Journal of Personality Assessment, *97*(6), 638–649, 2015

ISSN: 0022-3891 print / 1532-7752 online
DOI: 10.1080/00223891.2015.1044093

Routledge
Taylor & Francis Group

The Brief Aggression Questionnaire: Structure, Validity, Reliability, and Generalizability

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,[7] Benjamin S. Crosier,[8] C. Veronica Smith,[9] E. Layne Paddock,[10] John B. Nezlek,[11,12] Lee A. Kirkpatrick,[11] Angela D. Bryan,[13] and Renée J. Bator[14]

[1]*Department of Psychology, University of Florida*
[2]*Department of Psychology, University of Kentucky*
[3]*Department of Psychology, University of North Carolina at Wilmington*
[4]*Department of Psychology, University of Western Sydney, Australia*
[5]*Department of Psychology, University of Toronto, Canada*
[6]*Department of Business, University of Navarra, Pamplona, Navarra, Spain*
[7]*Department of Psychology, Stetson University*
[8]*Center for Technology and Behavioral Health, Dartmouth College*
[9]*Department of Psychology, University of Mississippi*
[10]*Chair of Work and Organizational Psychology, ETH-Zürich, Switzerland*
[11]*Department of Psychology, College of William and Mary*
[12]*University of Social Sciences and Humanities, Poznań, Poland*
[13]*Department of Psychology and Neuroscience, University of Colorado, Boulder*
[14]*Department of Psychology, State University of New York at Plattsburgh*

In contexts that increasingly demand brief self-report measures (e.g., experience sampling, longitudinal and field studies), researchers seek succinct surveys that maintain reliability and validity. One such measure is the 12-item Brief Aggression Questionnaire (BAQ; Webster et al., 2014), which uses 4 3-item subscales: Physical Aggression, Verbal Aggression, Anger, and Hostility. Although prior work suggests the BAQ's scores are reliable and valid, we addressed some lingering concerns. Across 3 studies (N = 1,279), we found that the BAQ had a 4-factor structure, possessed long-term test–retest reliability across 12 weeks, predicted differences in behavioral aggression over time in a laboratory experiment, generalized to a diverse nonstudent sample, and showed convergent validity with a displaced aggression measure. In addition, the BAQ's 3-item Anger subscale showed convergent validity with a trait anger measure. We discuss the BAQ's potential reliability, validity, limitations, and uses as an efficient measure of aggressive traits.

The reliability and validity of new measures must be tested rigorously and repeatedly if they are to be adopted by researchers. The case for brief self-report measures of aggression is no different. Webster et al. (2014) developed the 12-item Brief Aggression Questionnaire (BAQ) as a more efficient alternative to the 29-item Aggression Questionnaire (BPAQ; Buss & Perry, 1992). The BAQ uses the three highest loading items from each of the BPAQ's four subscales: Physical Aggression, Verbal Aggression, Anger, and Hostility. In five studies ($N \approx 4{,}000$), the BAQ was found to have (a) theoretically consistent patterns of convergent and discriminant validity with other self-report measures, (b) a four-factor structure using multiple factor analyses, (c) adequate information recovery using item response theory, (d) stable test–retest reliability across 3 weeks, and (e) convergent validity with behavioral measures of aggression (Webster et al., 2014). Although we recommend using the 29-item BPAQ in situations where time permits, we also believe that researchers face an increasing demand for efficient measures such as the BAQ in specific settings that require them, including experience sampling studies, daily diary studies, prescreening or mass-testing studies, longitudinal studies, field studies, and studies with special populations (see Widaman, Little, Preacher, & Sawalani, 2011). In addition, brief measures can help reduce respondent fatigue and inattentiveness. Thus, when used in conjunction with several other long-format questionnaires, the full 29-item BPAQ might add unnecessary items to a burgeoning item count that can become overly burdensome to respondents.

Although there is a clear trade-off between reliability and efficiency regarding the number of items per construct when creating brief measures, the BAQ uses three items per construct for three reasons. First, confirmatory factor analyses (CFAs) and item response theory (IRT) analyses found that the 12-item BAQ can efficiently recover test information about four latent aggressive traits with only three items per construct (Webster et al., 2014; see also Bryant & Smith, 2001). Second, because the BAQ sought to preserve the BPAQ's four-factor structure, including four or five items per construct would have needlessly ballooned the total number of

Received August 13, 2013; Revised February 21, 2015.
Address correspondence to Gregory D. Webster, Department of Psychology, University of Florida, P.O. Box 112250, Gainesville, FL 32611-2250; Email: gdwebs@ufl.edu

items by a factor of four, thus defeating the purpose of creating an efficient measure (i.e., 12 vs. 16 vs. 20 items out of 29). Third, three items per construct are often a necessary minimum for model identification and convergence when testing structural equation models (SEMs; Kline, 2011).

Despite these advances, the BAQ has a least four key limitations. First, prior assessments of the BAQ's structure have relied solely on principal axis factoring (PAF) and confirmatory factor analysis (CFA; Webster et al., 2014) without first presenting an exploratory factor analysis (EFA), which is often an initial step in scale construction to determine factor structure and assess item–factor pairings (Fabrigar, Wegner, MacCallum, & Strahan, 1999). Consequently, we present the first EFA of the BAQ's structure (Study 1). Second, because the BAQ has shown acceptable test–retest reliability for only a short time interval (3 weeks; Webster et al., 2014), we sought to address this concern by assessing the BAQ's test–retest reliability for a longer time interval (12 weeks; Study 1). Third, although the BAQ's Physical Aggression subscale relates positively to behavioral aggression (noise blasts in an ostensibly competitive two-person game; Webster et al., 2014), it remains unknown whether the BAQ relates to the time course of aggressive responding (noise blasts across 25 trials in the same game; Study 2). Specifically, we expect a Person (trait) × Situation (aggressive retaliation over time) interaction. At Trial 1, the BAQ should positively predict behavioral aggression (noise blasts) because the trait influence should be strongest when situation is weak (retaliation from the participant's ostensible partner has not yet occurred). By Trial 25, the BAQ should less reliably predict behavioral aggression because the trait influence should become comparatively weaker over time as the situation (aggressive retaliation across trials) grows stronger. Fourth, although the BAQ has shown acceptable psychometric properties in samples of U.S. undergraduates (Webster et al., 2014), its generalizability to more diverse, nonstudent samples remains unknown. In Study 3, we address this limitation by surveying a large and diverse international sample with a broader age range. In addition, we strove to expand the nomological network of the BAQ by examining its convergent and discriminant validity with trait anger and displaced aggression (Study 3).

Thus, whereas prior research established and justified item selection for the 12-item BAQ along with gender differences (Webster et al., 2014), this research focuses on addressing the limitations already listed and expanding the BAQ's validity and generalizability. In addition, given psychological science's renewed emphasis on replication and reproducibility (see Pashler & Wagenmakers's [2012] overview), we believe that replicating the BAQ's reliability and factor structure while addressing some of its lingering limitations is both necessary and important.

Study 1: Factor Structure and Test–Retest Reliability

The goals of Study 1 were twofold. First, we aimed to replicate and extend prior results regarding the BAQ's four-factor structure (Webster et al., 2014). Whereas prior studies have relied on PAF and CFA, Study 1 focuses on EFA as a necessary step in assessing structure in scale construction (Fabrigar et al., 1999). We also used multiple criteria to establish the plausibility of a four-factor BAQ model. Second, we sought to extend the BAQ's test–retest reliability. Establishing acceptable test–retest reliability is essential to developing new or brief scales because trait-level individual differences should be relatively stable over time. We measured the BAQ at two time points 12 weeks apart, which allowed us to test longer term test–retest reliability. Although prior research established the BAQ's test–retest reliability across 3 weeks (Webster et al., 2014), initially promising results could be due to memory biases or carryover effects characteristic of short time periods.

