = .92, IFI = .93, NNFI = .90; and Sample 5, $\chi^2(50, n = 170) = 119.4$, GFI = .90, RMSEA = .091, CFI = .92, NNFI = .89.

As a measure of each factor's internal consistency, we also computed Cronbach's alphas for unit-weighted scores on the refined AQ factors. For both Samples 4 and 5, respectively, reliability coefficients were generally acceptable for the refined factors of Physical Aggression (.79 and .80), Verbal Aggression (.83 and .80), Anger (.76 and .76), and Hostility (.75 and .70). These reliabilities are generally comparable to those for Samples 1 and 2, which used the longer, original form of the AQ (see Table 5).

We also used multigroup CFA to assess the generalizability of the refined measurement model across Samples 4 and 5 and across all five samples. Confirming replicability, the refined four-factor model produced equivalent loadings, $\Delta\chi^2(8, n = 341) = 4.7, p > .78$, and equivalent factor variances–covariances, $\Delta\chi^2(10, n = 341) = 5.8, p > .82$, for Samples 4 and 5. Analyzing the data of all five samples simultaneously, we found that the refined four-factor model produced (a) fully invariant loadings in Samples 1, 2, 4, and 5; and (b) 11 of 12 invariant loadings in Sample 3, $\Delta\chi^2(46, n = 1154) = 24.9, p > .99$. In addition, the overarching higher-order Aggression factor had the same relationships with the first-order AQ factors across Samples 1, 2, 4, and 5, $\Delta\chi^2(12, n = 848) = 14.5, p > .26$. Thus, the refined measurement model replicates with the shortened version of the AQ.

As a strong test of the comparability of the factors represented in the short and long forms of the AQ, we reran the analyses for Samples 4 and 5 fixing the factor intercorrelations in the model to the exact values found earlier for our American Sample 1 (see Table 5). Using the factor correlations from the earlier sample did not significantly worsen the fit of the model for either Sample 4, $\Delta\chi^2(6, n = 171) = 12.4, p > .05$, or Sample 5, $\Delta\chi^2(6, n = 170) = 11.8, p > .06$. Thus, shortening the AQ to 12 items does not appear to change the conceptual meaning of the underlying aggression subtraits.

GENERAL DISCUSSION

The most important finding of the present study is the discovery of an appropriate measurement model for the Aggression Questionnaire. Although Buss and Perry's (1992) original four-factor model has a strong theoretical foundation, it explains too little common variance in the 29 AQ items to serve as a formal measurement model. The refined model, in contrast, preserves the conceptual content of the original model, but improves its statistical precision. That the refined model replicates with both the original and short versions of the AQ, is equally applicable for both males and females, and holds across independent samples from three different countries (i.e., United States, England, and Canada) increases our confidence in the external validity of this measurement model.

Our data also provide strong support for the convergent and discriminant

validity of the refined Physical Aggression, Anger, and Hostility factors. Each of these factors correlated more strongly with the criterion measure to which it is presumed to correspond than to the other criterion measures. The Verbal Aggression factor, in contrast, showed no discriminant validity in either its original or refined forms, and was at least moderately correlated with each of the other AQ factors. Clearly, future work is needed to establish the construct validity of this particular subtrait.

Our data further demonstrate that the refined Hostility factor has greater discriminant validity than the original Hostility factor. The latter correlated equally with scores on both the Cook–Medley Hostility Scale and the MAI Anger Arousal subscale, whereas the former correlated more strongly with Cook–Medley scores than with the other criterion measures. It is instructive to note that Buss and Perry (1992) originally defined hostility as the cognitive component of aggression, consisting of “feelings of ill will and injustice” (p. 457). By this definition, hostility is a cognitive dimension that includes negative feelings. That the original Hostility factor had equivalent correlations with the hostility and anger criteria is entirely consistent with this conceptualization. The refined Hostility factor, in contrast, excludes items tapping jealousy (item 22), paranoia (item 28), and suspiciousness (items 27 and 29). Omitting these items appears to disentangle Hostility from Anger, giving the former a more exclusively cognitive focus.

