Using a Polygraph System for Evaluation of the Social Desirability Response Bias in Self-Report Measures of Aggression

Dmitri Poltavski  
*University of North Dakota*, dmitri.poltavski@UND.edu

Richard Van Eck  
*University of North Dakota*, richard.vaneck@und.edu

Austin T. Winger  
*University of North Dakota*, austin.winger@und.edu

Charles Honts

Follow this and additional works at: [https://commons.und.edu/tl-fac](https://commons.und.edu/tl-fac)

Part of the [Educational Leadership Commons](https://commons.und.edu/tl-fac)

Recommended Citation

[https://commons.und.edu/tl-fac/8](https://commons.und.edu/tl-fac/8)

This Article is brought to you for free and open access by the Department of Teaching and Learning at UND Scholarly Commons. It has been accepted for inclusion in Teaching & Learning Faculty Publications by an authorized administrator of UND Scholarly Commons. For more information, please contact [zeinebyousif@library.und.edu](mailto:zeinebyousif@library.und.edu).
Using a Polygraph System for Evaluation of the Social Desirability Response Bias in Self-Report Measures of Aggression

Dmitri Poltavski, Richard Van Eck, Austin T. Winger, Charles Honts

Abstract

Empirical evidence suggests that respondents systematically overreport socially desirable behaviors and systematically underreport socially undesirable behaviors. This “social desirability response bias (SDRB)” presents significant challenges for research that relies on self-report measures to assess behaviors that adhere to or violate social norms. The present study used a state-of-the-art polygraph system to examine SDRB in widely used aggression questionnaires, including the Buss–Perry Aggression Questionnaire-Short Form, the Indirect Aggression subscale of the Aggression Questionnaire by Western Psychological Services, and the Reactive–Proactive Questionnaire. Sixteen college students with no criminal record, no known prior history of aggressive behavior, and no reported drug abuse, responded to verbally administered forms of the instruments. Indirect aggression items produced the largest, statistically significant physiological response across the sampled channels. The magnitude of this response was negatively and significantly correlated with the self-report ratings of the frequency of such behaviors. The mean separation between physiological (skin conductance) and self-report responses for indirect aggression remained significant and consistent with correlational analyses when both types of responses were converted to the same scale and compared directly. Finally, the relative magnitude of skin conductance response for items assessing indirect aggression was significantly greater than the relative magnitude of skin conductance response for direct aggression. Overall, the study suggests that the use of automated state-of-the-art polygraph systems may potentially identify sensitive items on self-report instruments where social responsibility response bias is possible. Implications for the use of this procedure with such instruments are discussed.

Keywords

Polygraph, Aggression questionnaires, Social desirability response bias

With the explosion of online survey research in the past two decades, an increasing number of social scientists focus on investigation of social taboos, illegal or reprehensible behavior, and risky choices. Obtaining valid and reliable data based on such items has proven to be difficult because of differences in sensitivity levels between different questions (Krumpal 2013). Tourangeau and Yan (2007) suggest that there are three primary dimensions that can independently contribute to a question’s sensitivity: intrusiveness into one’s private life, threat of disclosure of the reported information to third parties, and social desirability. The latter refers to the tendency of respondents to admit to socially desirable traits and behaviors, while denying socially undesirable behaviors. Unlike ‘intrusiveness,’ the problem associated with ‘social desirability’ is not the sensitivity of a question, but rather the sensitivity of an answer with a greater endorsement of an attitude or behavior (a ‘yes’ response) corresponding to a greater likelihood of social disapproval (Fowler 1995). There is ample empirical evidence that respondents systematically overreport socially desirable behaviors and attitudes, while they systematically underreport socially undesirable ones (e.g. Barnett 1998; Lee 1993; Tourangeau et al. 2000). This phenomenon is sometimes referred to as social desirability response bias, or SDRB. For example, underreporting is quite common for socially undesirable behaviors such as illicit drug use, smoking, alcohol consumption, and abortion.
Respondents also tend to underreport crime victimization, aggressive behavior, certain types of income (e.g., welfare), and unpopular attitudes, such as racism and anti-Semitism (Krumpal 2013).

