AGGRESSIVE BEHAVIOR
Volume 40, pages 120–139 (2014)

The Brief Aggression Questionnaire: Psychometric and Behavioral Evidence for an Efficient Measure of Trait Aggression

Gregory D. Webster[1]*, C. Nathan DeWall[2], Richard S. Pond Jr.[3], Timothy Deckman[2], Peter K. Jonason[4], Bonnie M. Le[5], Austin Lee Nichols[6], Tatiana Orozco Schember[1], Laura C. Crysel[1], Benjamin S. Crosier[1], C. Veronica Smith[7], E. Layne Paddock[8], John B. Nezlek[9,10], Lee A. Kirkpatrick[9], Angela D. Bryan[11], and Renée J. Bator[12]

[1]*Department of Psychology, University of Florida, Gainesville, Florida*
[2]*Department of Psychology, University of Kentucky, Lexington, Kentucky*
[3]*Department of Psychology, University of North Carolina at Wilmington, Wilmington, North Carolina*
[4]*Department of Psychology, University of Western Sydney, Sydney, NSW, Australia*
[5]*Department of Psychology, University of Toronto, Toronto, ON, Canada*
[6]*Peking University HSBC Business School Shenzhen, Peking, China*
[7]*Department of Psychology, University of Mississippi, Oxford, Mississippi*
[8]*Lee Kong Chian School of Business, Singapore Management University, Singapore, Singapore*
[9]*Department of Psychology, College of William and Mary, Williamsburg, Virginia*
[10]*University of Social Sciences and Humanities, Poznań, Poland*
[11]*Department of Psychology and Neuroscience, University of Colorado at Boulder, Boulder, Colorado*
[12]*Department of Psychology, State University of New York at Plattsburgh, Plattsburgh, New York*

A key problem facing aggression research is how to measure individual differences in aggression accurately and efficiently without sacrificing reliability or validity. Researchers are increasingly demanding brief measures of aggression for use in applied settings, field studies, pretest screening, longitudinal, and daily diary studies. The authors selected the three highest loading items from each of the Aggression Questionnaire's (Buss & Perry, 1992) four subscales—physical aggression, verbal aggression, anger, and hostility—and developed an efficient 12-item measure of aggression—the Brief Aggression Questionnaire (BAQ). Across five studies ($N = 3{,}996$), the BAQ showed theoretically consistent patterns of convergent and discriminant validity with other self-report measures, consistent four-factor structures using factor analyses, adequate recovery of information using item response theory methods, stable test–retest reliability, and convergent validity with behavioral measures of aggression. The authors discuss the reliability, validity, and efficiency of the BAQ, along with its many potential applications. Aggr. Behav. 40:120–139, 2014.

Keywords: anger; aggression; Aggression Questionnaire; hostility; item response theory; measurement; short form

INTRODUCTION

Aggression researchers face an increasing need for efficient measures. From daily diary studies to longitudinal and field research, brief measures of popular psychological constructs are increasingly in high demand (Widaman, Little, Preacher, & Sawalani, 2011). Several research contexts exist that require researchers to accurately and efficiently assess individual differences in personality, emotion, and cognition, particularly when participants need to be assessed frequently and quickly, or are susceptible to fatigue. This need for efficient

Different analyses of some of the data from three of the six independent samples included in this article appear in published sources: Study 1 (Webster & Kirkpatrick, 2006; Webster, Kirkpatrick, Nezlek, Smith, & Paddock, 2007, Study 3); Study 2, Sample 2 (Webster & Bryan, 2007); and Study 3 (Webster, 2006, 2007; Webster & Bryan, 2007). Aside from descriptive statistics (*M*s, *SD*s, αs), the present results do not reproduce previously published results.

*Correspondence to: Gregory D. Webster, Department of Psychology, University of Florida, P.O. Box 112250, Gainesville, FL 32611-2250. E-mail: gdwebs@gmail.com

Received 26 December 2012; Accepted 8 August 2013

DOI: 10.1002/ab.21507
Published online 1 October 2013 in Wiley Online Library (wileyonlinelibrary.com).

measures has resulted in the validation of brief measures for some of the most widely used scales in psychology.

For example, in the Big Five personality factors, often measured with the 44-item Big-Five Inventory (BFI; John & Srivastava, 1999), and has inspired several brief measures such as the Ten-Item Personality Inventory (Gosling, Rentfrow, & Swann, 2003), the 10-item BFI (BFI-10; Rammstedt & John, 2007), and the 20-item Mini International Personality Item Pool–Five-Factor Model (Donnellan, Oswald, Baird, & Lucas, 2009). Similarly, the 40-item Narcissistic Personality Inventory (Raskin & Terry, 1998) has been reduced to a 16-item version (Ames, Rose, & Anderson, 2005), and the 19-item Impulsivity and Sensation Seeking scale (ImpSS; Zuckerman, Kuhlman, Joireman, Teta, & Kraft, 1993) has been reduced to eight items (ImpSS-8; Webster & Crysel, 2012). The Dark Triad—narcissism, psychopathy, and Machiavellianism—was reduced from three measures and 91 items to a single, 12-item scale (the "Dirty Dozen"; Jonason & Webster, 2010; Webster & Jonason, 2013; but also see Miller et al., 2012). There are even single-item measures of self-esteem (Robins, Hendin, & Trzesniewski, 2001) and need to belong (Nichols & Webster, 2013) that are valid measures of their respective 10-item scales. Thus, efficient measures are increasingly in demand.

But what about aggression? Aggression—behavior intended to harm other people who want to avoid the harm (Bushman & Huesmann, 2010) is a highly researched topic that transcends traditional academic boundaries. It is studied not only by anthropologists, sociologists, criminologists, and communication scientists but also by social, personality, developmental, and clinical psychologists. More specifically, *trait aggression* describes individual differences in thoughts (e.g., hostility), emotions (e.g., anger), and behavior (e.g., verbal and Physical Aggression) that are intended to harm another person. Regarding the Big Five personality traits, trait aggression often relates positively with neuroticism, inconsistently with extraversion, and negatively with agreeableness, openness, and conscientiousness (see Barlett & Anderson, 2012, for a brief review).

One challenge in studying this interdisciplinary phenomenon is accurately measuring individual differences in trait aggression. The most popular way to assess individual differences in aggression has been through self-report measures—a measurement technique with a long history. Below we provide a short history of self-report aggression measures and discuss the need for a new brief measure that produces valid, reliable scores. We then offer converging evidence from five studies showing the validity and reliability of the Brief Aggression Questionnaire (BAQ) and discuss its implications for aggression research.

Self-Report Measures of Aggression: A Brief History

One of the first and most widely used measures of aggression is the Buss–Durkee Hostility Inventory (BDHI; Buss & Durkee, 1957). The BDHI consists of 66 dichotomously scored true–false items that assess hostility across seven subscales (Assault, Indirect, Irritability, Negativism, Resentment, Suspicion, and Verbal; it also includes a nine-item Guilt subscale unrelated to the hostility items). When created, the BDHI was a combination of new, author-generated items and items taken from established scales that related to one of the measure's seven subscales. These 105 initial items were then administered to a sample of 159 college students, after which the authors used frequency (items that were neither endorsed too often or too rarely) and internal consistency to identify the items that best assessed hostility.

Despite the BDHI's success, its seven-subscale approach to measuring hostility seemed unnecessarily complex. Even Buss and Durkee (1957, pp. 347–348; see also Bushman, Cooper, & Lemke, 1991; Kernis, Grannemann, & Barclay, 1989) showed that these seven subscales could be reduced to two factors: aggressiveness (assault, indirect aggression, irritability, and Verbal Aggression) and Hostility (resentment and suspicion). To address these problems, Buss and Perry (1992) streamlined the BDHI to create its successor—the Buss–Perry Aggression Questionnaire (BPAQ)—by jettisoning several items and updating several more to form the 29-item BPAQ. Unlike its seven-factor predecessor, the BPAQ focused on four facets of aggression: Physical Aggression, Verbal Aggression, anger, and hostility. Across multiple studies, Buss and Perry found consistent support for a four-factor structure. In addition to being more efficient than the BDHI (66 items vs. 29 items—a 56% decrease), the BPAQ featured improved psychometric properties including higher internal consistency reliability.

Despite the BPAQ's acclaim and wide acceptance, with 29 items, it remains too long for research studies that require brevity. With the advent of mobile and web-based technologies, researchers sought more efficient measures that could be used for experience-sampling studies, longitudinal studies, special populations, participant pool prescreening, and mass testing (e.g., Amazon's Mechanical Turk; see Buhrmester, Kwang, & Gosling, 2011), To this end, Bryant and Smith (2001) developed a short form of the BPAQ (BPAQ-SF). In this present research, we develop and evaluate an alternative short form of the BPAQ called the BAQ.

Why Another Self-Report Measure of Aggression?

Constructing brief measures requires methodological and psychometric considerations to ensure the reliability and validity of test scores. One way to create a brief measure with reliable and valid scores is to take the "best" items from a longer "parent" measure. Identifying the best items requires considering multiple criteria including item-total correlations, factor loadings from principal axis factoring (PAF) or confirmatory factor analyses (CFAs), and the wording and face validity of items. The BPAQ-SF's 12 items were selected based on Bryant and Smith's (2001) preliminary study of 307 U.S. undergraduates; however, they largely replicated its four-factor structure in four other undergraduate samples from the United States, the United Kingdom, and Canada (total $N = 1{,}154$). In contrast, the BPAQ sampled 1,253 U.S. undergraduates from the same institution. A strength of the BPAQ is its stable factor loadings for item selection (one large sample), whereas a strength of the BPAQ-SF is its factor-structure generalizability (five moderate samples). Because our initial goal was optimal item selection, and because larger samples provide more stable factor loadings than smaller ones, we chose to use the BPAQ's results to select the "best" items from each of its four subscales for our brief measure, the BAQ.

Whereas the BPAQ included multiple reverse-scored items, the BPAQ-SF has none. Because we designed our measure to mirror the original BPAQ, the BAQ includes a reverse-scored item. Scales including at least one reverse-scored items are potentially advantageous for at least three reasons. First, respondents tend to *agree* with items when they have a *positive* connotation and—more relevant to self-reporting aggression—*disagree* with items when they have a *negative* connotation (i.e., acquiescence bias, Paulhus & Vazire, 2007). Second, respondents tend to answer in ways that they believe will please the investigators or confirm their hypotheses (i.e., positive response bias, demand characteristics, the "good subject effect"; Nichols & Maner, 2008; Orne, 1962). Third, respondents tend to show bias toward acceptable, normative social behavior or expectations (Paulhus, 2002), which is important for self-reporting aggression because it is viewed as socially undesirable, and so respondents may be more reluctant to report how aggressive they actually are. A reverse-scored item, which forces people to report their *lack* of aggression, can help reduce these biases. Thus, we created a measure as efficient and valid as the BPAQ-SF, but with one reverse-scored item.

The Present Research

We propose a new 12-item measure of aggression, the BAQ.[1] In a series of five studies ($N = 3{,}996$), we show that the BAQ gives reliable, valid scores (Studies 1 and 5), replicates the four-factor structure of the BPAQ using factor analyses (Studies 1 and 2), efficiently recovers information using item response theory (IRT; Study 3), has strong test–retest reliability (Study 4), and contains a Physical Aggression subscale that shows convergent validity with behavioral aggression (Study 5). We also show that the BAQ possesses acceptable convergent and discriminant validity with other measures (Studies 1 and 5). We argue that the BAQ produces reliable, valid scores and is efficient and practical to use in multiple settings.