What measurement model should researchers adopt in using the AQ, and how should they best score this instrument? Our findings clearly indicate that dispositional aggression (as measured by the AQ) is multidimensional. Thus, examining the four separate subtraits is more reasonable both conceptually and statistically than relying solely on a pooled “total score.” Indeed, by collapsing across multiple dimensions, AQ total score could conceivably obscure differences that exist for individual factors. This suggests that it would be unwise for researchers to use only the 29-item total score in quantifying responses to the AQ.

Our data also demonstrate that the refined four-factor model is superior to Buss and Perry’s (1992) original four-factor model in its overall goodness-of-fit to the data. In other words, the 12 AQ items constituting the refined model more closely reflect the underlying subtraits of physical aggression, verbal aggression, anger, and hostility than do the 29 items constituting the original model. Although the four original factors offer somewhat greater internal consistency in terms of Cronbach’s alpha, this apparent advantage disappears when the reliabilities of the refined factors are corrected for differences in the number of constituent items (adjusted α s = .88 – .92), using the Spearman–Brown prophecy formula (Magnusson, 1967). Thus, the refined model appears to provide a valid and reliable means of measuring the four aggression subtraits, using either the original 29-item AQ or its 12-item short form.

Whether one should adopt the four-factor model or its hierarchical counterpart depends on one's research objectives. The first-order model enables one to examine the relationships among each of the four AQ factors as well as the specific correlates of each subtrait of aggression. The higher-order model, in contrast, enables one to examine the common variance among the four AQ factors as an independent or dependent variable in its own right. In either case, the present study suggests that researchers can enhance both conceptual and predictive precision by using the refined measurement model for the AQ.

Our research conclusions are constrained by several important limitations. First, we used only self-report measures as criteria in assessing construct validity. Clearly, it would be advantageous to include behavioral and physiological criterion measures of physical and verbal aggression, anger, and hostility (cf. Leibowitz, 1968). This would increase confidence that the AQ actually measures these aggression subtraits. In addition, we used only college student samples across the three countries, and our results may not generalize to older adults or offender populations (cf. Williams et al., 1996).

In closing, we note two unresolved measurement issues concerning the AQ. The first issue concerns the nature of the response scale used to quantify AQ responses. Although we changed the original 5-point scale to a 6-point scale in the AQ short form, we preserved the original anchors for the scale endpoints: *extremely uncharacteristic of me* to *extremely characteristic of me*. However, the exact meaning of this scale remains somewhat unclear. Specifically, is an uncharacteristic–characteristic scale unipolar or bipolar (cf. Sudman & Bradburn, 1982)? Might respondents interpret “extremely uncharacteristic” to mean that the particular characteristic in question is so unlike them that its *opposite* actually holds true for them? Such bipolarity would be conceptually inconsistent with traditional theoretical formulations of aggression. In the future, it might be better to anchor the scale with *not at all characteristic of me* to *very much characteristic of me*, so as to avoid conceptual confusion.

A final measurement issue concerns the order of the items constituting the AQ. Buss and Perry (1992) did not report the order of the 29 items in the original AQ, but simply advised researchers to “make up their own . . . version by scrambling the items so that items do not pile up” (footnote 1, p. 453). Accordingly, different users have used different randomized orderings of the AQ items. This approach has drawbacks, however. Though apparently intended to enhance measurement validity, the practice of rerandomizing the order of the AQ items across studies makes it extremely difficult to conduct secondary analyses of existing AQ data sets, as done in the present study. When comparing data across studies, one must not only determine the specific order of the items that each researcher has used, but must also match up items across studies and then rearrange all AQ items in each data

set into one universal order, constructing a complex form of "Rosetta Stone" in the process. For example, AQ item 11 in our Sample 1 is (a) item 21 in Buss and Perry's (1992, p. 454) Table 1; (b) item 22 in Archer, Holloway, and McLoughlin's (1995) data set; (c) item 29 in Harris's (1995) data set; and (d) item 5 in our 12-item short form of the AQ. Rerandomizing the order of AQ items at each administration makes little sense, given the methodological sufficiency of a single random order. To facilitate comparison of results across studies, we strongly urge future researchers to adopt a standard random order of the AQ items (the note for Table 2 reports the order we have used for both the original and short forms of the AQ).