Method

Measures. To test factor structure in Study 1, we aggregated BAQ data from two independent samples to achieve a sufficient sample size (Samples 1 and 2 described later). Specifically, we sought a > 20:1 cases-to-items ratio, which is important for achieving stable estimates (e.g., Kline, 2013; but also see MacCallum, Widman, Zhang, & Hong, 1999). In both samples, participants responded to the 12 BAQ items using a 7-point scale ranging from 1 (*extremely uncharacteristic of me*) to 7 (*extremely characteristic of me*).

Sample 1. Participants were 125 undergraduates (56 men, 58 women, 11 did not report gender) enrolled in introductory psychology courses at a public university in Virginia who received course credit for their participation in an online questionnaire (ages: 18–22 years, $M = 19.10$, $SD = 1.22$). Regarding race and ethnicity, the sample was 77% White (non-Hispanic), 7% Asian American or Pacific Islander, 7% Black or African American, 3% Hispanic, and 4% other races or ethnicities. BAQ descriptive statistics and correlations appear in Table 1.

Table 1.—Brief Aggression Questionnaire (BAQ) descriptive statistics and zero-order correlations of observed scores for Study 1, Sample 1 (below diagonal) and Study 2 (above diagonal).

	Study 1 ($N = 125$)			Zero-Order Correlations					Study 2 ($N = 307$)		
BAQ Measure	*M*	*SD*	α	1	2	3	4	5	*M*	*SD*	α
1. Physical aggression	3.03	1.76	.84	—	.43	.28	.27	.78	2.75	1.65	.83
2. Verbal aggression	3.84	1.31	.66	.54	—	.31	.19	.69	3.56	1.24	.62
3. Anger	2.71	1.39	.81	.40	.51	—	.36	.66	2.31	1.16	.67
4. Hostility	3.07	1.35	.74	.37	.35	.48	—	.63	2.36	1.18	.65
5. BAQ mean	3.16	1.11	.86	.79	.78	.77	.70	—	2.74	0.91	.79

Note. All correlations significant at $p < .01$.

Sample 2. Participants were a convenience sample of 140 undergraduates enrolled in psychology classes at a public university in Florida. Each participant was asked to complete a paper version of the 12-item BAQ in class twice—12 weeks apart. We chose a 12-week interval for convenience because it corresponded to the second and penultimate weeks in a semester, and because it was long enough to avoid possible carryover effects associated with prior testing. Of the 140 participants, 130 (93%) and 123 (88%) completed questionnaires during Weeks 1 and 12, respectively; 113 (81%) participants completed both sessions. We used the sample from Week 1 for the EFA because it was larger than the sample from Week 12. Of the 113 participants recruited for the test–retest reliability analysis, 88 were women and 25 were men; ages ranged from 18 to 29 years ($M = 20.31$, $SD = 1.54$). Race or ethnicity information was not collected for this sample.

Results and Discussion

Factor structure. To assess factor structure while exceeding a 20:1 cases-to-items ratio, we aggregated Sample 1 ($n = 125$) and Sample 2, Week 1 ($n = 130$) for the EFA ($N = 255$). Using M*plus* 6.1 (Muthén & Muthén, 2010), we specified an EFA with up to four factors using the default oblique geomin rotation and full maximum likelihood estimation. The EFA procedure estimated models with one to four factors; fit indexes appear in Table 2. As expected, model fit improved significantly with each additional factor (via $\Delta\chi^2$), and only the four-factor model yielded "good" fit indexes (i.e., comparative fit index [CFI] and Tucker–Lewis Index [TLI] $\geq .95$; root mean square error of approximation [RMSEA] and standardized root mean square residual [SRMR] $\leq .05$).

We also assessed the appropriateness of the BAQ's expected four-factor structure using multiple methods because each one has strengths and weaknesses. First, as stated earlier, going from three to four factors produced fit indexes widely considered to be in the acceptable range. Second, using a scree plot, the eigenvalues > 1.0 criterion (i.e., Kaiser–Guttman criterion [Guttman, 1954; Kaiser, 1960]) also suggested a four-factor solution (Figure 1). In contrast, parallel analysis (Horn, 1965) suggested a three-factor solution. Parallel analysis assumes eigenvalues from a random matrix with the same number of items and sample size as the observed eigenvalues (see Hayton, Allen, & Scarpello, 2004).

TABLE 2.—Study 1: Exploratory factor analysis results for the Brief Aggression Questionnaire.

Models or Differences	χ^2	*df*	CFI	TLI	RMSEA	90% CI LL	90% CI UL	SRMR
1 factor	439.54	54	.656	.580	.167	.153	.182	.105
2 factors	210.46	43	.851	.771	.124	.107	.141	.075
2 vs. 1 difference	229.08	11						
3 Factors	121.42	33	.921	.842	.103	.083	.122	.052
3 vs. 2 difference	89.04	10						
4 Factors	39.29	24	.986	.963	.050^{ns}	.018	.077	.021
4 vs. 3 difference	82.13	9						

Note. $N = 255$. Fit indexes and suggested acceptable fit thresholds: Comparative fit index (CFI) and Tucker–Lewis Index (TLI): > 90. Root mean square error of approximation (RMSEA) and standardized root mean square residual (SRMR): < .08. LL and UL = lower and upper limits for RMSEA. All χ^2 and RMSEA statistics were significant at $p < .05$ except *ns*.

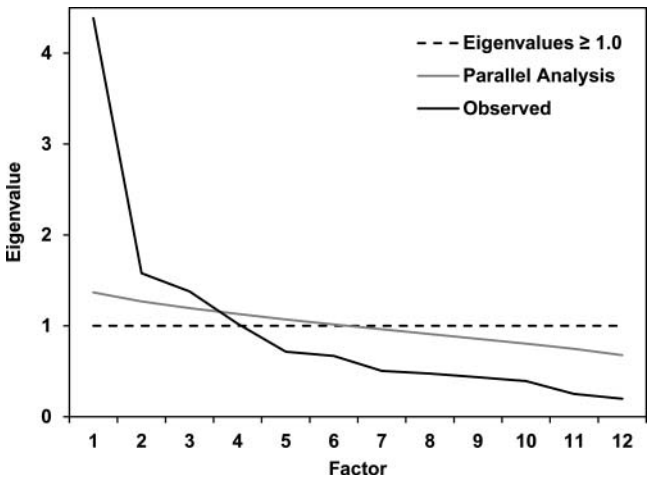

FIGURE 1.—Scree plots of eigenvalues by number of factors: Observed and two threshold criteria.

Finally, we also used regression-based iterative outlier analyses to identify eigenvalues that departed significantly from linearity in the scree plot. This involved a series of simple regressions in which we regressed eigenvalues onto a number of factors while using established methods to identify the largest outlier (via Studentized deleted residual [SDR] and Cook's *D*; see Judd, McClelland, & Ryan, 2009, pp. 301–305), remove it, and then rerun the model without the largest outlier. The stopping rule was the absence of outliers (both SDR and Cook's *D*). After five iterations, outliers became absent, thereby suggesting a purely linear relationship between eigenvalues and number of factors, and thus supporting a four-factor model (Table 3; Figure 1). To summarize, a parallel analysis supported a three-factor model, and, showing some consensus, three methods—fit index thresholds, eigenvalues ≥ 1.0, and iterative outlier analyses—supported a four-factor model.