STUDY 1: SCALE DEVELOPMENT AND PRELIMINARY TESTS

Study 1 had two goals. First, we sought to develop a new, brief measure of aggression—the BAQ—and test its hypothesized four-factor structure and its convergent and discriminant validity with both its "parent" (BPAQ) and "grandparent" (BDHI) measures. Second, we sought to compare the convergent and discriminant validity of the BAQ and the BPAQ-SF.

Method

Participants. Participants were 109 (55 men, 54 women; $M_{age} = 20.0$ years, $SD = 1.2$) introductory psychology students at a public university in Virginia who received course credit for participating in an online study.[2]

Measures and procedure. Aggression was measured using the 29-item version of the BPAQ (Buss & Perry, 1992). The BPAQ contains both the 12 items that constitute the BPAQ-SF (Bryant & Smith, 2001) and our 12-item measure, the BAQ. For the BAQ, however, we used Buss and Perry's (1992, p. 459) factor analytic results of their combined sample of 1,253 participants to choose the three highest-loading items within each of their four subscales (Table I, "rank" column)—a method that benefited from the increased reliability of their large sample. As shown in Table I, only half (6 of 12) of these items overlapped with those of the BPAQ-SF. Within each subscale, the BPAQ-SF and the BAQ shared only one or two items—never zero or three. We used a response scale ranging from 1 (*extremely uncharacteristic of me*) to 5 (*extremely characteristic of me*).

[1] Although the Brief Aggression Questionnaire has neither been published previously nor undergone formal psychometric evaluation, it has been used effectively in prior published studies (Jonason & Webster, 2010; Webster, 2006, 2007; Webster & Bryan, 2007; Webster & Crysel, 2012; Webster et al., 2007).

[2] We did not collect or could not locate ethnicity data for this sample.

TABLE I. The 29-Item Aggression Questionnaire (Buss & Perry, 1992), Its 12-Item Short Form (Bryant & Smith, 2001; italicized items) and the 12-Item Brief Aggression Questionnaire (boldface items)

	Loading	Rank
Physical aggression		
1. Once in a while I can't control the urge to strike another person.[a]	0.61	
2. Given enough provocation, I may hit another person.	0.84	1
3. If someone hits me, I hit back.[a]	0.64	
4. I get into fights a little more than the average person.[a]	0.51	
5. If I have to resort to violence to protect my rights, I will.[a]	0.65	2.5
6. There are people who pushed me so far that we came to blows.[a]	0.65	2.5
7. I can think of no good reason for hurting another person.*[a]	0.63	
8. I have threatened people I know.	0.52	
9. I have become so mad that I have broken things.	0.52	
Verbal aggression		
1. I tell my friends openly when I disagree with them.	0.46	2.5
2. I often find myself disagreeing with people.[v]	0.40	
3. When people annoy me, I may tell them what I think of them.[v]	0.46	2.5
4. I can't help getting into arguments when people disagree with me.[v]	0.38	
5. My friends say that I'm somewhat argumentative.	0.51	1
Anger		
1. I flare up quickly but get over it quickly.	0.51	
2. When frustrated, I let my irritation show.	0.44	
3. I sometimes feel like a powder keg ready to explode.[i]	0.43	
4. I am an even-tempered person.*	0.65	3
5. Some of my friends think I'm a hothead.	0.61	
6. Sometimes I fly off the handle for no good reason.	0.71	2
7. I have trouble controlling my temper.	0.72	1
Hostility		
1. I am sometimes eaten up with jealousy.[r]	0.43	
2. At times I feel I have gotten a raw deal out of life.[r]	0.55	
3. Other people always seem to get the breaks.[r]	0.61	2
4. I wonder why sometimes I feel so bitter about things.	0.50	
5. I know that "friends" talk about me behind my back.[s]	0.48	
6. I am suspicious of overly friendly strangers.[s]	0.44	
7. I sometimes feel that people are laughing at me behind my back.[s]	0.65	1
8. When people are especially nice, I wonder what they want.[s]	0.56	3

Notes. *Reverse-scored item. Items developed from the [a]Assault, [v]Verbal Aggression, [i]Irritability, [r]Resentment, or [s]Suspicion subscales of Buss–Durkee (1957) Hostility Inventory.

Among the BAQ items that differed from those of the BPAQ-SF, those selected for the BAQ appeared to be more face-valid measures of their respective constructs. For example, because threats are not—strictly speaking—Physical Aggression, the BAQ's item "If I have to resort to violence to protect my rights, I will" is arguably a more face-valid measure of Physical Aggression than the BPAQ-SF's "I have threatened people I know." Similarly, of the five BPAQ Verbal Aggression items, only one does *not* describe a verbal interaction between two people; "I often find myself disagreeing with people" could just as easily describe a thought, attitude, or personality trait (disagreeableness) rather than Verbal Aggression per se. The BPAQ-SF uses this item; the BAQ does not. Regarding anger, the BPAQ-SF uses the ambiguous item "I flare up quickly but get over it quickly," which appears to relate more closely to anger management or self-regulation than anger itself; in its place, the BAQ uses "I am an even-tempered person" (reverse-scored). For the Hostility items, only one of the BPAQ-SF's items describes hostility directed toward others; for the other two items, the target of hostility is ambiguous. In contrast, all three of the BAQ's hostility items measure Hostility directed toward other people (Table I).

To assess convergent and discriminant validity of our new measure, we also measured the BPAQ's predecessor, the BDHI (Buss & Durkee, 1957), both because it remains highly cited today and because it provides multiple facets of hostility that should relate differentially to the BPAQ's four subscales. The BDHI includes 66 items with a dichotomous, true–false response format and eight subscales: Assault, Indirect hostility, Irritability, Negativity, Resentment, Suspicion, Verbal hostility, and Guilt (scored separately). Items included, "If somebody hits me first, I let them have it" (Assault) and "Almost every week I see someone I dislike" (Resentment). Total and subscale scores were computed as the mean of all "hostile" responses with higher numbers reflecting greater hostility (range: 0–1).

TABLE II. Studies 1 and 5 (Above and Below the Diagonal, Respectively): Statistics and Correlations for Three Aggression Questionnaire Scales

	BPAQ					BPAQ-SF					BAQ					Study 1		
Scale	P	V	A	H	T	P	V	A	H	T	P	V	A	H	T	*M*	*SD*	α
BPAQ																		
Physical	—	.49	.49	.27	.78	**.92**	.41	.44	.27	.72	**.93**	.46	.47	.19	.76	2.17	0.85	.87
Verbal	.48	—	.62	.29	.72	.49	**.92**	.44	.18	.71	.45	**.93**	.50	.25	.76	2.96	0.86	.79
Anger	.49	.53	—	.60	.85	.46	.66	**.91**	.48	.85	.35	.52	**.88**	.48	.77	2.19	0.73	.80
Hostility	.34	.33	.51	—	.72	.26	.31	.55	**.88**	.68	.17	.21	.58	**.94**	.65	2.41	0.83	.85
Total	.81	.71	.81	.73	—	.73	.69	.75	.61	**.96**	.66	.64	.78	.62	**.96**	2.38	0.63	.91
BPAQ-SF																		
Physical	**.92**	.47	.46	.34	.76	—	.42	.40	.25	.74	**.90**	.44	.45	.20	.73	1.93	1.05	.84
Verbal	.44	**.91**	.54	.37	.68	.41	—	.48	.17	.73	.35	**.74**	.52	.28	.66	2.66	1.01	.82
Anger	.50	.50	**.93**	.50	.78	.46	.50	—	.44	.78	.30	.35	**.86**	.44	.67	2.19	0.87	.67
Hostility	.34	.37	.49	**.87**	.68	.34	.41	.49	—	.64	.19	.30	.49	**.82**	.56	2.39	0.97	.73
Total	.73	.74	.79	.68	**.96**	.74	.76	.80	.74	—	.62	.19	.79	.59	**.91**	2.29	0.70	.83
BAQ																		
Physical	**.94**	.44	.41	.27	.72	**.92**	.37	.43	.27	.67	—	.62	.34	.12	.71	2.23	1.06	.78
Verbal	.45	**.95**	.47	.24	.63	.45	**.75**	.44	.28	.63	.43	—	.42	.18	.73	3.06	0.93	.67
Anger	.39	.37	**.87**	.42	.65	.34	.39	**.79**	.41	.63	.28	.31	—	.48	.77	1.93	0.84	.79
Hostility	.34	.27	.45	**.90**	.65	.34	.30	.45	**.75**	.61	.27	.19	.36	—	.61	2.30	0.89	.65
Total	.81	.72	.77	.63	**.96**	.79	.64	.74	.59	**.91**	.78	.69	.66	.63	—	2.38	0.66	.81
Study 5																		
Mean	2.70	3.27	2.51	2.42	2.68	2.26	2.74	2.34	2.33	2.42	2.75	3.56	2.31	2.36	2.74			
SD	1.22	1.14	1.07	1.12	0.88	1.42	1.31	1.25	1.34	1.01	1.65	1.24	1.16	1.18	0.91			
α	.87	.78	.81	.85	.91	.76	.81	.72	.80	.86	.83	.62	.67	.65	.79			

Notes. Study 1: $N = 109$, rs $\geq .19$ are $P < .05$. Study 5: $N = 307$, all rs are $P < .05$.
BPAQ, Buss–Perry Aggression Questionnaire; SF, Short Form; BAQ, Brief Aggression Questionnaire; Boldface, convergent validity correlations.

Results and Discussion

Preliminary analyses. Descriptive statistics for—and correlations among—all three scales appear in Table II (above diagonal). Note that lower scale reliabilities (αs) are to be expected for three-item scales because alphas are a function of not only the average inter-item correlation but also the number of items (i.e., scale length; Schmitt, 1996). For example, a scale with a mean inter-item correlation of .30 would have an alpha of .81 with 10 items, but only .56 with three items.

"Parent–child" validity. We compared the extent to which the two brief measures (BPAQ-SF, BAQ) accurately recovered the scores from the full version of their "parent" measure, the BPAQ (Table II, above diagonal). If our three-item BAQ subscales more accurately represented the four BPAQ dimensions than the three-item BPAQ-SF subscales, then the convergent validity correlations should be stronger for the BAQ than the BPAQ-SF, and the off-diagonal, discriminant validity correlations should be weaker for the BAQ than the BPAQ-SF. As expected, the convergent validity correlations were slightly stronger for the BAQ than those of the BPAQ-SF, but not significantly. Also as expected, off-diagonal correlations among the abbreviated and full versions of the four subscales were lower for the BAQ's three-items subscales than those of the BPAQ-SF, but not significantly.

Principal axis factoring. We next ran PAF analyses with oblique rotation on both the BPAQ-SF and our new BAQ measure (Table A shows pattern matrices; see online Supporting Information). (Tables A–D and Figs. A–C appear in the online Supporting Information document.) Based on theory and the BPAQ's four-factor structure, we specified a four-factor structure for both the BPAQ-SF and the BAQ (Table B shows component correlations; online Supporting Information). Only the BAQ, however, had four factors with eigenvalues ≥ 1.0; the BPAQ-SF's fourth factor had an eigenvalue of only 0.9. Nevertheless, all items loaded primarily on their hypothesized axis for both measures. This analysis suggests that both brief measures can recover the BPAQ's four-factor structure using only 12 items.