Finally, we alert researchers to the fact that Western Psychological Services (Los Angeles, CA) has recently developed a new 34-item version of the original 29-item AQ (Buss & Warren, 2000). This new instrument contains all but one of the original Buss-Perry items (14 of which have been modified to improve readability), along with a new, 6-item Indirect Aggression subscale. Although Indirect Aggression was initially a component of the Buss-Durkee Hostility Inventory (Buss & Durkee, 1957), Buss and Perry (1992) omitted this dimension of aggression in their original AQ. Researchers wishing to use the short form of the AQ should note that the questions in the new AQ have been arranged so that the first 12 items represent the indicators constituting our refined measurement model. Advantages to using this new 34-item AQ (Buss & Warren, 2000), as opposed to the original 29-item AQ (Buss & Perry, 1992), include the availability of (a) more extensive evidence supporting construct validity; (b) large-scale, national norms; (c) a reliable subscale to tap Indirect Aggression; and (d) a single, fixed order for the AQ items.

REFERENCES

- Archer, J., Kilpatrick, G., & Bramwell, R. (1995). Comparison of two aggression inventories. *Aggressive Behavior*, **21**, 371-380.
- Archer, J., Holloway, R., & McLoughlin, K. (1995). Self-reported physical aggression among young men. *Aggressive Behavior*, **21**, 325-342.
- Bagozzi, R. P. (1993). Assessing construct validity in personality research: Applications to measures of self-esteem. *Journal of Research in Personality*, **27**, 49-87.
- Bagozzi, R. P., & Edwards, J. R. (1998). A general approach for representing constructs in organizational research. *Organizational Research Methods*, **1**, 45-87.
- Bagozzi, R. P., & Heatherton, T. F. (1994). A general approach to representing multifaceted personality constructs: Application to state self-esteem. *Structural Equation Modeling*, **1**, 35-67.
- Barefoot, J. C., Dahlstrom, W. G., & Williams, R. B., Jr. (1983). Hostility, CHD incidence and total mortality: A 25-year follow-up study of 255 physicians. *Psychosomatic Medicine*, **45**, 59-63.
- Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, **107**, 238-246.

- Bentler P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, **88**, 588–606.
- Bernstein, I. A., & Gesn, P. R. (1997). On the dimensionality of the Buss/Perry Aggression Questionnaire. *Behavior Research and Therapy*, **35**, 563–568.
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.
- Browne, M. W., & Cudeck, R. (1989). Single-sample cross-validation indices for covariance structures. *Multivariate Behavioral Research*, **24**, 445–455.
- Bryant, F. B., & Baxter, W. J. (1997). The structure of positive and negative automatic cognition. *Cognition and Emotion*, **11**, 225–258.
- Bryant, F. B., & Yarnold, P. R. (1995). Principal-components analysis and exploratory and confirmatory factor analysis. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding multivariate statistics* (pp. 99–136). Washington, DC: American Psychological Association.
- Bryant, F. B., Yarnold, P. R., & Grimm, L. G. (1996). Toward a measurement model of the Affect Intensity Measure: A three-factor structure. *Journal of Research in Personality*, **30**, 223–247.
- Byrne, B. M., Shavelson, R. J., & Muthen, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, **105**, 456–466.
- Buss, A. H. (1961). *The psychology of aggression*. New York: Wiley.
- Buss, A. H., & Durkee, A. (1957). An inventory for assessing different kinds of hostility. *Journal of Consulting Psychology*, **21**, 343–349.
- Buss, A. H., Fischer, H., & Simmons, A. J. (1962). Aggression and hostility in psychiatric patients. *Journal of Consulting*, **26**, 84–89.
- Buss, A. H., & Perry, M. (1992). The Aggression Questionnaire. *Journal of Personality and Social Psychology*, **63**, 452–459.
- Buss, A. H., & Warren, W. L. (2000). *The Aggression Questionnaire manual*. Los Angeles: Western Psychological Services.
- Cook, W. W., & Medley, D. M. (1954). Proposed Hostility and Pharisaic-Virtue scales for the MMPI. *Journal of Applied Psychology*, **38**, 414–418.
- Felsten, G., & Hill, V. (1998). Aggression Questionnaire hostility scale predicts anger in response to mistreatment. *Behavior Research and Therapy*, **37**, 87–97.
- Harburg, E., Erfurt, J. C., Hauenstein, L. S., Chape, C., Schull, W. J., & Schork, M. A. (1973). Socio-ecological stress, suppressed hostility, skin color, and black-white male blood pressure. *Psychosomatic Medicine*, **35**, 276–296.
- Harris, J. A. (1995). Confirmatory factor analysis of the Aggression Questionnaire. *Behavior Research and Therapy*, **33**, 991–993.
- Harris, J. A. (1997). A further evaluation of the Aggression Questionnaire: Issues of validity and reliability. *Behavior Research and Therapy*, **35**, 1047–1053.
- Hoelter, J. W. (1983). The analysis of covariance structures: Goodness-of-fit indices. *Sociological Methods and Research*, **11**, 325–344.
- Hu, L.-T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, **3**, 424–453.
- Hull, J. G., Lehn, D. A., & Tedlie, J. C. (1991). A general approach to testing multifaceted personality constructs. *Journal of Personality and Social Psychology*, **61**, 932–945.
- Joreskog, K. G., & Sorbom, D. G. (1996). *LISREL 8: User's reference guide*. Chicago: Scientific Software.