Table 4 shows the factor structure matrix from the four-factor EFA. All items loaded at .50 or greater on their expected factors with one exception: "My friends say that I'm somewhat argumentative" loaded more strongly on Anger (.54) than its predicted factor, Verbal Aggression (.44). Thus, with one exception, the BAQ items loaded on the four factors related to their respective constructs: Physical Aggression, Verbal Aggression, Anger, and Hostility. Nevertheless, our moderate sample size ($N \approx 250$) might have contributed to this discrepancy. We revisit and address this concern in Study 3, where we use CFAs with larger samples (*N*s > 500; see

TABLE 3.—Study 1: Iterative outlier analyses of eigenvalues regressed on number of factors.

Factor	Studentized Deleted Residual (*t*)	Cook's *D*
1	14.83*	2.01[a]
2	2.13*	0.76
3	4.23*	1.52[a]
4	5.35*	1.76[a]
5	0.19	0.02

[a]Cook's *D*s ≥ 1.0 are considered outliers.
*$p < .05$, one-tailed.

TABLE 4.—Study 1: Results of exploratory factor analysis (EFA) for four factors.

Brief Aggression Questionnaire Subscales and Items	Factor 1	2	3	4
Physical aggression				
2. Given enough provocation, I may hit another person.	**.96**	.42	.32	.35
5. If I have to resort to violence to protect my rights, I will.	**.76**	.35	.30	.37
6. There are people who pushed me so far that we came to blows.	**.60**	.40	.21	.46
Anger				
4. I am an even-tempered person.[a]	.11	**.52**	–.04	.02
6. Sometimes I fly off the handle for no good reason.	.39	**.86**	.14	.47
7. I have trouble controlling my temper.	.48	**.90**	.16	.45
Verbal aggression				
1. I tell my friends openly when I disagree with them.	.24	.06	**.82**	.05
3. When people annoy me, I may tell them what I think of them.	.36	.37	**.54**	.24
5. My friends say that I'm somewhat argumentative.	.42	**.54**	.44	.28
Hostility				
3. Other people always seem to get the breaks.	.25	.33	.04	**.79**
7. I sometimes feel that people are laughing at me behind my back.	.28	.49	–.09	**.54**
8. When people are especially nice, I wonder what they want.	.30	.33	.12	**.60**

Note. $N = 255$. Factor loadings > .50 are shown in bold. Oblique rotation was used.
[a]Reverse-scored item.

also Webster et al., 2014) while also exploring the BAQ's generalizability.

Test–retest reliability. Descriptive statistics and test–retest correlations for the BAQ at both time points appear in Table 5. As expected, test–retest reliability correlations were strong and significant, ranging from .68 to .80 among the four subscales, and as high as .81 for the BAQ mean. Overall, these findings show that the BAQ has good test–retest reliability even over a longer time interval of 12 weeks, suggesting that it measures stable aggressive traits.

Because traditional test–retest correlations can confound temporal reliability with measurement error (Watson, 2004), we also examined latent test–retest correlations using an autoregressive latent-variable test–retest reliability model, which separated measurement error from temporal reliability (see Khoo, West, Wu, & Kwok, 2006, pp. 305–307; see also Nichols & Webster, 2015). This involved modeling latent variables for the 12-item BAQ for each of its four three-item subscales at Times 1 and 2, allowing for equal loadings and correlated residuals for the same items at different time points, and examining the correlation associated with regressing Time 2's latent variable onto Time 1's latent variable. We tested these models as a series of CFAs in M*plus* 6.1 (Muthén & Muthén, 2010) using full maximum likelihood estimation. These CFAs showed that the latent BAQ and its latent subscales had strong and significant test–retest reliability correlations ranging from .83 to .90 (Table 6). In addition, the models fit the data well (Table 6), except for the 12-item latent BAQ, which would typically be modeled with a four-factor approach (vs. a unidimensional one) in most CFA or SEM contexts (see Study 3). Thus, modeling measurement error improved the temporal reliability of the BAQ and its subscales.

TABLE 5.—Study 1, Sample 2: Descriptive statistics and zero-order correlations of observed scores for the Brief Aggression Questionnaire.

	Descriptive Statistics			Time 1					Time 2			
	M	*SD*	α	1	2	3	4	5	6	7	8	9
Time 1												
1. Physical	2.14	1.17	.72									
2. Verbal	3.84	1.20	.66	.41								
3. Anger	2.19	0.95	.74	.29	.19							
4. Hostility	2.71	1.05	.57	.44	.04^{ns}	.27						
5. Mean	2.72	0.74	.76	.81	.65	.61	.63					
Time 2												
6. Physical	2.51	1.30	.76	**.80**	.51	.25	.25	.69				
7. Verbal	3.76	1.19	.73	.33	**.76**	.19	–.06^{ns}	.48	.46			
8. Anger	2.32	0.95	.72	.37	.34	**.74**	.20	.59	.43	.35		
9. Hostility	3.04	1.10	.64	.20	–.05^{ns}	.28	**.68**	.39	.07^{ns}	–.03^{ns}	.30	
10. Mean	2.90	0.76	.77	.66	.60	.51	.39	**.81**	.77	.68	.74	.47

Note. All correlations were significant at $p < .05$ except *ns*. $N = 113$. Twelve-week test–retest reliability correlations are shown in bold.

STUDY 2: PREDICTIVE VALIDITY WITH BEHAVIORAL AGGRESSION OVER TIME

Having found some additional support for the BAQ's four-factor structure and evidence of good test–retest reliability (Study 1), we next addressed questions about the BAQ's predictive validity regarding behavioral aggression over time. In Study 2, we reanalyzed data from Webster et al.'s (2014)

TABLE 6.—Study 1: Latent test–retest reliabilities (r_{xx}) and fit indexes based on confirmatory factor analyses for the Brief Aggression Questionnaire (BAQ).

Models	r_{xx}	χ^2	*df*	CFI	TLI	RMSEA	90% CI LL	90% CI UL	SRMR
Physical Aggression	.88	5.44	7	1.0	1.0	.000	.000	.089	.034
Verbal Aggression	.90	5.47	7	1.0	1.0	.000	.000	.089	.038
Anger	.89	6.45	7	1.0	1.0	.000	.000	.099	.050
Hostility	.83	11.17	7	.979	.954	.065	.000	.133	.054
BAQ (12 items)	.87	591.38*	250	.757	.732	.099*	.089	.109	.131

Note. $N = 140$. Fit indexes and suggested acceptable fit thresholds. Comparative fit index (CFI) and Tucker–Lewis Index (TLI): > 90. Root mean square error of approximation (RMSEA) and standardized root mean square residual (SRMR): < .08. LL and UL = lower and upper limits for RMSEA.
*$p < .05$.

Study 5, in which participants completed the 29-item BPAQ before taking part in a 25-trial competitive reaction-time computer game, in which the winners of each trial (randomly assigned to win 13 of 25 trials) got to blast the losers (ostensibly in an adjacent room) with bursts of white noise. Although Webster et al.'s Study 5 showed that the BAQ's Physical Aggression subscale performed as well as or slightly better than that of the BPAQ, change over time—or across trials—was not examined. In Study 2, we predicted a Time × BAQ interaction for noise blast duration and intensity: Trait BAQ scores should positively predict aggressive behavior more strongly at Trial 1 than at Trial 25. In other words, we predicted a classic Person × Situation interaction (Funder, 2008; Krueger, 2009; Swann & Seyle, 2005; Webster, 2009): Trait effects will be strongest when the situation is weak (Trial 1), but weaken as the situation grows stronger (i.e., following perceived iterative aggressive retaliation; Trial 25). Decomposing this same interaction from another angle (see Aiken & West, 1991), we also predicted that people with high trait aggression (1 *SD* above the BAQ mean) would continue to respond aggressively regardless of retaliation over trials (i.e., stronger trait → less situational influence), whereas people with low trait aggression (1 *SD* below the BAQ mean) would increase their aggressive responding because of retaliation over trials (i.e., weaker trait → more situational influence). We remained agnostic, however, as to whether this interaction would be (a) stronger for noise blast duration or intensity, or (b) driven by any particular BAQ subscale.