Convergent and discriminant validity. We next tested the convergent and discriminant validity of the BAQ and the BPAQ-SF using the seven BDHI subscales as criterion measures (Tables III and IV). The correlation matrix in Table IV is equivalent to the heteromethod block of a multitrait–multimethod matrix (MTMM; Campbell & Fiske, 1959). Both brief measures were correlated with their respective aggression constructs (i.e., strong, positive rs showing convergent validity) and were not correlated with unrelated constructs. Moreover, each brief measure reproduced the pattern of significance in correlations between the 29-item

TABLE III. Study 1: Descriptive Statistics and Correlations for the Buss–Durkee Hostility Inventory

	M	*SD*	α	Items	1	2	3	4	5	6	7	8
1. BDHI total	0.40	0.16	.89	66								
2. Physical assault	0.33	0.24	.76	10	.66							
3. Verbal hostility	0.53	0.22	.72	13	.70	.51						
4. Irritability	0.43	0.23	.68	11	.79	.35	.41					
5. Resentment	0.31	0.23	.57	8	.69	.30	.22	.61				
6. Suspicion	0.27	0.21	.66	10	.65	.27	.28	.46	.56			
7. Indirect hostility	0.50	0.26	.71	9	.74	.32	.44	.56	.45	.37		
8. Negativism	0.37	0.26	.47	5	.50	.27	.23	.35	.29	.20	.37	
9. Guilt	0.49	0.22	.56	9	.21	−.02	.03	.21	.22	.29	.19	.15

Notes. $N = 109$. Correlations ≥.20 in absolute magnitude were significant (Ps < .05, two-tailed).

TABLE IV. Study 1: Partial Multitrait-Multimethod Matrix (MTMM) Showing Convergent (On-Diagonal) and Discriminant (Off-Diagonal) Validity Correlations Between the 29-Item Buss–Perry (1992) Aggression Questionnaire and Its Subscales (Top), and Two Competing 12-Item Versions of the Same with 3-Item Subscales: Bryant and Smith's (2001) Version (Middle) and Our Version (Bottom)

	Total	Physical aggression	Verbal aggression	Anger	Hostility
Aggression Questionnaire (Buss & Perry, 1992)					
BDHI total	**.86**	.63	.59	.75	.66
Physical assault	.68	**.85**	.39	.44	.29
Verbal hostility	.62	.55	**.75**	.49	.19†
Irritability	.69	.35	.48	**.73**	**.62**
Resentment	.59	.27	.20	.55	**.75**
Suspicion	.59	.26	.32	.49	**.73**
Indirect hostility	.54	.38	.34	.55	.41
Negativism	.29	.21	.11‡	.26	.27
Guilt	.15‡	−.02‡	−.01‡	.19	.30
Aggression Questionnaire Short Form (Bryant & Smith, 2001)					
BDHI total	**.81**	.56	.57	.65	.57
Physical assault	.60	**.73**	.32	.41	.25
Verbal hostility	.59	.49	**.69**	.33	.15‡
Irritability	.68	.30	.49	**.67**	.54
Resentment	.60	.26	.23	.53	**.73**
Suspicion	.52	.23	.32	.43	.53
Indirect hostility	.54	.34	.36	.47	.40
Negativism	.24	.19	.09‡	.22	.21
Guilt	.14‡	−.03‡	.02‡	.18†	.26
Brief Aggression Questionnaire					
BDHI total	**.80**	.51	.51	.66	.58
Physical assault	.67	**.81**	.38	.39	.23
Verbal hostility	.63	.51	**.68**	.39	.17†
Irritability	.62	.23	.39	**.63**	.54
Resentment	.51	.16†	.11‡	.54	**.68**
Suspicion	.56	.19†	.28	.48	**.69**
Indirect hostility	.45	.23	.27	.45	.34
Negativism	.23	.14‡	.10‡	.21	.21
Guilt	.09‡	−.08‡	−.03‡	.10‡	.31

Notes. All correlations were significant at $P < .05$ except †$P < .10$ and ‡*ns*. Total–subscale correlations >.80 and subscale–subscale correlations >.60 appear in boldface.

BPAQ and the BDHI with one exception: for the 29-item BPAQ, Hostility was strongly related to Resentment and Suspicion in the BDHI (i.e., rs > .70). For the BPAQ-SF, however, Hostility was related to only Resentment, whereas the BAQ showed strong correlations with both Resentment and Suspicion. Because both the Resentment and Suspicion subscales of the BDHI were highly correlated with the BPAQ's Hostility subscale, we chose to include both as "on-diagonal," convergent measures of the Hostility subscales for the BPAQ-SF and BAQ in subsequent analyses.

If our BAQ scales more accurately represent the BDHI dimensions than the BPAQ-SF, then the pattern of convergent validity suggested by the pattern of correlations between the BDHI and the full BPAQ should be stronger

for the BAQ (vs. the BPAQ-SF), and the off-diagonal, discriminant validity-correlations should be weaker for the BAQ (vs. the BPAQ-SF). As expected, the on-diagonal, convergent validity correlations were just slightly stronger for the BAQ scales ($M_r = 0.70$) compared to those of the BPAQ-SF ($M_r = 0.68$), but not significantly, $t(4) = 0.74$, $P = .50$, $d = 0.37$ (Table IV). Also as expected, discriminant validity correlations between the abbreviated aggression measures and the BDHI subscales (excluding Guilt) were significantly lower for the BAQ subscales ($M_r = 0.31$) compared to the BPAQ-SF ($M_r = 0.33$), $t(22) = -2.29$, $P = .032$, $d = -0.49$.

Summary. Study 1 showed three key findings. First, the BAQ's subscales demonstrated the predicted pattern of convergent and discriminant validity with the subscales of both the BPAQ and the BDHI (its "parent" and "grandparent" measures, respectively). Second, a PAF analysis illustrated that the BAQ replicates the four-factor structure of the BPAQ. Third, the BAQ showed a slightly cleaner factor structure—and slightly better convergent and discriminant validity—than the BPAQ-SF. To be sure, although the direction of these comparisons was consistent with our predictions, the extent of most of these differences was non-significant. Overall, the BAQ did a reasonable job of efficiently measuring the four factors from the original PBAQ. We return to comparing the BPAQ-SF and the BAQ in Studies 4 and 5. Meanwhile, in Study 2, we sought to replicate the BAQ's four-factor structure using confirmatory factor analyses (CFAs) in two large samples.

STUDY 2: MEASUREMENT MODELS

Although the PAF results in Study 1 supported a four-factor solution for the BAQ, the sample size was modest for factor analytic methods. To this end, in Study 2, we sought to use two large independent samples ($Ns > 500$) from different U.S. regions to test the four-factor structure of the BAQ using a series of CFAs. Crucially, CFAs produce goodness-of-fit indexes (e.g., χ^2), which allow for different nested measurement models to be compared empirically.

We compared three maximum-likelihood measurement models of the BAQ (Fig. 1). First, we tested a four-factor model—one factor per subscale. Because we believe that the four-factor model will be widely adopted, we tested the extent to which gender differences moderated item-based factor loadings as a set. We did this because men tend to report and enact more aggression—particularly unprovoked Physical Aggression—than women (Archer, 2004; Bettencourt & Miller, 1996; Eagly & Steffen, 1986). Second, we tested a hierarchical model in which the four subscale-based latent factors loaded on a second-order latent factor of global aggression. Third, we tested a one-factor model, in which all 12 items loaded on a single latent factor of global aggression. Because we developed the BAQ to optimize subscale items (vs. the scale as a whole), we expected this model to fit the data less well than the other two models.

Method

In both samples, participants were asked to complete the 12-item BAQ as part of a mass-testing session at the start of a semester. Samples 1 and 2 used 5- and 9-point response scales, respectively, with anchors of *extremely uncharacteristic of me* to *extremely characteristic of me*.[3]

Sample 1. Participants were 552 undergraduates (235 men, 316 women, and 1 unknown gender) enrolled in introductory psychology courses at a public university in Virginia who received course credit for completing online mass-testing measures (82% non-Hispanic White or European American; ages: 18–26 years, $Mdn = 19$, $M = 18.9$, $SD = 1.1$).

Sample 2. Participants were 1,000 undergraduates (381 men, 609 women, and 10 unknown gender) enrolled in introductory psychology courses at a public university in Colorado who received course credit for completing bubble-sheet mass-testing measures (83% non-Hispanic White or European American; ages: 16–25 years, $Mdn = 19$, $M = 19.0$, $SD = 1.3$).

Results and Discussion

Descriptive statistics for Samples 1 and 2 appear in Table V. The fit statistics for all four measurement models (CFAs) for both Samples 1 and 2 appear in Table VI and Figure 1. In both samples, the four-factor model fit the data adequately and better than each of the other two measurement models. Because these models were nested, and because the four-factor model was the most complex, comparatively simpler models significantly worsened fit. In broad terms of absolute goodness-of-fit, the four-factor and hierarchical models showed adequate—but not good—fit, and the one-factor model showed poor fit. In terms of comparative fit, the four-factor model fit the data better than the hierarchical model, which fit better than the one-factor model.

Given the meta-analytic evidence for gender differences in aggression (Archer, 2004; Bettencourt & Miller, 1996; Eagly & Steffen, 1986), we next tested the extent to which the item loadings as a set differed by gender in the four-factor model across both samples (e.g., Webster & Bryan, 2007). For each sample, we first tested unconstrained models that freed all parameters to differ by gender, and then tested models that constrained the

[3] We varied response scale length (5-, 7-, 9-, and 10-point) in this study and subsequent ones. This did not affect the BAQ's structural characteristics and is consistent with prior research that systematically varied self-report response scale length (see Dawes, 2008). Items in both Study 2 samples were slightly positively skewed but met maximum likelihood assumptions.

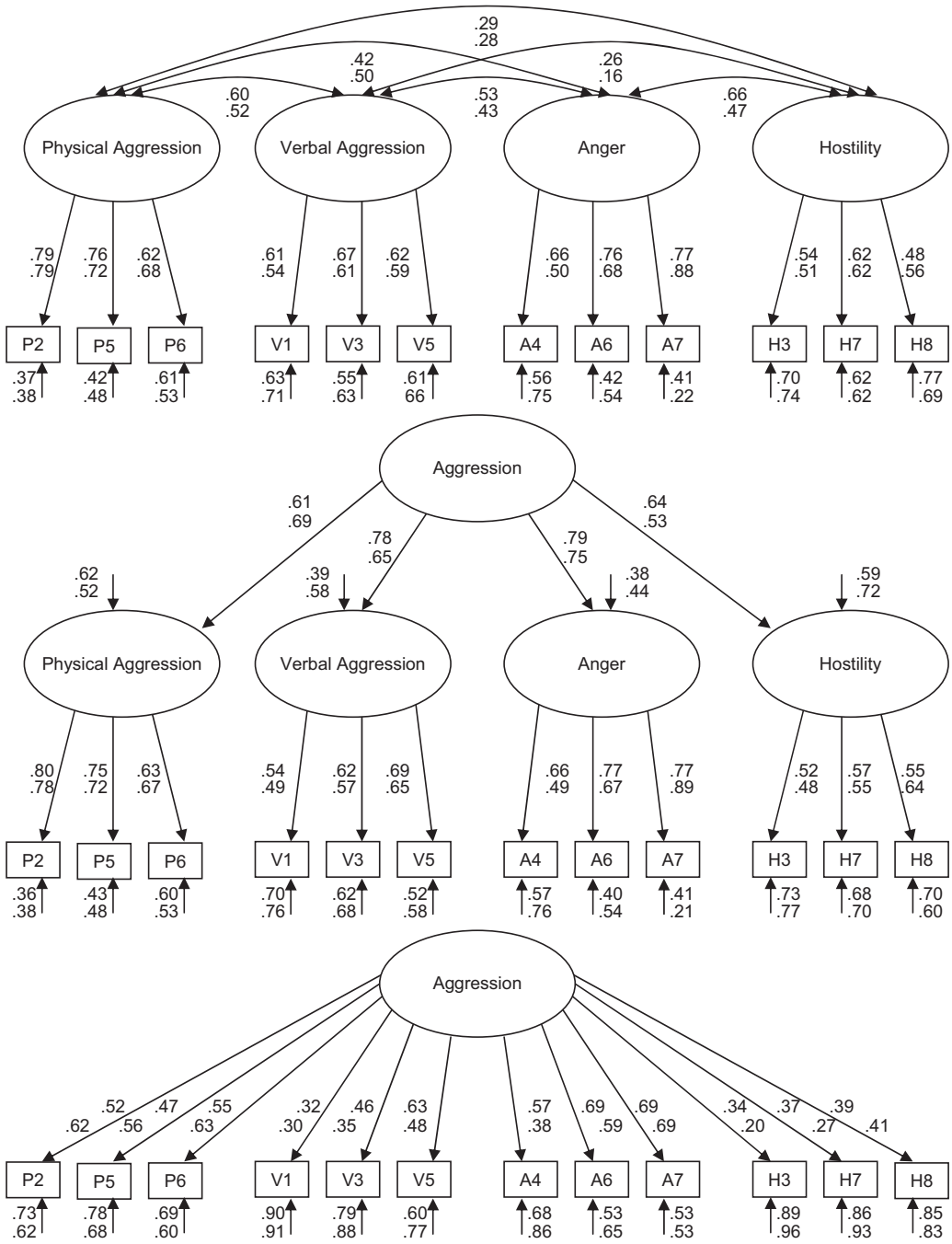

Fig. 1. Study 2, Samples 1 and 2: confirmatory factor analysis (CFA) results for the four-factor (top), hierarchical (middle), and one-factor (bottom) models. Upper and lower standardized coefficient pairs correspond to Samples 1 and 2 (*Ns* = 552 and 1,000). All residual error terms were left uncorrelated.