- Judd, C. M., Jessor, R., & Donovan, J. E. (1986). Structural equation models and personality research. *Journal of Personality*, **54**, 149–198.
- Kline, R. B. (1998). *Principles and practice of structural equation modeling*. New York: Guilford.
- Leibowitz, G. (1968). Comparison of self-report and behavioral techniques of assessing aggression. *Journal of Consulting and Clinical Psychology*, **32**, 21–25.
- MacCallum, R. C. (1986). Specification searches in covariance structure modeling. *Psychological Bulletin*, **100**, 107–120.
- MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modification in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, **111**, 490–504.
- Magnusson, D. (1967). *Test theory*. Reading, MA: Addison–Wesley.
- Maiuro, R. D., Cahn, T. S., Vitaliano, P. P., Wagner, B. C., & Zegree, J. B. (1988). Anger, hostility, and depression in domestically violent versus generally assaultive men and non-violent control subjects. *Journal of Consulting and Clinical Psychology*, **56**, 17–23.
- Marsh, H. W., Balla, J. R., & Hau, K.-T. (1996). An evaluation of incremental fit indices: A clarification of mathematical and empirical properties. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 315–353). Mahwah, NJ: Erlbaum.
- Maruyama, G. M. (1998). *Basics of structural equation modeling*. Thousand Oaks, CA: Sage.
- McDonald, R. P., & Marsh, H. W. (1990). Choosing a multivariate model: Noncentrality and goodness of fit. *Psychological Bulletin*, **107**, 247–255.
- Meesters, C., Muris, P., Bosma, H., Schouten, E., & Beuving, S. (1996). Psychometric evaluation of the Dutch version of the Aggression Questionnaire. *Behavior Research and Therapy*, **34**, 839–843.
- Nichols, J. G., Licht, B. G., & Pearl, R. A. (1982). Some dangers of using personality questionnaires to study personality. *Psychological Bulletin*, **92**, 572–580.
- Novaco, R. (1975). *Anger control: The development and evaluation of an experimental treatment*. Lexington, MA: Lexington Books.
- Shekelle, R. B., Gale, M., Ostfeld, A. M., & Paul, O. (1983). Hostility, risk of coronary heart disease, and mortality. *Psychosomatic Medicine*, **45**, 109–114.
- Siegel, J. M. (1986). The Multidimensional Anger Inventory. *Journal of Personality and Social Psychology*, **51**, 191–200.
- Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, **25**, 173–180.
- Sudman, S., & Bradburn, N. M. (1982). *Asking questions: A practical guide to questionnaire construction*. San Francisco: Jossey–Bass.
- Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, **38**, 1–10.
- Velicer, W. F., Gova, J. M., Cherico, N. P., & Corriveau, D. P. (1985). Item format and the structure of the Buss–Durkee Hostility Inventory. *Aggressive Behavior*, **11**, 65–82.
- Williams, T. W., Boyd, J. C., Cascardi, M. A., & Poythress, N. (1996). Factor structure and convergent validity of the Aggression Questionnaire in an offender population. *Psychological Assessment*, **8**, 398–403.
- Yarnold, P. R., Bryant, F. B., & Grimm, L. G. (1987). Comparing the long and short forms of the Student Jenkins Activity Survey. *Journal of Behavioral Medicine*, **10**, 75–90.
- Zillmann, D. (1979). *Hostility and aggression*. Hillsdale, NJ: Earlbaum.