Method

Participants. Participants were 307 undergraduates enrolled in introductory psychology courses at a public university in Florida who received course credit in exchange for their participation (91 men, 216 women; ages: 18–41 years, $M = 19.34$, $SD = 2.27$).

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

Procedure. The procedure was a modified Taylor (1967) aggression paradigm, where participants were ostensibly paired in adjacent rooms and competed against each other to be the quickest responder on each reaction-time trial; the winner of each trial could deliver a white noise blast to his or her partner. In reality, participants completed their reaction-time trials against a computer program set to mimic another person's actions. Participants controlled the duration (0.0–5.0 sec) and intensity (0–105 dB) of the noise (about the volume of a smoke alarm) corresponding to response scales ranging from 0 (*low*) to 10 (*high*), which was consistent with similar research using this paradigm with 10- or 11-point response scales (e.g., DeWall, Bushman, Giancola, & Webster, 2010; see Webster et al., 2014). Participants' noise-blast duration ($M = 4.92$, $SD = 2.20$) and intensity ($M = 5.18$, $SD = 2.26$) scores were positively correlated ($r = .86$, $p < .001$). This laboratory procedure provides a valid and established measure of behavioral aggression (e.g., Anderson & Bushman, 1997; Giancola & Chermack, 1998).

Multilevel modeling. We used the multilevel modeling program Hierarchical Linear Modeling 6.0 (HLM; Raudenbush, Bryk, Cheong, & Congdon, 2006) because multiple trials (25) were nested within each participant. Using HLM's default restricted maximum likelihood estimation and robust standard errors, multilevel modeling 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 BAQ. Specifically, Level 1 of our multilevel was

$$\text{Duration or Intensity}_{ti} = \pi_{0i} + \pi_{1i}(\text{Trial} - 13)_i + e_{ti},$$

where Duration or Intensity$_{ti}$ represents the noise blast duration or intensity (separate models) given out at time t by individual i. Each person's duration or intensity scores are modeled by π_{0i}, which represents the mean or intercept for each person at the midpoint of the 25 trials (Trial 13), and π_{1i}(Trial – 13)$_i$, which represents the change-over-time (or trial) slope for each person. The error term, e_{ti}, represents the residual Level-1 variance.

We modeled the Level-1 random-effects intercepts and slopes for each person simultaneously at Level 2 as a function of their respective BAQ score (grand-mean-centered):

$$\pi_{0i} = \beta_{00} + \beta_{01}(\text{BAQ} - \text{mean}) + r_{0i},$$

$$\pi_{1i} = \beta_{10} + \beta_{11}(\text{BAQ} - \text{mean}) + r_{1i}.$$

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 duration or intensity score (at Trial 13 and at the mean BAQ score). The β_{01}(BAQ – mean) coefficient represents the moderating effect of individual differences in the BAQ on people's overall mean duration or intensity score. In contrast, π_{1i} represents each person's duration- or intensity-over-time slope, and the β_{10} coefficient represents the average of these respective slopes (at the mean BAQ score). The focal effect, which is the β_{11}(BAQ – mean) coefficient, represents the moderating effect of the BAQ on people's change-over-time slopes in duration or intensity. The error terms, r_{0i} and r_{1i}, represent the residual Level-2 variances in people's intercepts and slopes, respectively.

Results and Discussion

See Table 1 for BAQ mean and subscale correlations and descriptive statistics. Average noise-blast duration was related to the BAQ, $\beta_{01} = 0.46$, $t(305) = 3.81$, $p < .001$, $r_p = .21$ (see also Webster et al., 2014). Regarding the focal analyses, noise-blast duration increased over time for the average participant, $\beta_{10} = 0.041$, $t(305) = 5.19$, $p < .001$, $r_p = .28$; however, the change-over-time slopes were unrelated to BAQ scores, $\beta_{11} = -0.0016$, $t(305) = -0.16$, $p = .87$, $r_p = -.01$.

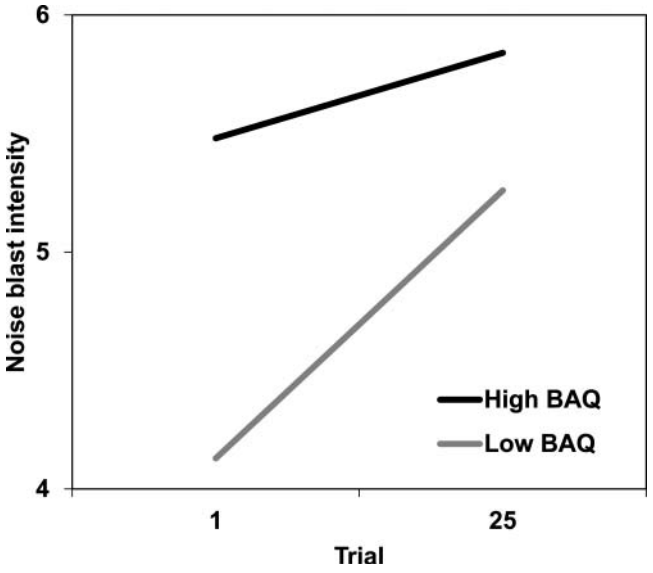

FIGURE 2.—Study 2: Noise-blast intensity as a function of trial (1–25) and mean score on the 12-item Brief Aggression Questionnaire (BAQ; Low = –1 *SD*, High = +1 *SD*).

Average noise-blast intensity was related to the BAQ mean, $\beta_{01} = 0.53$, $t(305) = 4.35$, $p < .001$, $r_p = .24$ (see also Webster et al., 2014). Regarding the focal analyses, noise-blast intensity increased over time for the average participant, $\beta_{10} = 0.031$, $t(305) = 3.88$, $p < .001$, $r_p = .22$; and the change-over-time slopes were negatively related to BAQ scores, $\beta_{11} = -0.018$, $t(305) = -1.96$, $p = .050$, $r_p = -.11$ (Figure 2). In addition, controlling for gender reduced this effect only slightly, $\beta_{11} = -0.017$, $t(304) = -1.86$, $p = .063$, $r_p = -.11$; and replacing the 12-item BAQ with the 29-item BPAQ showed that the full measure did not significantly moderate the change-over-time slopes, $\beta_{11} = -0.015$, $t(305) = -1.64$, $p = .101$, $r_p = -.09$. We also ran a follow-up model that replaced the BAQ mean with its four subscales; none was a significant predictor of noise blast intensity change (ps > .08, $|r_p|$s < .10). Simple effects tests (see Aiken & West, 1991) on the cross-level interaction in Figure 2 showed the predicted effect: At Trial 1, noise-blast intensity was positively and significantly related to the BAQ mean, $\beta_{01} = 0.74$, $t(305) = 4.62$, $p < .001$, $r_p = .26$; but by Trial 25, this was no longer the case, $\beta_{01} = 0.32$, $t(305) = 1.94$, $p = .053$, $r_p = .11$. On an exploratory basis, we also examined the other two simple effects: Noise-blast intensity increased significantly over time for participants scoring 1 *SD* below the mean, $\beta_{10} = 0.047$, $t(305) = 4.08$, $p < .001$, $r_p = .23$, but the same increase was not significant for participants scoring 1 *SD* above the mean, $\beta_{10} = 0.015$, $t(305) = 1.34$, $p = .18$, $r_p = .08$.