item loadings to be the same for both genders (e.g., "I am an even-tempered person" would be constrained to have equal loadings on the Anger factor for both men and women). Tests showed no significant difference in either sample, suggesting that—as a set—item loadings did not differ by gender. Although there are mean-level differences in aggression (particularly Physical Aggression), we can assume some structural equivalency regarding how men and women respond to the BAQ items given its four-factor structure.

STUDY 3: ITEM RESPONSE THEORY: INFORMATION AND DIFFERENTIAL RESPONDING

Studies 1 and 2 each relied on Classical Test Theory (CTT) techniques to assess items, scales, and subscales

TABLE V. Study 2: Descriptive Statistics (Samples 1 and 2)

	Sample 1 ($N = 552$)				Sample 2 ($N = 1,000$)			
	M	*SD*	α	MIC	*M*	*SD*	α	MIC
Physical aggression	2.58	1.04	.77	.52	3.63	2.15	.77	.52
Verbal aggression	3.08	0.84	.66	.40	5.30	1.70	.60	.60
Anger	2.23	0.83	.77	.54	3.19	1.69	.71	.46
Hostility	2.70	0.78	.56	.30	3.93	1.61	.58	.32
Total	2.65	0.62	.80	.25	4.02	1.21	.76	.21

Notes. Samples 1 and 2 used 1–5 and 1–9 response scales, respectively. MIC, mean inter-item correlation. An MIC of .30 yields αs = .81 and .56 for 10- and 3-item scales, respectively.

TABLE VI. Study 2: Confirmatory Factor Analysis Results

						90% CI		
Models or differences	χ^2	*df*	CFI	TLI	RMSEA	LL	UL	SRMR
Sample 1 ($N = 552$)								
Measurement models								
1. Four-factor	221.77	48	0.902	0.866	0.081	0.070	0.092	0.066
2. Hierarchical	270.45	50	0.876	0.837	0.089	0.079	0.100	0.072
2 vs. 1 Difference	48.68	2						
3. One-factor	677.95	54	0.650	0.572	0.145	0.135	0.155	0.098
3 vs. 2 Difference	407.50	4						
Four-factor: gender								
Unconstrained	297.66	96	0.882	0.838	0.087	0.076	0.099	0.073
Constrained items	304.34	104	0.883	0.852	0.084	0.073	0.095	0.075
Difference	6.68^{ns}	8						
Sample 2 ($N = 1,000$)								
Measurement models								
1. Four-factor	277.90	48	0.914	0.882	0.069	0.061	0.077	0.056
2. Hierarchical	302.14	50	0.906	0.876	0.071	0.063	0.079	0.057
2 vs. 1 Difference	24.24	2						
3. One-factor	1,041.30	54	0.632	0.550	0.135	0.128	0.142	0.091
3 vs. 2 Difference	739.16	4						
Four-factor: gender								
Unconstrained	308.22	96	0.913	0.881	0.067	0.059	0.075	0.056
Constrained items	323.44	104	0.910	0.886	0.065	0.057	0.073	0.059
Difference	15.22^{ns}	8						

Notes. Fit indexes (and suggested acceptable-fit cut-offs; see Browne & Cudeck, 1993; Hu & Bentler, 1999; Kline, 2011, for a review and critique): CFI, comparative fit index (≥0.90); TLI, Tucker–Lewis index (non-normed fit index or NNFI; ≥0.90); RMSEA, root mean square error of approximation (≤0.08); LL and UL, lower and upper limits; SRMR, standardized root mean square residual (≤0.08). Diff., difference. All χ^2 and RMSEA statistics were significant at $P < .05$ except *ns*. Samples 1 and 2 used 5- and 9-point response scales, respectively.

(e.g., PAF, CFA). IRT offers key improvements over CTT; it is more analytically flexible and gives researchers more information (see Fraley, Waller, & Brennan, 2000; Morizot, Ainsworth, & Reise, 2007). In Study 3, we expected our IRT models to show that the BAQ's subscales efficiently recover information about their respective underlying traits. Specifically, we expected that Physical Aggression would show the greatest gender differences, given meta-analytic gender differences in unprovoked behavioral aggression (Archer, 2004; Bettencourt & Miller, 1996; Eagly & Steffen, 1986). We expected gender differences in the other three subscales to be smaller than that for Physical Aggression.

Method

Participants. Participants were 1,790 undergraduates (720 men, 1,064 women, and 6 unknown gender) enrolled in introductory psychology courses at a public university in Colorado who received course credit for participation (82% non-Hispanic White or European American; ages: 17–36 years, $Mdn = 19$, $M = 18.9$, $SD = 1.4$).

Measures and procedure. As part of mass-testing sessions, participants were asked to complete the 12-item BAQ using a 10-point response scale from 1 (*extremely uncharacteristic of me*) to 10 (*extremely characteristic of me*).

Results and Discussion

Item-level correlations and descriptive statistics. Table C (see online Supporting Information) shows the item-level correlations and descriptive statistics for all 12 BAQ items by gender, with women's and men's statistics shown above and below the diagonal, respectively (scale- and subscale-level descriptive statistics are shown in Table VII). Overall, the correlation matrix conformed to the expected pattern, with same-subscale correlations being higher than other correlations. Means and *SD*s suggested that men tended to score higher on items relating to verbal and Physical Aggression, whereas gender differences for Anger and Hostility were comparatively weaker in magnitude.

Item response theory results: Discrimination, difficulty, and information. Because IRT models assume unidimensionality, we began by running separate graded-response polytomous models (analogous to two-parameter logistic models—2PLMs—for dichotomous data) on each of the four BAQ subscales. The discrimination (α) and difficulty (β_{1-9}) parameters from these four models appear in Table D (online Supporting Information). Item *discrimination* is the degree to which an item can differentiate between people with similar levels of the same latent trait. Item *difficulty* is the amount of the latent trait necessary to have a 50% chance of endorsing the item. The number of difficult parameters is one less than the number of categorical response options (i.e., β_{1-9} for a 10-point Likert scale). The discrimination parameters (α) ranged from 0.58 to 2.45, and most were over 1.0, indicating that these items were good at discriminating between people along their respective latent trait measures. The nine difficulty parameters for each item (β_{1-9}) suggested that aggression items tend to be difficult to endorse, perhaps because aggression is often considered a socially undesirable trait. Because of the large sample size ($N = 1{,}790$) none of the four IRT models fit the data well using null-hypothesis testing, χ^2s(969) > 1,119, Ps < .001.

Another key concept in IRT is item- and scale-level *information*, which is related to the concept of precision-of-measurement in CTT. Whereas CTT assumes the same reliability across all levels of a trait (Cronbach's α), IRT relaxes this assumption, presuming that measurement precision can vary across levels of a latent trait, with greater precision near the middle (vs. the ends) of the latent trait where there is more information. The corresponding scale information curves (SICs) for each subscale are shown in Figure A (top, see online Supporting Information). First, both Anger and Physical Aggression had similar amounts of information, and both had more information than Hostility and Verbal Aggression, which had similar information amounts. Second, in terms of information profiles, anger and Physical Aggression were slightly positively skewed, suggesting that these subscales did a better job at discriminating at higher levels of the latent trait than at its lower levels. In contrast, Hostility and Verbal Aggression were more

TABLE VII. Study 3: Descriptive and Information Statistics for the Brief Aggression Questionnaire and Its Subscales by Participant Gender

	Descriptives		Reliability		Information		Gender diff.	
Subscale	*M*	*SD*	α	MIC	Total	/Item	*t*	*d*
All								
BAQ total	4.30	1.34	.80	.24	43.13	3.59	11.53*	0.55
Anger	3.52	1.84	.78	.55	13.74	4.58	0.27	0.01
Physical	3.73	2.37	.80	.58	12.82	4.27	19.84*	0.94
Hostility	4.35	1.63	.57	.31	7.66	2.55	2.23*	0.11
Verbal	5.60	1.81	.63	.37	8.91	2.97	7.27*	0.34
Men								
BAQ total	4.73	1.36	.80	.24	27.19	2.27		
Anger	3.54	1.88	.78	.55	13.18	4.39		
Physical	4.95	2.44	.80	.57	13.30	4.43		
Hostility	4.45	1.64	.54	.28	7.16	2.39		
Verbal	5.97	1.75	.58	.33	8.43	2.81		
Women								
BAQ total	4.01	1.25	.79	.24	26.20	2.18		
Anger	3.51	1.82	.78	.55	14.42	4.81		
Physical	2.90	1.92	.72	.47	10.05	3.35		
Hostility	4.28	1.62	.59	.32	7.87	2.62		
Verbal	5.35	1.80	.64	.37	9.15	3.05		

Notes. /Item, per item; Diff., difference; MIC, mean inter-item correlation (e.g., an MIC of .30 yields αs = .81 and .56 for 10- and 3-item scales, respectively).
*$P < .05$.

symmetrical, and had broader information profiles (thicker tails), which suggests they did a comparatively better job of discriminating at both extremes of their respective latent traits.

These SICs are impressive given that each subscale has only three items. Table VII shows the total information (area under the curve; AUC) for each SIC for each subscale, and the total information per item. Information in IRT is related to reliability in CTT; this relationship was confirmed in the present study: scale reliability coefficients (αs) were significantly correlated with total information from the four subscales, $r(2) = .98$, $P = .02$. Thus, subscales with greater internal consistency had larger information profiles.