Thus, supporting our prediction and showing a classic Person × Situation interaction, the BAQ mean was sensitive to trait differences in aggression in predicting initial noise-blast intensity (but not duration), and this relationship waned across time (repeated trials) as people became immersed in a strong, competitive situation. From another perspective, people with high BAQ scores maintained their aggressive behavior (via high-intensity noise blasts) unabated throughout the 25 trials, whereas people with low BAQ scores were more susceptible to the demands of the situation, starting with lower intensity noise blasts, but then increasing them over 25 trials. Whereas prior research has examined the BAQ's association with mean levels of noise blast duration and intensity (averaged across 25 trials; see Webster et al., 2014), this research extends these findings to show that the BAQ—but not the BPAQ—predicted initial differences in behavioral aggression (noise blast intensity) and its increase over time across 25 trials in a controlled laboratory experiment.

STUDY 3: EXPANDED GENERALIZABILITY AND CONVERGENT VALIDITY

Although the BAQ's reliability and validity have received some support here (Studies 1 & 2) and elsewhere (Webster et al., 2014), it has yet to be determined the extent to which its psychometric properties generalize to more diverse, nonstudent samples. Specifically, nearly all participants to this point have been U.S. undergraduates, or more generally, "W.E.I.R.D." people (i.e., Western, educated, industrialized, rich, and democratic; Henrich, Heine, & Norenzayan, 2010). To address this and other concerns, we conducted Study 3 with three goals in mind.

First, we tested a series of CFAs to compare the fit of the BAQ's predicted four-factor model to a hierarchical model (four latent BAQ subscales load onto a second-order latent aggression factor) and to a unidimensional model (12 items load directly onto one latent aggression factor). Because the BAQ was developed to optimize subscale items (vs. mean score), we expected the unidimensional model to fit the data worse than other models.

Second, because we believe that the four-factor model will be widely adopted, we tested the extent to which gender (men vs. women), first-language (English vs. other), and sample (student vs. non-student) differences moderated the item factor loadings (metric invariance) and intercepts (scalar invariance). We did this because (a) men tend to report and enact more aggression—particularly unprovoked physical aggression—than women (Archer, 2004; Bettencourt & Miller, 1996; Eagly & Steffen, 1986), and (b) prior research using the BAQ had U.S. undergraduate participants whose first language was English. Nevertheless, we expected the BAQ to show partial metric and scalar invariance across these three group comparisons.

Third, we also sought to expand the nomological network of the BAQ and its subscales by examining its correlations with trait anger and displaced aggression. Trait anger is important to understanding the BAQ's Anger subscale, especially because prior research (Webster et al., 2014) stressed validating the BAQ's Physical Aggression subscale. Specifically, we expected the Trait Anger Scale (described later) would correlate more highly with the BAQ's Anger subscale than its other three subscales. We also measured a brief version of the three-factor Displaced Aggression Questionnaire (DAQ; described later). Testing relationships among the BAQ and DAQ (and their subscales) is important because doing so would show the first evidence of convergent validity between the two multifaceted trait aggression measures. Although we expected positive correlations among the BAQ and DAQ and their subscales,

we remained agnostic regarding the precise pattern of correlations among subscales.

Method

Participants and procedure. Participants were 611 people recruited using Amazon's Mechanical Turk (MTurk; see Buhrmester, Kwang, & Gosling, 2011; Goodman, Cryder, & Cheema, 2013; Hauser & Schwarz, in press). MTurk is a marketplace for work that brings together researchers and willing participants who wish to be paid for research (e.g., completing surveys, computerized experiments). MTurk has been widely adopted by psychology departments and business schools throughout the world. For example, by 2010, 16 of the top 30 of business schools in the United States (53%) were using MTurk to collect research data (Goodman et al., 2013). Many experiments pay participants between 5 cents and 50 cents. In Study 3, we paid participants 25 cents for completing an online survey. To see our survey, we required MTurk participants to have human intelligence task (HIT) approval rates $\geq$ 95% based on their prior performance on other tasks or studies (i.e., in previous tasks or studies, they completed the assigned task or study to the satisfaction of the researcher or requester); this assured some quality control. Indeed, MTurk participants with such HIT rates often outperform traditional participant pools in survey attentiveness (Hauser & Schwarz, in press). For additional information on MTurk and the demographics of its users, see Buhrmester et al. (2011).

Using three items designed to detect inattentive participants (e.g., "Please select option 2 for this item") interspersed throughout the online survey, we identified and excluded seven participants with inattentive response patterns. This left a sample of 604 people (336 men, 260 women, 4 transgender, 4 did not report gender; ages: 18–79 years, $M = 29.53$, $SD = 10.03$), representing more than 40 countries and 40 languages. The five most frequent countries were India (48.7%), the United States (31.3%), Canada (4.0%), Pakistan (2.6%), and the Philippines (1.3%); all other countries represented less than 1.0% of the sample. Although all participants could read English, 48% of them reported English as their first language. Ethnically, the sample was 94.0% non-Hispanic; racially, the sample was 48.8% Indian, 36.6% White, 4.3% unspecified Asian, 2.3% Hispanic, 2.2% Black, 2.2% Filipino, 1.7% unspecified, 1.0% Chinese, 0.5% American Indian, and 0.5% Japanese. Regarding education, 0.2% completed only primary or grammar school, 7.4% completed secondary or high school, 23.6% had some college or university; and 5.7%, 41.8%, 17.4%, and 3.8% had earned an associate's, bachelor's, master's, or doctoral or professional degree, respectively. Participants completed the online survey and demographic questions in a timely fashion, averaging 8.5 min ($\approx$ 10 sec per item).

Measures.

Brief Aggression Questionnaire: Participants completed the 12-item BAQ using a 7-point response scale ranging from 1 (*extremely uncharacteristic of me*) to 7 (*extremely characteristic of me*).

Trait Anger Scale: Participants also provided data on the 15-item Trait Anger Scale (Spielberger, Jacobs, Russell, & Crane, 1983; e.g., "I have a fiery temper") using a 4-point scale ranging from 1 (*almost never*) to 4 (*almost always*).

Displaced Aggression Questionnaire: Participants also completed an abbreviated, nine-item measure of displaced aggression based on items drawn from the DAQ (Denson, Pedersen, & Miller, 2006). Just as Webster et al. (2014) chose the highest loading items from the BPAQ's subscales to make the BAQ, we chose the three highest loading items (reported in Denson et al., 2006) from each of the DAQ's three subscales—Angry Rumination, Revenge Planning, and Displaced Aggression—to make an abbreviated measure. The DAQ's three factor-based subscales measure theoretically informative dimensions of displaced aggression: affective (Angry Rumination), cognitive (Revenge Planning), and behavioral (Displaced Aggression). We chose to measure displaced aggression because it should be positively correlated with trait aggression as reflected in the BAQ; however, we remained agnostic regarding relationships among specific subscales.

Results and Discussion

Confirmatory factor analyses. We first ran a series of CFAs in M*plus* 6.1 (Muthén & Muthén, 2010) using full maximum likelihood estimation to test the factor structure of the BAQ using our nonstudent sample. The fit statistics for all three measurement models (CFAs) appear in Table 7. The four-factor model fit the data well and better than each of the other two measurement models. Thus, comparatively simpler

TABLE 7.—Study 3: Confirmatory factor analysis results for the Brief Aggression Questionnaire.