Differential item and scale functioning. We next examined differential item functions (DIF) in a series of models for each subscale and the total BAQ, and then constructed differential scale functions (DSF) for each subscale and the total BAQ by gender (Table VIII; Figs. A–C; see online Supporting Information). Table VIII shows a consistent pattern of DIF. First, allowing for DIF in the total BAQ did not fit the data well. Recall that a key assumption of IRT is unidimensionality, which is not met with the BAQ because it has four dimensions or factors. Thus, the poor fit here has more to do with multidimensionality than it does with gender-based DIF. Second, every subscale but Physical Aggression produced fit indexes suggesting that their respective models fit the data well when allowing for gender-based DIF. In contrast, although Physical Aggression produced acceptable fit indexes for the CFI and TLI, it was the only subscale to produce a significant RMSEA, suggesting poor fit. Although this DIF model accounted for gender differences, the model misfit may relate to unmeasured individual differences that could further delineate differences in men's Physical Aggression. This is because (a) the fit statistic (χ^2) was nearly three times greater for men (100) than for women (34) and (b) the *SD* for men's Physical Aggression is more than 0.50 larger than that for either women's Physical Aggression or men's *SD*s on the other three BAQ subscales. Thus, men in particular vary a great deal in Physical Aggression, and gender differences alone cannot explain this variability (see Archer & Mehdikhani, 2003, for meta-analytic evidence).

Figures A–C (see online Supporting Information) show some interesting patterns of differential scale responding by gender—key differences that would likely be masked using CTT alone. First, the DSF for total BAQ —which should be interpreted with caution because it is not a unidimensional construct—shows that the SICs for both men and women are slightly positively skewed, but more so for men (Fig. A, bottom, see online Supporting Information). This suggests that the BAQ does a slightly better job of distinguishing between individuals at higher levels of latent aggression than it does at its lower levels; however, this is especially true for men, who tend to report more aggression overall. In addition, men have a higher peak in information and greater AUC, suggesting that the BAQ's scores are slightly more reliable, precise, and informative for men than for women, despite its less symmetrical SIC.

TABLE VIII. Study 3: Differential Items Functioning by Gender for the Brief Aggression Questionnaire (BAQ) and Its Four Subscales

						90% CI			
Models	χ^2	*df*	CFI	TLI	RMSEA	LL	UL	P_{close}	WRMR
BAQ total	3,522.27	203	0.724	0.821	0.135	0.131	0.139	.000	4.73
Women	1,679.61								
Men	1,842.66								
Anger	40.46	23	0.996	0.999	0.029	0.013	0.044	.992	1.04
Women	17.22								
Men	23.25								
Physical	134.25	23	0.969	0.992	0.074	0.062	0.086	.001	1.83
Women	34.32								
Men	99.93								
Hostility	66.47	23	0.949	0.987	0.046	0.033	0.059	.672	1.47
Women	25.40								
Men	41.07								
Verbal	23.40^{ns}	23	1.0	1.0	0.004	0.000	0.028	1.0	0.73
Women	8.33								
Men	15.06								

Notes. $N = 1{,}790$. Fit indexes (and suggested acceptable-fit cut-offs; see Browne & Cudeck, 1993; Hu & Bentler, 1999; Kline, 2011, for a review and critique): CFI, comparative fit index ($\geq$0.90); TLI, Tucker–Lewis index (non-normed fit index or NNFI; $\geq$0.90); RMSEA, root mean square error of approximation ($\leq$0.08); LL and UL, lower and upper limits; WRMR, weighted root mean square residual. All χ^2 statistics were significant at $P < .05$ except *ns*.

The SICs for Anger and Hostility show similar patterns of DSF (Fig. B, see online Supporting Information). First, the SICs for both genders were slightly but equally positively skewed; this skew was less than that for verbal or Physical Aggression. Second, although both genders' SICs peaked in the same locations, women had a higher peak, and thus a greater amount of information was recovered.

The SICs for verbal and Physical Aggression also showed similar patterns of DSF (Fig. C, see online Supporting Information). First, the SICs for both men and women were somewhat positively skewed, but not equally so; men's information profile was more positively skewed than women's for each subscale. This suggests that the upper categories of the response scale were more diagnostic in differentiating men along these latent traits than the lower categories. Moreover, the skew—and the gender difference in the skew—was greater than that for either Anger or Hostility. Second, men's and women's SICs peaked in different locations, with men's peaks being roughly 0.50 and 0.75 *SD*s above those for women for verbal and Physical Aggression on their respective latent traits. Women had a higher peak and more total information than men for Verbal Aggression, whereas men had a higher peak and more total information than women for Physical Aggression.

Recommendations. Gender differences may be an inherent property of the underlying latent trait (Study 3 results) rather than the items themselves (Study 2 results). Meta-analytic evidence suggests that men are more physically aggressive than women (Archer, 2004; Bettencourt & Miller, 1996; Eagly & Steffen, 1986). That this is borne out in DSF may be a reflection of this fact rather than a systematic bias in the measure. In addition, prior aggression measures (e.g., BDHI, BPAQ) have used the same items for men and women, and researchers who have used these scales often control for gender in their analysis to account for gender differences (e.g., Webster, Kirkpatrick, Nezlek, Smith, & Paddock, 2007). For these reasons, we recommend that researchers consider controlling for gender differences when using the BAQ and particularly its Physical Aggression subscale, not necessarily because the BAQ items are biased, but because the latent traits themselves tend to reflect real underlying gender differences.

STUDY 4: TEST–RETEST RELIABILITY AND CONVERGENT VALIDITY

Having confirmed the structure and psychometric robustness of the BAQ in Studies 1–3, we examined its test–retest reliability in Study 4. Establishing test–retest reliability is essential to developing new or brief scales because trait-level individual differences should be stable over time. To this end, we measured the 29-item BPAQ at two time points 3 weeks apart, which allowed us to compare the test–retest correlations between the BAQ and the BPAQ-SF.

Method

Participants were 238 undergraduates enrolled in introductory psychology classes at a public university in Kentucky who were asked to complete the 29-item BPAQ (1—*extremely uncharacteristic of me* to 7—*extremely characteristic of me*) twice—3 weeks apart—as part of a broader field study. Of these, 207 (87%) had complete data for all four subscales at Times 1 and 2 (58 men, 148 women, 1 unknown; 85% non-Hispanic White or European American; ages: 18–31 years, $Mdn = 18$, $M = 18.6$, $SD = 1.2$). During each laboratory session, participants completed the BPAQ on desktop computers in private, individual cubicles.

Results and Discussion

Descriptive statistics for all three scales at both time points are in Table IX. Test–retest reliability correlations (Table X, left columns) were somewhat stronger on average for the BAQ subscales ($M_r = .64$) than those of the BPAQ-SF ($M_r = .59$), but not significantly, $t(3) = 1.94$, $P < .15$, $d = 1.12$; however, the pattern was consistent: the BAQ had stronger test–retest correlations for all four subscales and the total score than did the BPAQ-SF. Just as key, the *average strength* of the test–retest reliability correlations for the BPAQ subscales ($M_r = .65$) did not differ from that of the BAQ, $t(3) = 0.55$, $P = .62$, $d = 0.32$, but did differ marginally from that of the BPAQ-SF, $t(3) = 2.66$, $P < .08$, $d = 1.54$. The *pattern* of subscale test–retest correlations between the BAQ and the BPAQ were remarkably similar, $r(2) = .94$, $P < .06$; this was not true for patterns between the BPAQ-SF and the BPAQ or BAQ, rs$(2) < .79$, Ps $> .21$. Thus, both in terms of average strength (magnitude) and pattern, the BAQ's test–retest correlations more closely mirrored those of its parent measure—the BPAQ—than did those for the BPAQ-SF.

Following the recommendations of Smith, McCarthy, and Anderson (2000), we also examined "parent–child" (or whole–part) convergent validity correlations among the BPAQ and its two short forms across time. This was done because administering the short-form items nested within the long-form can artificially inflate correlations between the short- and long-forms since the items in the short-form contribute to the correlation twice. To address this concern, we assessed convergent validity correlations for (a) Time-2 BPAQ with Time-1 BPAQ-SF and Time-1 BAQ and (b) Time-1 BPAQ with Time-2 BPAQ-SF and Time-2 BAQ (Table X, right columns). In every case, the BAQ had higher "parent–child" convergent validity correlation with the BPAQ than did the BPAQ-

TABLE IX. Study 4: Descriptive Statistics at Times 1 and 2 for the Buss–Perry Aggression Questionnaire (BPAQ), Its Short Form (BPAQ-SF), and the Brief Aggression Questionnaire (BAQ)

	BPAQ			BPAQ-SF			BAQ		
	M	*SD*	α	*M*	*SD*	α	*M*	*SD*	α
Time 1									
Physical aggression	2.39	1.12	.86	2.04	1.29	.80	2.31	1.53	.86
Verbal aggression	3.11	1.04	.76	2.65	1.13	.76	3.31	1.16	.64
Anger	2.48	1.03	.81	2.17	1.12	.71	2.38	1.12	.65
Hostility	2.65	1.14	.85	2.48	1.39	.84	2.65	1.27	.71
Total	2.61	0.84	.91	2.33	0.93	.86	2.66	0.86	.79
Time 2									
Physical aggression	2.33	1.09	.85	1.90	1.31	.86	2.17	1.44	.85
Verbal aggression	2.78	1.17	.85	2.35	1.21	.86	3.00	1.30	.78
Anger	2.32	0.98	.82	2.00	1.08	.77	2.33	1.03	.59
Hostility	2.51	1.22	.90	2.41	1.50	.87	2.46	1.29	.76
Total	2.46	0.89	.93	2.17	1.01	.90	2.49	0.91	.83

Notes. $N = 207$. Response scale: 1–7.

TABLE X. Study 4: Test–Retest Correlations and "Parent–Child" (or Part–Whole) Convergent Validity Correlations Over Time for the Buss–Perry Aggression Questionnaire (BPAQ), Its Short Form (BPAQ-SF), and the Brief Aggression Questionnaire (BAQ)

				"Parent–child" correlations			
	Test–retest correlations			$BPAQ_1$		$BPAQ_2$	
	BPAQ	BPAQ-SF	BAQ	$BPAQ\text{-}SF_2$	BAQ_2	$BPAQ\text{-}SF_1$	BAQ_1
Physical aggression	.713	.658	.671	.655	.665	.656	.660
Verbal aggression	.550	.539	.572	.473	.533	.498	.538
Anger	.658	.539	.668	.547	.631	.586	.587
Hostility	.670	.628	.656	.606	.631	.597	.640
Total	.678	.619	.658	.623	.657	.642	.654

Notes. $N = 207$. Response scale: 1–7. Test–retest interval: 3 weeks. Times 1 and 2 indicated by subscripts.

SF. When we submitted these correlations to a 2 (BPAQ Time: 1 vs. 2) × 2 (Measure: BAQ vs. BPAQ-SF) repeated-measures ANOVA, a significant main effect emerged for Measure, $F(1, 3) = 15.18$, $P < .03$, $\eta^2 = .84$, showing that the BAQ's subscales had significantly higher convergent validity correlations with the BPAQ's subscales ($M_r = .61$) than did the BPAQ-SF's subscales ($M_r = .58$); neither the main effect of BPAQ Time nor the BPAQ Time × Measure interaction was significant, Ps $> .38$, η^2s $< .26$.

At both time points (Time 1 and Time 2), the off-diagonal correlations among the four BAQ subscales were significantly lower for the BAQ (M_{r}s $= .28, .36$) than those of the BPAQ-SF (M_{r}s $= .45, .53$), ts(5) ≤ -2.55, Ps $\leq .05$, ds ≤ -1.44 (Table XI). These findings suggest that the BAQ does a better job of measuring the four aggression domains as independent facets than the BPAQ-SF.

STUDY 5: CONVERGENT VALIDITY WITH BEHAVIORAL AGGRESSION

Study 5 had two goals. First, because we measured the 29-item version of the BPAQ, we again examined which of the two brief measures—the BAQ or the BPAQ-SF—was more closely related to its parent measure. Second, we examined the convergent validity of the BAQ's Physical Aggression subscale with measures of behavioral aggression. We expected that the BAQ would perform as well as the BPAQ-SF in convergent validity with the aggression measures.