Models or Differences	χ^2	*df*	CFI	TLI	RMSEA	90% CI LL	90% CI UL	SRMR
Measurement models								
1. Four-factor	131.11	48	.959	.943	$.054^{ns}$	.043	.065	.050
2. Hierarchical	146.93	50	.952	.936	$.057^{ns}$	.046	.067	.052
2 vs. 1 difference	15.82	2						
3. Unidimensional	617.23	54	.720	.657	.132	.122	.141	.086
3 vs. 2 difference	470.30	4						
Four-factor: Gender								
1. Configural invariance[a]	202.34	96	.945	.924	$.061^{ns}$	.049	.073	.056
2. Full metric invariance[b]	210.49	104	.945	.930	$.059^{ns}$	.047	.070	.059
2 vs. 1 difference	8.15^{ns}	8						
3. Full scalar invariance[c]	219.39	112	.945	.935	$.057^{ns}$	.046	.068	.061
3 vs. 2 difference	8.90^{ns}	8						
Four-factor: First language								
1. Configural invariance[a]	205.33	96	.945	.924	.062	.050	.073	.058
2. Full metric invariance[b]	216.35	104	.943	.928	$.060^{ns}$	.049	.071	.062
2 vs. 1 difference	11.02^{ns}	8						
3. Full scalar invariance[c]	248.28	112	.931	.919	.064	.053	.075	.065
3 vs. 2 difference	31.39	8						
4. Partial scalar invariance	227.57	110	.940	.928	$.060^{ns}$	.049	.071	.065
4 vs. 2 difference	11.22^{ns}	6						
Four-factor: Sample[d]								
1. Configural invariance[a]	214.55	96	.934	.909	.069	.057	.081	.060
2. Full metric invariance[b]	222.96	104	.934	.916	.066	.054	.078	.062
2 vs. 1 difference	8.41^{ns}	8						
3. Full scalar invariance[c]	261.82	112	.917	.902	.072	.060	.083	.066
3 vs. 2 difference	38.86	8						
4. Partial scalar invariance	233.99	109	.931	.916	.066	.055	.078	.062
4 vs. 2 difference	11.03^{ns}	5						

Note. Fit indexes and suggested acceptable fit thresholds: Comparative fit index (CFI) and Tucker–Lewis Index (TLI): > 90. Root mean square error of approximation (RMSEA) and standardized root mean square residual (SRMR): < .08. LL and UL = lower and upper limits for RMSEA. All χ^2 and RMSEA statistics were significant at $p < .05$ except *ns*.

[a]Equal form. [b]Equal factor loadings. [c]Equal intercepts. $N = 603$ except [d]$N = 520$.

models significantly worsened the fit. In broad terms of absolute goodness of fit, the four-factor and hierarchical models each showed good fit, whereas the unidimensional model showed poor fit. In terms of comparative fit, however, and replicating Webster et al.'s (2014) results, the four-factor model fit the data better than the hierarchical model (i.e., $\Delta\chi^2$), which fit better than the one-factor model.

Using multiple-group CFAs (Brown, 2006; Kline, 2011), we next tested for metric (item loadings) and scalar (item intercepts) invariance in the BAQ's four-factor model for three grouping variables: gender (men vs. women), first language (English vs. non-English), and sample (student vs. nonstudent; Table 7). We first tested a fully unconstrained (configural invariance or equal form) model that freed all parameters to differ by grouping variable, and then tested models that constrained the item loadings to be the same for both groups (metric invariance or equal factor loadings), followed by also constraining the item intercepts to be the same for both groups (scalar invariance or equal intercepts); we fixed the factor variances at 1.0 for the reference groups in the metric and scalar invariance models (see Brown, 2006, pp. 236–304). For gender, tests comparing these three models showed no significant differences (Table 7), suggesting both full metric and scalar invariance (equal factor loadings and intercepts).

For first language, we grouped the sample into respondents who answered "Yes" to the question, "Is English your first language?" (48%) and those who answered "No." Results indicated that the BAQ items showed full metric invariance for first language, but not full scalar invariance (equal factor loadings, but not equal intercepts; Table 7). To address this concern, we examined modification indexes iteratively, allowing the item with the highest intercept difference across groups to vary freely. After freeing the intercepts for two items, partial scalar invariance was achieved (equal intercepts for 10 items). Specifically, people whose first language was English scored higher on "Other people seem to get the breaks," but people whose first language was not English scored higher on "I have trouble controlling my temper."

To compare student with nonstudent samples, we created a new data set with the 213 U.S. and Canadian MTurk participants and the 307 students from Study 2 ($N = 520$). Similar to the prior analyses for first language, results indicated that the BAQ items showed full metric invariance, but not full scalar invariance (equal factor loadings, but not equal intercepts; Table 7). To this end, we again examined modification indexes iteratively, allowing the item with the highest intercept difference across groups to vary freely. After freeing the intercepts for three items, partial scalar invariance was achieved (equal intercepts for nine items). Specifically, nonstudents (vs. students) scored higher on the items "There are people who pushed me so far that we came to blows"; "When people annoy me, I may tell them what I think of them"; and "I am an even-tempered person" (prior to reverse-scoring).

Collectively, these CFAs suggest that the BAQ's four-factor structure holds not only for U.S. undergraduates (see Webster et al., 2014), but also for a more diverse sample of nonstudents. The four-factor model again produced the best fitting solution. Showing full metric invariance (equal factor loadings), factor loadings as a set did not vary significantly between men and women, between English- and non-English-speaking people, or between North American students and nonstudents. Although we showed full scalar invariance (equal intercepts) for gender, we only showed partial scalar invariance for first language and sample after freeing a few item intercepts. Thus, the BAQ showed evidence of measurement consistency across three broad categorical attributes: gender, English as a first language, and student versus nonstudent. Although more diverse than most U.S. student samples, MTurk participants still represent a select group of English-speaking people with Internet access. Thus, our findings are a preliminary—and necessary—step toward broadening the BAQ's generalizability.

Convergent validity. Having established some degree of measurement invariance in our nonstudent sample, we next created mean scores by averaging across items for each scale or subscale. Descriptive statistics and correlations among measures are shown in Table 8. One purpose of Study 3 was to test the convergent validity of the BAQ's Anger subscale with that of an established anger measure—the Trait Anger Scale. Indeed, this pair of anger measures had the highest correlation ($r = .58$), which was significantly higher than the correlations between the Trait Anger Scale and each of the other

TABLE 8.—Study 3: Descriptive statistics and zero-order correlations of observed scores for the Brief Aggression Questionnaire (BAQ), the Abbreviated Displaced Aggression Questionnaire (DAQ), and the Trait Anger Scale.

	Descriptive Statistics			Zero-Order Correlations								
Variable	*M*	*SD*	α	1	2	3	4	5	6	7	8	9
BAQ												
1. Anger	3.11	1.23	.67									
2. Physical Aggression	2.75	1.40	.82	.50								
3. Hostility	3.67	1.25	.65	.35	.35							
4. Verbal Aggression	4.15	1.24	.66	.25	.38	.19						
5. BAQ Mean	3.42	0.91	.81	.73	.80	.66	.63					
DAQ												
6. Angry Rumination	3.96	1.54	.88	.35	.29	.49	.18	.46				
7. Revenge Planning	3.21	1.54	.87	.56	.53	.47	.32	.67	.53			
8. Displaced Aggression	2.84	1.42	.86	.49	.32	.35	.17	.47	.32	.39		
9. DAQ Mean	3.34	1.17	.87	.60	.49	.56	.29	.68	.80	.83	.71	
Trait Anger Scale	2.05	0.47	.86	.58	.51	.45	.33	.66	.46	.55	.48	.64

Note. *N*s = 602–604. All *p*s < .001.

three BAQ subscales ($zs \geq 2.16$, $ps \leq .03$; see Lee & Preacher, 2013). After correcting for attenuation (unreliability) in both measures—by dividing the correlation by the square root of the product of each measure's reliability coefficient (α; Spearman, 1904)—this correlation increased ($r' = .76$). Thus, the BAQ's three-item Anger subscale explained 34% (or 58% after correcting for attenuation) of the variance in the 15-item Trait Anger Scale (or vice versa).

As an exploratory exercise, we also measured brief, three-item versions of the DAQ's three subscales. As expected, and showing some convergent validity, each subscale was positively and significantly correlated with each of the four BAQ subscales; however, there was some substantial variability in the strength of these correlations. Specifically, large correlations ($rs \approx .50$) linked (a) Anger with Revenge Planning and Displaced Aggression, (b) Physical Aggression with Revenge Planning, and (c) Hostility with Angry Rumination and Revenge Planning. Moderate correlations ($rs \approx .30$) linked (a) Anger with Angry Rumination, (b) Physical Aggression with Angry Rumination and Displaced Aggression, (c) Hostility with Displaced Aggression, and (d) Verbal Aggression with Revenge Planning. Small-to-moderate correlations ($rs \approx .20$) linked Verbal Aggression with Angry Rumination and Displaced Aggression. Of the four BAQ subscales, Anger was most closely related to the three DAQ subscales ($rs = .35$–$.56$), suggesting that this affective component of aggression might be more closely linked to displaced aggression than either the cognitive (Hostility) or behavioral (Physical Aggression and Verbal Aggression) forms of aggression. This analysis also suggests that, similar to the BAQ, the DAQ can be measured using three-item scales, although further testing will be necessary.

General Discussion

Self-report measures of aggression are necessary for assessing individual differences in aggressive traits. Short-form measures are increasingly in demand (Widaman et al., 2011), and aggression researchers need brief self-report scales to measure aggressive traits in contexts that place premiums on time or space (e.g., daily diary studies, field studies, special populations, mass testing or prescreening questionnaire packets). The 12-item BAQ fulfills this need by providing an efficient measure of anger, hostility, and verbal and physical aggression.

Three studies, including 1,279 participants, offered some converging evidence supporting the structure, validity, reliability, and generalizability of the BAQ's scores. Study 1 showed some consistency in the BAQ's structure as a four-factor aggression measure; however, one method—parallel analysis—supported a three-factor solution. Study 1 also showed that the BAQ had acceptable 12-week test–retest reliability, using both traditional and latent-variable methods. Using a behavioral aggression measure (noise blasts), Study 2 showed that the BAQ interacted with time (trial) in a retaliatory aggression experiment, such that BAQ mean scores positively predicted initial aggression (noise-blast intensity) more strongly at Trial 1 than at Trial 25. Study 3 confirmed the BAQ's four-factor structure and highlighted its partial metric and scalar invariance (equal factor loadings and intercepts) across gender, English as a first language, and student versus nonstudent groups using a large, diverse sample.

Studies 2 and 3 also expanded the BAQ's nomological network by showing some theoretically consistent patterns of convergent validity with behavioral aggression over time (trials), trait anger, and displaced aggression. Specifically, these findings improve on prior BAQ research (Webster et al., 2014), which reported evidence for convergent validity among the BAQ, its precursors, and behavioral aggression averaged across time (trials). In conjunction with prior work (Webster et al., 2014; see also Jonason & Webster, 2010; Webster, 2006, 2007; Webster & Bryan, 2007; Webster & Crysel, 2012; Webster, Kirkpatrick, Nezlek, Smith, & Paddock, 2007), these studies add to a growing literature of evidence supporting the BAQ as a psychometrically robust measure of individual differences in trait aggression.

Limitations and Directions for Future Research

Our results provide new and converging evidence supporting the structure, validity, reliability, and generalizability of BAQ scores that was absent from Webster et al.'s (2014) development of the BAQ. Despite the consistency of these findings, at least seven limitations remain that might serve to inspire future research.

First, regarding item selection, although the BAQ maximized within-factor loadings (Webster et al., 2014), it largely ignored cross-factor loadings, some of which were nontrivial (Table 4). These larger-than-desired cross-loadings pose a concern to both the convergent and discriminant validity of the BAQ's subscales. For instance, in latent-variable models (e.g., CFAs, SEMs) where the cross-loadings are set to zero, the BAQ's interfactor correlations could be unexpectedly high, leading to possible reductions in construct validity. Thus, researchers should be aware of this fact, especially when using the BAQ's subscales in latent-variable contexts.

Second, because we sought to create and validate three-item measures, construct underrepresentation is a concern. As noted earlier, creating brief measures requires a trade-off between efficiency (number of items) and breadth (representing a wide swath of the construct). Because we chose to emphasize brevity, and hence efficiency, the BAQ's subscales likely lack the breadth of their parent versions in the BPAQ. Nevertheless, because the BAQ continues to measure all four of the BPAQ's original factor-based subscales, the breadth of the mean BAQ score adequately represents that of its parent measure, the BPAQ ($rs = .96$; Webster et al. 2014).

Third, although Study 1 supported BAQ's four-factor structure across three methods—fit index thresholds, eigenvalues ≥ 1.0, and iterative outlier analyses—Study 1 also supported a three-factor solution using parallel analysis. Thus, although there was some consensus, some room for debate continues on the BAQ's factor structure, owing to its multifaceted nature as either a global aggression measure or one with four subscales. Study 3's results also supported the BAQ's four-factor structure, and although it fit the data better than a hierarchical factor model (via $\Delta\chi^2$), the fit of both models was acceptable (via other fit indexes). Again, the use of the BAQ as either a global or four-subscale measure appears justifiable; however, further research might be needed to resolve some of the BAQ's structural ambiguity.

Fourth, because Cronbach's alpha (α) is positively correlated with number of scale items (holding mean interitem

correlation [MIC] constant; see Cortina, 1993, p. 101), we expected the BAQ's subscales to have acceptable—but not excellent—internal consistency. This research produced 16 BAQ subscale αs, and all were acceptable (> .50; see Schmitt, 1996) for three-item measures, and these corresponded with respectable MICs (> .30; see Cortina, 1993). Thus, researchers should be aware of the internal consistency trade-off that accompanies the increased efficiency of the BAQ's three-item subscales. Indeed, researchers might wish to consider accounting for measurement error in the BAQ's three-item subscales by disattenuating correlations (as we did for the Trait Anger Scale) or by using latent-variable models (e.g., SEM).

Fifth, although we sought to support the BAQ's validity and reliability in our studies, we focused on test–retest reliability (Study 1), and on predictive and convergent validity (Studies 2 & 3). Although the BAQ's test–retest reliability is good (Study 1; Webster et al., 2014), its internal consistency reliability (α) is acceptable at best (as noted earlier). Although the BAQ showed evidence of predictive validity with noise-blast intensity (Study 2; and duration in Webster et al., 2014, Study 5), effect sizes were small, and generalizability to other behavioral aggression measures remains untested. In addition, although we expanded the BAQ's nomological network, thorough convergent (and discriminant) validity tests for its Hostility and Verbal Aggression subscales are lacking. Future research could also examine other forms of validity not tested here (e.g., diagnostic, ecological). For example, peer assessments of trait aggression using the BAQ might be especially informative (in conjunction with self-reports).

Sixth, although we assessed the BAQ in both student (Studies 1 & 2) and nonstudent (Study 3) samples, we did not sample any special populations (e.g., children, at-risk youth, prisoners, clinical samples) in which the BAQ might be used in the future. Thus, although the BAQ showed some convergence between North American student and nonstudent samples, we do not know the extent to which the BAQ's measurement properties also extend to various special populations. In addition, although Study 3 expanded the BAQ's generalizability to a more diverse, nonstudent sample, the extent to which it generalizes to other nations, cultures, and age groups remains unclear because we did not have the broad sample sizes necessary to adequately test these possibilities. Thus, further research on a global scale might be necessary for establishing the BAQ's international, cross-cultural, and generational generalizability.