Method

Participants. Participants were 307 undergraduates enrolled in introductory psychology courses at a public university in Kentucky who received course credit for participation (91 men, 216 women; ages: 18–41 years, $Mdn = 19$, $M = 19.3$, $SD = 2.3$).[2] Participants arrived to the study in two- to eight-person groups and completed the study on desktop computers in private, individual cubicles during a single laboratory session.

Measures. Participants completed the 29-item BPAQ using a 7-point response scale (1—*extremely uncharacteristic of me* to 7—*extremely characteristic of me*).

Procedure. We told participants that, to obtain a good measure of their reaction time, they would be paired with one of the other participants present and instructed

TABLE XI. Study 4: Partial Correlation Matrix: Test–Retest Reliability Correlations (Bolded) for the Buss–Perry Aggression Questionnaire—Short Form (Above the Diagonal) and the Brief Aggression Questionnaire (Below the Diagonal)

	Time 1					Time 2				
	Physical	Verbal	Anger	Hostility	Total	Physical	Verbal	Anger	Hostility	Total
Time 1										
Physical	—	.41	.47	.28	.71	**.66**	.29	.38	.20	.47
Verbal	.39	—	.60	.43	.78	.32	**.54**	.36	.28	.46
Anger	.26	.21	—	.47	.82	.36	.33	**.54**	.35	.49
Hostility	.21	.26	.33	—	.74	.21	.24	.32	**.63**	.46
Total	.74	.67	.63	.65	—	.51	.45	.52	.49	**.62**
Time 2										
Physical	**.67**	.26	.23	.14	.51	—	.56	.68	.38	.81
Verbal	.24	**.57**	.08‡	.22	.41	.47	—	.66	.37	.79
Anger	.26	.18	**.67**	.29	.50	.39	.22	—	.49	.86
Hostility	.22	.16	.27	**.66**	.48	.32	.33	.39	—	.73
Total	.50	.41	.41	.45	**.66**	.79	.72	.65	.71	—

Notes. All correlations were significant at $P < .05$ except ‡*ns*. $N = 207$. The test–retest interval was 3 weeks.

to compete with that participant on a competitive reaction-time game. The task was a modified version of the Taylor Aggression Paradigm (Taylor, 1967), in which both members of each pair ostensibly competed against each other over who could respond more quickly, with the winner delivering a tone blast to their partner. In reality, participants completed a reaction-time task against a computer program, which was programmed to mimic another person's actions. Of the 25 trials, participants won 13; the order of wins and loses was randomized for each participant. Prior to each trial, participants set the intensity of the noise (0–105 dB, about the volume of smoke alarm) and selected the duration of how long their partner would suffer (0.0–5.0 sec). This task is a well-validated measure of laboratory aggression (e.g., Anderson & Bushman, 1997; Giancola & Chermack, 1998). During debriefing, none of the participants expressed suspicion.

Results and Discussion

Preliminary analyses. Because we used the 29-item BPAQ, we compared the BAQ and BPAQ-SF, and assessed the extent to which they accurately reflected the BPAQ and its four subscales. Table II (below diagonal) shows the descriptive statistics for—and the correlations among—the subscales for the BPAQ, BPAQ-SF, and BAQ. As expected, the BAQ subscales had significantly lower intercorrelations ($M_r = .31$) than the BPAQ-SF subscales ($M_r = .44$), $t(5) = -3.50$, $P = .02$, $d = -1.57$. Once again, the BAQ subscales were more mutually independent than those of the BPAQ-SF.

We next examined a partial MTMM involving the extent to which the 29-item BPAQ correlated with the BPAQ-SF and the BAQ (Table II, below diagonal). As expected, convergent validity correlations were slightly stronger for the BAQ ($M_r = .92$) than the BPAQ-SF ($M_r = .91$), but not significantly, $t(3) = 0.49$, $P = .66$, $d = 0.28$; however, when we included only those subscales with positively valenced items for a fairer comparison (i.e., excluding Anger), the convergent validity correlations were marginally stronger for the BAQ ($M_r = .93$) than the BPAQ-SF ($M_r = .90$), $t(2) = 3.69$, $P < .07$, $d = 2.61$. When the last result above was combined with its parallel result from Study 1 in a 2 (Study: 1 vs. 5) × 2 (Measure: BAQ vs. BPAQ-SF) repeated-measures ANOVA, a significant main effect emerged for Measure, $F(1, 2) = 19.84$, $P < .05$, $\eta^2 = .91$, showing that the BAQ's subscales had significantly higher convergent validity correlations with the BPAQ's subscales ($M_r = .933$) than the BPAQ-SF's subscales had with the BPAQ's subscales ($M_r = .905$); neither the main effect of Study nor the Study × Measure interaction was significant, Ps $> .80$, η^2s $< .04$. Also as expected, off-diagonal correlations among the abbreviated and full versions of the four subscales were significantly lower for the BAQ ($M_r = .38$) than those for the BPAQ-SF ($M_r = .45$), $t(11) = -5.05$, $P < .001$, $d = -1.52$. Together, these results suggest the BAQ's subscales show better convergent and discriminant validity with its parent measure's subscales (i.e., the BPAQ) than do the BPAQ-SF's subscales; however, the former was only true for positively valenced subscales.

Convergent and discriminant validity. The behavioral aggression data consisted of noise blast duration ($M = 4.92$, $SD = 2.20$) and intensity ($M = 5.18$, $SD = 2.26$) across 25 trials. Possible responses ranged from 0 to 10 for both measures, which were strongly and positively correlated, $r(305) = .86$, $P < .001$. Using the noise blast data, we constructed four unique but related measures of behavioral aggression: *total aggression*

(mean intensity and duration, standardized and then summed), *unprovoked aggression* (intensity and duration from first trial standardized and then summed), *extreme aggression* (number of times people selected extreme intense [9 or 10] noise; $M = 6.65$, $SD = 6.95$; Bushman, Ridge, Das, Key, & Busath, 2007), and *aggressive energy* (multiplying intensity with the square root of duration for each trial, then averaging the products across the 25 trials; $M = 10.83$, $SD = 7.69$; Carnagey & Anderson, 2005).

As expected, convergent validity correlations were indeed significantly stronger for the BAQ's Physical Aggression subscale ($M_r = .24$) than they were for the BPAQ-SF's ($M_r = .21$), $t(3) = 8.87$, $P < .01$, $d = 5.12$ (Table XII). Also as expected, discriminant validity correlations between behavioral aggression and the three non-Physical Aggression subscales were significantly lower for the BAQ subscales ($M_r = .08$) than BPAQ-SF's ($M_r = .11$), $t(27) = -5.36$, $P < .001$, $d = -1.03$. Because the behavioral measures were intercorrelated (rs = .61 to .88), the paired t-tests may be liberal. Nevertheless, the overall pattern of correlations showed that the BAQ's Physical Aggression subscale was a valid measure of Physical Aggression (and that the other subscales were not) and that the BAQ had significantly better convergent and discriminant validity than the BPAQ-SF.

Multilevel modeling. We used multilevel modeling (MLM) because multiple trials (25) were nested within each participant. Using maximum likelihood estimation, MLM allows for the simultaneous modeling of within- and between-person effects (Nezlek, 2008, 2011; Raudenbush & Bryk, 2002). Specifically, within-person (or between-trial) variance in noise blast duration or intensity was modeled at Level 1, and between-person variance in noise blast duration or intensity was modeled at Level 2 as a function of individual differences in the self-report aggression measures (e.g., the BAQ). For example, the results described in the bottom half of Table XIII, Step 2, were based on a MLM where the Level 1 model was

$$\text{Intensity}_{ti} = \pi_{0i} + e_{ti},$$

where Intensity_{ti} represents the noise blast intensity given out at Time t by Person i. The Level 1 model is a null model, meaning that each person's intensity scores are modeled only by π_{0i}, which represents the mean or intercept for each person. The Level 1 residual variance is captured by the error term e_{ti}.

In MLM, the Level-1 means or intercepts for each person are modeled simultaneously at Level 2 as a function of the four BAQ subscales (grand-mean-centered) and gender (uncentered; coded −0.5 = women, 0.5 = men):

$$\begin{aligned}\pi_{0i} = {} & \beta_{00} + \beta_{01}(\text{Physical}) + \beta_{02}(\text{Verbal}) \\ & + \beta_{03i}(\text{Anger}) + \beta_{04}(\text{Hostility}) + \beta_{05}(\text{Gender}) \\ & + r_{0i}.\end{aligned}$$

Here, π_{0i} again represents the mean or intercept for each person. The β_{00} coefficient represents the grand mean—the between-person average of each person's average intensity score. The coefficients β_{01}–β_{04}

TABLE XII. Study 5: Convergent and Discriminant Validity Correlations Among All Three Self-Report Aggression Measures and Four Behavioral Aggression Measures

	Total	Physical aggression	Verbal aggression	Anger	Hostility
29-Item Aggression Questionnaire (Buss & Perry, 1992)					
Aggression					
Total	.22	**.26**	.16	.10†	.12
Unprovoked	.13	**.18**	.06‡	.00‡	.12
Extreme	.23	**.30**	.18	.10†	.08‡
Energy	.18	**.25**	.08‡	.06‡	.11†
12-Item Aggression Questionnaire Short Form (Bryant & Smith, 2001)					
Aggression					
Total	.21	**.22**	.17	.10†	.14
Unprovoked	.11	**.14**	.06‡	.02‡	.12
Extreme	.23	**.27**	.19	.10†	.11
Energy	.16	**.22**	.09‡	.06‡	.12
12-Item Brief Aggression Questionnaire					
Aggression					
Total	.21	**.25**	.13	.08‡	.10†
Unprovoked	.13	**.16**	.06‡	.00‡	.12
Extreme	.23	**.29**	.15	.07‡	.07‡
Energy	.18	**.24**	.07‡	.04‡	.11†

Notes. $N = 307$. All correlations were significant at $P < .05$ except †$P < .10$ and ‡*ns*. Predicted correlations appear in boldface.

TABLE XIII. Study 5: Multilevel Model Results of Taylor Aggression Paradigm Noise Blast Duration and Intensity (25 Trials) as Functions of Different Aggression Questionnaire Subscales and Gender

	BPAQ			BPAQ-SF			BAQ		
	Coef.	*t*	r_p	Coef.	*t*	r_p	Coef.	*t*	r_p
Duration									
Step 1									
Physical	.382	3.32**	.19	.234	2.46*	.14	.254	3.08**	.17
Verbal	.141	1.06	.06	.169	1.37	.08	.056	0.52	.03
Anger	−.206	−1.34	−.08	−.129	−1.04	−.06	−.028	−0.23	−.01
Hostility	.168	1.36	.08	.143	1.33	.08	.091	0.83	.05
Step 2									
Physical	.161	1.22	.07	.063	0.57	.03	.105	1.12	.06
Verbal	.159	1.20	.07	.171	1.45	.08	.108	1.01	.06
Anger	−.060	−.41	−.02	−.017	−.14	−.01	.064	0.55	.03
Hostility	.191	1.56	.09	.159	1.52	.09	.095	0.86	.05
Gender	1.124	3.42***	.19	1.205	3.65***	.21	1.151	3.47***	.20
Intensity									
Step 1									
Physical	.489	4.13***	.23	.343	3.62**	.20	.339	4.10***	.23
Verbal	.113	0.87	.05	.154	1.26	.07	.027	0.25	.01
Anger	−.145	−.89	−.05	−.110	−.82	−.05	.007	0.06	.00
Hostility	.065	0.52	.03	.078	0.72	.04	.027	0.24	.01
Step 2									
Physical	.268	1.96	.11	.175	1.60	.09	.191	2.00*	.11
Verbal	.132	1.02	.06	.155	1.35	.08	.079	0.74	.04
Anger	.001	0.01	.00	.001	0.01	.00	.099	0.83	.05
Hostility	.088	0.73	.04	.094	0.90	.05	.031	0.83	.02
Gender	1.124	3.45***	.19	1.191	3.64***	.21	1.146	3.48***	.20

Notes. $N = 307$. Coef., unstandardized regression coefficient (slope); r_p, partial correlation.
*$P < .05$.
**$P < .01$.
***$P < .001$.

represent the moderating effects of their respective BAQ subscales, controlling for each other and gender differences (β_{05}). In ordinary least-squares regression terms, this is analogous to performing a multiple regression in which each person's mean intensity scores were regressed onto their BAQ subscale score and gender. Thus, the β_{01}–β_{04} coefficients represent their respective partial relationships with noise blast intensity. Similarly, the β_{05} coefficient represents the moderating effects of participant gender—the extent of gender differences in noise blast intensity, controlling for the BAQ subscales. The Level-2 residual variance is captured by the error term r_{0i}.