Seventh, whether the BAQ—and particularly the Physical Aggression subscale—can assess proclivities toward extreme violent behavior remains an open question. Prior research found that scores on the BAQ's Physical Aggression subscale correlated positively with a validated laboratory aggression measure—blasting a stranger with aversive white noise (Webster et al., 2014). Although noise blasts show aggressive behavior, they are clearly not violent aggression. Testing whether the BAQ can predict extreme violence might be a goal of future research.

Implications and Recommendations

A broad theoretical and methodological implication of our findings is that psychological constructs can be measured effectively by using brief, efficient scales. When time constraints are placed on participants, or when space constraints are placed on researchers (e.g., number of items allowed), abbreviated measures can and should be used. Although the number of scale items contributes to internal consistency reliability (i.e., Cronbach's α), the main contributor is the average interitem correlation. If the average interitem correlation is reasonable, and if the items used are not redundant, then brief unidimensional scales are possible. Although longer, original scales are generally preferable when time and space are not concerns, abbreviated versions of those scales can be used to measure the constructs of interest. Specifically, we recommend that aggression researchers use the original 29-item BPAQ when time and space allows; however, given constrained time and space, we recommend they use the 12-item BAQ. Overall, our findings add to a growing literature that emphasizes the development of concise, efficient, self-report measures of psychological constructs, with the goal of meeting researchers' contemporary needs (e.g., Ames, Rose, & Anderson, 2005; Bryant & Smith, 2001; Donnellan, Oswald, Baird, & Lucas, 2009; Fraley, Waller, & Brennan, 2000; Gosling, Rentfrow, & Swann, 2003; Jonason & Webster, 2010; Nichols & Webster, 2013, 2014, 2015; Rammstedt & John, 2007; Robins, Hendin, & Trzesniewski, 2001; Webster & Crysel, 2012; Webster et al., 2014; Webster & Jonason, 2013; Widaman et al., 2011; Wood, Nye, & Saucier, 2010).

ACKNOWLEDGMENTS

Different analyses of some of the data included in Study 2 of this article have been published (Webster et al., 2014, Study 5). Aside from parts of the Study 2 Method section, and some zero-order correlations and descriptive statistics (*M*s, *SD*s, αs), these results do not reproduce previously published findings.

REFERENCES

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. 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

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

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press.

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

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

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*, 98–104. doi:10.1037/0021-9010.78.1.98

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., Bushman, B. J., Giancola, P. R., & Webster, G. D. (2010). The big, the bad, and the boozed-up: Weight moderates the effect of alcohol on aggression. *Journal of Experimental Social Psychology*, *46*, 619–623. doi:10.1016/j.jesp.2010.02.008

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

Fabrigar, L. R., Wegner, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*, 272–299. doi:10.1037/1082-989X.4.3.272

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

Funder, D. C. (2008). Persons, situations, and person–situation interactions. In O. P. John, R. Robins, & L. Pervin (Eds.), *Handbook of personality* (3rd ed., pp. 568–582). New York, NY: Guilford Press.

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

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making*, 26, 213–224. doi:10.1002/bdm.1753

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

Guttman, L. (1954). Some necessary and sufficient conditions for common factor analysis. *Psychometrika*, *19*, 149–161. doi:10.1007/BF02289162

Hauser, D. J., & Schwarz, N. (in press). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*. doi:10.3758/s13428-015-0578-z

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods*, *7*, 191–205. doi:10.1177/1094428104263675

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*, 61–83.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *32*, 179–185. doi:10.1007/BF02289447

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

Judd, C. M., McClelland, G. H., & Ryan, C. S. (2009). *Data analysis: A model comparison approach* (2nd ed.). New York, NY: Routledge.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *20*, 141–151. doi:10.1177/001316446002000116

Khoo, S.-T., West, S. G., Wu, W., & Kwok, O.-M. (2006). Longitudinal methods. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 301–317). Washington, DC: American Psychological Association.

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

Kline, R. B. (2013). Exploratory and confirmatory factor analysis. In Y. Petscher & C. Schatsschneider (Eds.), *Applied quantitative analysis in the social sciences* (pp. 171–207). New York, NY: Routledge.

Krueger, J. I. (2009). A componential model of situation effects, person effects, and situation-by-person interaction effects on social behavior. *Journal of Research on Personality*, *43*, 127–136. doi:10.1016/j.jrp.2008.12.042

Lee, I. A., & Preacher, K. J. (2013, September). *Calculation for the test of the difference between two dependent correlations with one variable in common* [Computer software]. Retrieved from http://quantpsy.org

MacCallum, R. C., Widman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*, 84–99.

Muthén, L. K., & Muthén, B. O. (2010). *Mplus user's guide* (6th ed.). Los Angeles, CA: Author.

Nezlek, J. B. (2008). An introduction to multilevel modeling for social and personality psychology. *Social and Personality Psychology Compass*, *2*, 842–860. doi:10.1111/j.1751-9004.2007.00059.x

Nezlek, J. B. (2011). Multilevel modeling for social and personality psychology. In J. B. Nezlek (Ed.), *Sage library in social and personality psychology methods*. London, England: Sage.

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

Nichols, A. L., & Webster, G. D. (2014). The single-item Need for Consistency Scale. *Individual Differences Research*, *12*, 50–58.

Nichols, A. L., & Webster, G. D. (2015). Designing a brief measure of social anxiety: Psychometric support for a three-item version of the Interaction Anxiousness Scale (IAS-3). *Personality and Individual Differences*, *79*, 110–115. doi:10.1016/j.paid.2015.01.043

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Psychological Science*, *7*, 528–530. doi:10.1177/1745691612465253

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

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

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., & Congdon, R. T., Jr. (2006). *HLM 6: Hierarchical linear and nonlinear modeling*. Lincolnwood, IL: Scientific Software International.

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. doi:10.1037/1040-3590.8.4.350

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, *15*, 72–101. doi:10.2307/1422689

Spielberger, C. D., Jacobs, G., Russell, S., & Crane, R. S. (1983). Assessment of anger: The State–Trait Anger Scale. In J. N. Butcher & C. D. Spielberger (Eds.), *Advances in personality assessment* (Vol. 2, pp. 161–201). Hillsdale, NJ: Erlbaum.

Swann, W. B., & Seyle, C. (2005). Personality psychology's comeback and its emerging symbiosis with social psychology. *Personality and Social Psychology Bulletin*, *31*, 155–165. doi:10.1177/0146167204271591

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

Watson, D. (2004). Stability versus change, dependability versus error: Issues in the assessment of personality over time. *Journal of Research in Personality*, *38*, 319–350. doi:10.1016/j.jrp/2004.03.001

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. (2009). The person–situation interaction is increasingly outpacing the person–situation debate in the scientific literature: A 30-year analysis of publication trends, 1978–2007. *Journal of Research in Personality*, *43*, 278–279. doi:10.1016/j.jrp.2008.12.030

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., DeWall, C. N., Pond, R. S., Jr., Deckman, T., Jonason, P. K., Le, B. M., Bator, R. J. (2014). The Brief Aggression Questionnaire: Psychometric and behavioral evidence for an efficient measure of trait aggression. *Aggressive Behavior*, *40*, 120–139. doi:10.1002/ab.21507

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.

Wood, D., Nye, C. D., & Saucier, G. (2010). Identification and measurement of a more comprehensive set of person-descriptive markers from the English lexicon. *Journal of Research in Personality*, *44*, 258–272. doi:10.1016/j.jrp.2010.02.003