Table XIII shows the MLM results for noise blast duration (top) and intensity (bottom) for the BPAQ (left), the BPAQ-SF (middle), and the BAQ (right), each subdivided into two steps (subscales vs. subscales and gender), for a total of 12 independent models. These models are simply hierarchical regression models (i.e., multiple "steps"; see Aiken & West, 1991) in an MLM context. At Step 1, the four subscales were entered simultaneously to predict variance in duration or intensity. At Step 2, participants' gender was added; thus, the Step 2 results show the effects for each subscale controlling for gender (and the other three subscales).

Overall, a clear pattern of results emerged for noise blast duration. In Step 1, the Physical Aggression subscales of all three scales were significantly and positively related to noise blast duration; however, the BAQ was more strongly related ($r_p = .17$, $P < .01$) than the BPAQ-SF ($r_p = .14$, $P < .05$).[4] In Step 2, controlling for gender reduced these effects to non-significance.

A similarly clear pattern emerged for noise blast intensity. In Step 1, the Physical Aggression subscales of all three scales were significantly and positively related to noise blast intensity; however, the BAQ was more strongly related ($r_p = .23$) than the BPAQ-SF ($r_p = .20$, Ps $< .01$). In Step 2, controlling for gender reduced these effects to non-significance for the BPAQ and the BPAQ-SF, but not the BAQ ($r_p = .11$, $P < .05$). Only the BAQ had a Physical Aggression subscale that was

[4] Although we report partial correlations (r_{ps}) for comparison (and for consistency with Studies 1–4), they should be interpreted with caution because such estimates of effect size in MLM may not reflect those from ordinary least squares (see Kreft & de Leeuw, 1998, p. 119).

significantly related to behavioral aggression after controlling for gender.

We next evaluated brief versions of Physical Aggression subscale independently (Table XIV). In separate models, the BAQ was more strongly related to both noise blast duration (r_ps = .23 vs. .20) and intensity (r_ps = .28 vs. .27) than the BPAQ-SF. We then let the two brief Physical Aggression subscales compete to explain variance in noise blast duration and intensity (Table XIV). The BAQ was more strongly related to both noise blast duration (r_ps = .10 vs. −.02) and intensity (r_ps = .10 vs. .00) than the BPAQ-SF. Comparing the magnitudes of these coefficients using a multivariate hypothesis testing procedure (see Raudenbush & Bryk, 2002) showed that the BAQ's Physical Aggression subscale was more strongly related to both noise blast duration ($\chi^2_{(1)} = 26.59$) and intensity ($\chi^2_{(1)} = 15.29$) than the BPAQ-SF's (Ps < .001).

Because these two brief measures were correlated .92, the results should be interpreted with caution because suppression effects are possible (MacKinnon, Krull, & Lockwood, 2000). For noise blast duration, a suppression effect is likely because the direction of the relation for the BPAQ-SF switched from positive to negative, whereas the unstandardized regression coefficient became larger instead of smaller for the BAQ. Yet, these results suggest the BAQ's Physical Aggression subscale is a stronger correlate of behavioral aggression than the BPAQ-SF's.

This raises the key issue of just how correlated the BAQ and BPAQ-SF subscales are with each other. Given that each subscale shares one or two items in common, it is not surprising that these correlations ranged from .72 to .93 across Studies 1, 4, and 5.[5] In other words, one measure can explain between 51% and 86% of the variance in the other. Within each study, we tested if the average of the four subscales differed from a correlation of 1.0 using a series of one-sample *t*-tests on both raw and (*r*-to-*z*) transformed correlations. If the average correlation is significantly different from 1.0, then one cannot claim that the subscales of the two short forms are completely redundant. Raw correlations for Studies 1, 4, and 5 were significantly lower than 1.0 on average, ts(3) ≤ −4.49, Ps ≤ .02; ds ≤ −2.59. Transformed correlations for Studies 1, 4, and 5 were also significantly lower than 2.56 (the *r*-to-*z* transform of .99) on average, ts(3) ≤ −9.04, Ps ≤ .003; ds ≤ −5.22. Thus, although the BAQ and BPAQ-SF subscales share items, the unique variance that they do *not* share can be meaningful, as shown by the differences in their Physical Aggression subscales in predicting behavioral aggression in Study 5.

TABLE XIV. Study 5: Multilevel Model Results of Taylor Aggression Paradigm Noise Blast Duration and Intensity (25 Trials) as Functions of Different Physical Aggression Subscales: Separate and Multiple Regressions

	Duration			Intensity		
	Coef.	*t*	r_p	Coef.	*t*	r_p
Separate regressions						
Physical aggression						
BPAQ-SF	.292	3.58*	.20	.382	4.88*	.27
BAQ	.284	4.05*	.23	.354	5.17*	.28
Multiple regression						
Physical aggression						
BPAQ-SF	−.092	−0.40	−.02	.011	0.05	.00
BAQ	.357	1.82†	.10	.346	1.77†	.10

Notes. N = 307. Coef., unstandardized regression coefficient (slope); r_p, partial correlation; BPAQ-SD, Buss–Perry Aggression Questionnaire; BAQ, Brief Aggression Questionnaire.
*P < .001.
†P < .10.

GENERAL DISCUSSION

Aggression occupies a classic area of study inside and outside the social sciences. Yet the most widely used self-report aggression measure—the BPAQ—does not always accommodate aggression researchers' contemporary needs. Aggression researchers need a brief self-report aggression measure with scores that are reliable and valid, and one that can facilitate efficient testing in multiple settings. The current investigation sought to meet this need in the aggression literature by providing psychometric evidence of the BAQ's validity and reliability.

Five studies, including nearly 4,000 participants (over 1,500 more participants than the BPAQ and BPAQ-SF studies combined), offered converging evidence of the validity and reliability of the BAQ's scores. The BAQ's scores showed consistent patterns of convergent and discriminant validity with other self-report measures. Using PAF and CFAs, we showed that the BAQ accurately reflects the four-factor structure of the 29-item BPAQ. We used IRT to show that the BAQ items efficiently recover information and effectively discriminate among people along their respective latent trait measures. The BAQ's test–retest reliability is also strong. We provided evidence that the BAQ's Physical Aggression subscale is associated with validated laboratory measures of behavioral aggression. Although the BAQ outperformed the BPAQ-SF in most validity tests, the BPAQ-SF often outperformed the BAQ in internal

[5] Correlations between the Brief Aggression Questionnaire and the Buss–Perry Aggression Questionnaire Short Form appear in Table II for Studies 1 and 5. For Study 4, the average of Time 1 and Time 2 correlations were .93, .72, .76, .78, and .92 for physical aggression, verbal aggression, anger, hostility, and total, respectively (N = 207, ps < .05).

consistency reliability (Cronbach's α). We recommend that researchers interested in assessing aggression take these strengths and weaknesses into account when choosing a brief measure. Nevertheless, both brief measures showed good psychometric properties. In sum, the current suite of studies provided confirmation of the validity and reliability of the BAQ's scores as a brief aggression measure.

A broader implication of this research is that psychological constructs can be measured adequately by using parsimonious scales. Aggression is multiply determined, stemming from individual differences, situational factors, and their interaction (DeWall, Anderson, & Bushman, 2011). To capture the complexity of aggression, researchers began by developing lengthy self-report aggression measures that tapped many different components of aggression. Our research questions the need for long self-report measures to assess complex psychological constructs, such as aggression, especially when time or space is limited. The current research shows that the BAQ provides the most efficient multifaceted measure of aggression currently available.

Our research adds to a recent chorus of scholars who emphasize the need for developing brief, efficient self-report measures of psychological constructs with the goal of meeting researchers' contemporary needs (Ames et al., 2005; Donnellan et al., 2009; Gosling et al., 2003; Jonason & Webster, 2010; Rammstedt & John, 2007; Robins et al., 2001; Widaman et al., 2011). The BAQ may provide new opportunities to assess aggression in novel settings while decreasing participant fatigue. In so doing, researchers may continue to unravel the mystery of individual differences in aggression and how its negative social consequences can be prevented. Nevertheless, when efficient measures are *not* necessary, we recommend researchers continue to use the 29-item BPAQ, which has shown its worth as a reliable and valid measure of aggression both in our studies and in over two decades of prior research.

Despite the consistency of the results, there are some limitations of the current studies. First, the samples in the research consisted of young American college students with a preponderance of women and non-Hispanic Whites or European Americans. Although this limitation has no bearing on the BAQ's psychometric properties, it reduces the external validity of some findings. Moreover, gender affected neither the BAQ's item-level properties as a set (Study 2) nor its ability to predict behavioral aggression (Study 5). The only noteworthy gender effects were not found in the BAQ's structure, but rather in levels of latent aggression endorsed by the BAQ's respondents, as reflected in the DSF of the IRT analyses (Study 3), which also had greater power to detect gender effects ($N = 1{,}790$) than other samples. Future studies should consider testing the BAQ's properties in samples with greater diversity in age, nationality, education, race, ethnicity, and socioeconomic background.

Second, regarding reliability, the BPAQ-SF and its subscales had better internal consistency reliabilities (αs) than the BAQ and its subscales, except for physical aggression. Because α is a function of the number of items and the mean inter-item correlation (MIC), and because both the BAQ and BPAQ-SF have three items per subscale, we know the MICs were lower for the BAQ than the BPAQ-SF. MICs can be inflated by having redundant items and diminished by having a broader item pool. We sought to create a brief aggression measure that was efficient without sacrificing breadth, and that included a reverse-scored item. For example, the BAQ's Hostility subscale draws on both resentment *and* suspicion items from the BPAQ (and BDHI; see Tables I and IV) and the BAQ's Anger subscale includes a reverse-scored item; the BPAQ-SF does neither. Thus, given the BAQ's greater attention to breadth and valance, it is not surprising that it has lower but still acceptable αs (vs. the BPAQ-SF). What the BAQ may have lacked in internal consistency, it made up for in test–retest reliability and convergent validity. Similar trade-offs among breadth, brevity, and reliability can be found in two-item personality measures (see John & Soto, 2007).

Third, although we showed that the BAQ's Physical Aggression subscale correlated with a behavioral aggression measure in Study 5, additional research needs to be done to establish the predictive validity of the BAQ's other three subscales. For example, other laboratory experiments could seek to manipulate or measure behavioral correlates of the Verbal Aggression subscale (e.g., intensity and duration of shouting at another person, shouting latency in a verbal argument, etc.). Researchers could also examine the extent to which the BAQ's subscales correlate with indirect or displaced aggression (e.g., Denson, Pedersen, & Miller, 2006). Together with self-reports, peer reports of aggression might provide an additional way to further establish the BAQ's validity (e.g., relatives, teachers, friends, roommates, romantic partners).

For aggression research to flourish, researchers require measures that meet their current needs. Existing self-report aggression measures are either too long or lack psychometric rigor, making it desirable to develop a brief self-report aggression measure with an unrivaled body of supporting evidence. The current research meets this need with the development of the BAQ. This new self-report measure can facilitate greater amounts of aggression research, which may in turn assist in understanding why people behave aggressively and how such aggression can be prevented.

REFERENCES

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interaction*. Thousand Oaks, CA: Sage.

Ames, D. R., Rose, P., & Anderson, C. P. (2005). The NPI-16 as a short measure of narcissism. *Journal of Research in Personality*, *40*, 440–450. DOI: 10.1016/j.jrp.2005.03.002

Anderson, C. A., & Bushman, B. J. (1997). External validity of "trivial" experiments: The case of laboratory aggression. *Review of General Psychology*, *1*, 19–41. DOI: 10.1037/1089-2680.1.1.19

Archer, J. (2004). Sex differences in aggression in real-world settings: A meta-analytic review. *Review of General Psychology*, *8*, 291–322. DOI: 10.1037/1089-2680.8.4.291

Archer, J., & Mehdikhani, M. (2003). Variability among males in sexually selected attributes. *Review of General Psychology*, *7*, 219–236. DOI: 10.1037/1089-2680.7.3.219

Barlett, C. P., & Anderson, C. A. (2012). Direct and indirect relations between the Big 5 personality traits and violent behavior. *Personality and Individual Differences*, *52*, 870–875. DOI: 10.1016/j.paid.2012.01.029

Bettencourt, B. A., & Miller, N. (1996). Gender differences in aggression as a function of provocation: A meta-analysis. *Psychological Bulletin*, *119*, 422–447. DOI: 10.1037/0033-2909.119.3.422

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Bryant, F. B., & Smith, B. D. (2001). Refining the architecture of aggression: A measurement model for the Buss–Perry Aggression Questionnaire. *Journal of Research in Personality*, *35*, 138–167. DOI: 10.1006/jrpe.2000.2302

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data? *Perspectives on Psychological Science*, *6*, 3–5. DOI: 10.1177/1745691610393980

Bushman, B. J., Cooper, H. M., & Lemke, K. M. (1991). Meta-analysis of factor analyses: An illustration using the Buss–Durkee Hostility Inventory. *Personality and Social Psychology Bulletin*, *17*, 344–349. DOI: 10.1177/0146167291173015

Bushman, B. J., & Huesmann, L. R. (2010). Aggression. In S. T. Fiske, D. T. Gilbert, & L. Gardner (Eds.), *The handbook of social psychology* (5th ed., pp. 833–863). Hoboken, NJ: Wiley.

Bushman, B. J., Ridge, R. D., Das, E., Key, C. W., & Busath, G. L. (2007). When god sanctions killing: Effect of scriptural violence on aggression. *Psychological Science*, *18*, 204–207. DOI: 10.1111/j.1467-9280.2007.01873.x

Buss, A. H., & Durkee, A. (1957). An inventory for assessing different kinds of hostility. *Journal of Consulting Psychology*, *21*, 343–349. DOI: 10.1037/h0046900

Buss, A. H., & Perry, M. (1992). The Aggression Questionnaire. *Journal of Personality and Social Psychology*, *63*, 452–459. DOI: 10.1037/0022-3514.63.3.452

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin*, *56*, 81–105. DOI: 10.1037/h0046016

Carnagey, N. L., & Anderson, C. A. (2005). The effects of reward and punishment in violent video games on aggressive affect, cognition, and behavior. *Psychological Science*, *16*, 882–889. DOI: 10.1111/j.1467-9280.2005.01632.x

Dawes, J. (2008). Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point, and 10-point scales. *International Journal of Market Research*, *50*, 61–77.

Denson, T. F., Pedersen, W. C., & Miller, N. (2006). The Displaced Aggression Questionnaire. *Journal of Personality and Social Psychology*, *90*, 1032–1051. DOI: 10.1037/0022-3514.90.6.1032

DeWall, C. N., Anderson, C. A., & Bushman, B. J. (2011). The General Aggression Model: Theoretical extensions to violence. *Psychology of Violence*, *1*, 245–258. DOI: 10.1037/a0023842

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2009). The mini-IPIP scales: Tiny-yet-effective measures of the Big Five factors of personality. *Psychological Assessment*, *18*, 192–203. DOI: 10.1037/1040-3590.18.2.192

Eagly, A. H., & Steffen, V. J. (1986). Gender and aggressive behavior: A meta-analytic review of the social psychological literature. *Psychological Bulletin*, *100*, 309–330. DOI: 10.1037/0033-2909.100.3.309

Fraley, R. C., Waller, N. G., & Brennan, K. A. (2000). An item response theory analysis of self-report measures of adult attachment. *Journal of Personality and Social Psychology*, *78*, 350–365. DOI: 10.1037/0022-3514.78.2.350

Giancola, P. R., & Chermack, S. T. (1998). Construct validity of laboratory aggression paradigms: A response to Tedeschi and Quigley (1996). *Aggression and Violent Behavior*, *4*, 237–253. DOI: 10.1016/S1359-1789(97)00004-9

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. Jr. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, *37*, 504–528. DOI: 10.1016/S0092-6566(03)00046-1

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria verse new alternatives. *Structural Equation Modeling*, *6*, 1–55. DOI: 10.1080/10705519909540118

John, O. P., & Soto, C. J. (2007). On the importance of being valid: Reliability and the process of scale construction. In R. W. Robins, R. C. Fraley, & R. F. Kruger (Eds.), *Handbook of research methods in personality psychology* (pp. 461–494). New York: Guilford.

John, O. P., & Srivastava, S. (1999). The Big-Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin, & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 102–138). New York: Guildford Press.

Jonason, P. K., & Webster, G. D. (2010). The Dirty Dozen: A concise measure of the Dark Triad. *Psychological Assessment*, *22*, 420–432. DOI: 10.1037/a0019265

Kernis, M. H., Grannemann, B. D., & Barclay, L. C. (1989). Stability and level of self-esteem as predictors of anger arousal and hostility. *Journal of Personality and Social Psychology*, *56*, 1013–1022. DOI: 10.1037/0022-3514.56.6.1013

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York: Guilford.

Kreft, I. G. G., & de Leeuw, J. (1998). *Introducing multilevel modeling*. Thousand Oaks, CA: Sage.

MacKinnon, D. P., Krull, J. L., & Lockwood, C. M. (2000). Equivalence of the mediation, confounding, and suppression effect. *Prevention Science*, *1*, 173–181. DOI: 10.1023/A:1026595011371

Miller, J. D., Few, L. R., Seibert, L. A., Watts, A., Zeichner, A., & Lynam, D. R. (2012). An examination of the Dirty Dozen measure of psychopathy: A cautionary talk about the costs of brief measures. *Psychological Assessment*, *24*, 1048–1053. DOI: 10.1037/a0028583

Morizot, J., Ainsworth, A. T., & Reise, S. P. (2007). Toward modern psychometrics: Application of item response theory models in personality research. In R. W. Robins, R. C. Fraley, & R. F. Kruger (Eds.), *Handbook of research methods in personality psychology* (pp. 407–423). New York: Guilford.

Nezlek, J. B. (2008). An introduction to multilevel modeling for social and personality psychology. *Social and Personality Psychology Compass*, *2*, 824–860. DOI: 10.1111/j.1751-9004.2007.00059.x

Nezlek, J. B. (2011). *Multilevel modeling for social and personality psychology.* Thousand Oaks, CA: Sage.

Nichols, A. L., & Maner, J. K. (2008). The good-subject effect: Investigating participant demand characteristics. *Journal of General Psychology*, *135*, 151–165. DOI: 10.3200/GENP.135.2.151-166

Nichols, A. L., & Webster, G. D. (2013). The single-item need to belong scale. *Personality and Individual Differences*, *55*, 189–192. DOI: 10.1016/j.paid.2013.02.018

Orne, M. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, *17*, 776–783. DOI: 10.1037/h0043424

Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. In H. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 67–88). Hillsdale, NJ: Erlbaum.

Paulhus, D. L., & Vazire, S. (2007). The self-report method. In R. W. Robins, R. C. Fraley, & R. F. Kruger (Eds.), *Handbook of research methods in personality psychology* (pp. 224–239). New York: Guilford.

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*, *41*, 203–212. DOI: 10.1016/j.jrp.2006.02.001

Raskin, R., & Terry, H. (1998). A principal-components analysis of the narcissistic personality inventory and further evidence of its construct validity. *Journal of Personality and Social Psychology*, *54*, 890–902. DOI: 10.1037/0022-3514.54.5.890

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage.

Robins, R. W., Hendin, H. M., & Trzesniewski, K. H. (2001). Measuring global self-esteem: Construct validation of a single-item measure and the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin*, *27*, 151–161. DOI: 10.1177/0146167201272002

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, *8*, 350–353.

Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. *Psychological Assessment*, *12*, 102–111. DOI: 10.1037//1040-3590.12.1.102

Taylor, S. P. (1967). Aggressive behavior and physiological arousal as a function of provocation and the tendency to inhibit aggression. *Journal of Personality*, *35*, 297–310. DOI: 10.1111/j.1467-6494.1967.tb01430.x

Webster, G. D. (2006). Low self-esteem *is* related to aggression, but especially when controlling for gender: A replication and extension of Donnellan et al. (2005). *Representative Research in Social Psychology*, *29*, 12–18.

Webster, G. D. (2007). Is the relationship between self-esteem and Physical Aggression necessarily U-shaped? *Journal of Research in Personality*, *41*, 977–982. DOI: 10.1016/j.jrp.2007.01.001

Webster, G. D., & Bryan, A. D. (2007). Sociosexual attitudes and behaviors: Why two factors are better than one. *Journal of Research in Personality*, *41*, 917–922. DOI: 10.1016/j.jrp.2006.08.007

Webster, G. D., & Crysel, L. C. (2012). "Hit me, maybe, one more time": Brief measures of Impulsivity and Sensation Seeking and their prediction of blackjack bets and sexual promiscuity. *Journal of Research in Personality*, *46*, 591–598. DOI: 10.1016/j.jrp.2012.07.001

Webster, G. D., & Jonason, P. K. (2013). Putting the "IRT" in "Dirty": Item response theory analyses of the Dark Triad Dirty Dozen—An efficient measure of narcissism, psychopathy, and Machiavellianism. *Personality and Individual Differences*, *54*, 302–306. DOI: 10.1016/j.paid.2012.08.027

Webster, G. D., Kirkpatrick, L. A., Nezlek, J. B., Smith, C. V., & Paddock, E. L. (2007). Different slopes for different folks: Self-esteem instability and gender as moderators of the relationship between self-esteem and attitudinal aggression. *Self and Identity*, *6*, 74–94. DOI: 10.1080/15298860600920488

Widaman, K. F., Little, T. D., Preacher, K. J., & Sawalani, G. M. (2011). On creating and using short forms of scales in secondary research. In K. H. Trzesniewski, M. B. Donnellan, & R. E. Lucas (Eds.), *Secondary data analysis: An introduction for psychologists* (pp. 39–62). Washington, DC: American Psychological Association.

Zuckerman, M., Kuhlman, D. M., Joireman, J., Teta, P., & Kraft, M. (1993). A comparison of three structural models for personality: The big three, the big five, and the alternative five. *Journal of Personality and Social Psychology*, *65*, 757–768. DOI: 10.1037/0022-3514.65.4.757

Supporting Information

Additional Supporting Information may be found in the online version of this article.