The Bogus Pipeline, or BPL, is a procedure that leads participants to believe they are being monitored by a lie detector. When a BPL is used while questioning participants about social desirability issues, it reliably reduces socially desirable responses and increases honesty, particularly for issues of fact, such as the level of sexual activity (Alexander and Fisher 2003; Fisher 2013), child abuse (Ganon et al. 2007) or sexual assault (Strang and Peterson 2017).

It is important to note that physiological data are not collected during these BPL studies. Rather, the participant is led to believe in the infallibility of the BPL procedure to determine the truthfulness of a response. Often, this procedure serves as an effective countermeasure against SDRB and, in the case of criminal offenders, helps to overcome the ‘threat-of-disclosure’ aspect of socially sensitive responses. In most research situations, however, the BPL methodology is not feasible. Therefore, understanding the propensity for dishonesty on self-report instruments may make an important contribution to research investigating sensitive topics, breach of social norms, socially undesirable traits, and other behaviors.

While the BPL represents an elaborate simulation of polygraph testing, the actual polygraph procedure is still often used by law enforcement and government agencies, as well as private industries in the context of detection of deception. Modern polygraph systems are highly deployable, sophisticated, and automated. Most of the criticism of the polygraph procedure raised in the psychophysiological community has been associated with deployment of specific protocols for detection of deception (e.g., Comparative Question Test or the Concealed Information Test; see Synnott et al. 2015, for a review). The accuracy of algorithms based on estimation of probability of deception is largely dependent on the specific polygraph examination protocol used, selection bias, assessment of ground truth, differences among examiners, examiner–examinee interactions, and delays in the timing of questions (Slavkovic 2002).

The advantage of modern automated polygraph systems for research purposes comes from the fact that they also provide a relative magnitude of a physiological response for each channel, based on automated standardization of raw scores and range-adjustment of subsequent z-scores. These statistics can be obtained separately for any type of question and are not based on the response difference between any two given questions (e.g., relevant vs. control). Utilization of these measures during polygraph testing is, thus, protocol-independent and helps avoid major issues that have beset polygraph research in the past (Synnott et al. 2015).

In the current study we tested whether a state-of-the-art automated polygraph system could be a useful tool in evaluating the risk of response bias on questionnaires investigating socially reprehensible and potentially criminal behaviors (i.e., different types of aggressive behavior). Such measures commonly use an Inconsistent Responding Index to safeguard against inaccurate self-reports of a particularly sensitive behavior. However, if inaccurate responses are consistently provided across the items in a given instrument (e.g., assessing behaviors that are highly socially desirable or undesirable, like aggression), such measures are prone to failure. Specifically, we hypothesized that elevated relative response magnitudes (RRMs) to questions assessing an individual’s history of different types of aggressive behaviors may be informative of SDRB, if observed in the context of self-reported absence or low occurrence of such behaviors by the individual.
Methods

Participants

Using an online research participation system, 16 participants (11 men and 5 women) were recruited from the undergraduate student population at a midwestern research university. The mean age of the participants was 20.31 years old ($SD = 1.62$). All of the participants were enrolled in psychology classes during the spring semester of 2016, and they received extra credit points in their courses for their participation in the study. All of the participants reported being in good health and did not report any history of mental illness, criminal behavior, or drug use/abuse. At the time of the study, none of the participants reported use of prescription medications. This sample was drawn from residents of a county, which averaged 4 juvenile detentions per 10,000 for violent crime between 1997 and 2015, whereas the national average for this period is approximately 13 (National Center for Juvenile Justice 2017). This suggests the population is less aggressive/violent than the national average. Further, violent crime statistics for the university campus and immediately adjacent community for the years 2014–2016 show an average of 0.95 violent crimes per 10,000, suggesting that the student population is significantly less aggressive/violent than the county and national average. Figure 1 presents the number of each type of violent crime per capita by year and for all 3 years.

Fig. 1 Violent crime rates for the University of North Dakota campus and surrounding community per 10,000 for years 2014–2016

Instruments

Buss–Perry Aggression Questionnaire (BPAQ)

For the purposes of the present study we used a 12-item Buss–Perry Aggression Questionnaire-Short Form (BPAQ-SF) based on the original 29-item BPAQ (Buss and Perry 1992). The 12 items assess the same four subtypes of aggression as the original BPAQ, which include physical aggression, verbal aggression, anger, and hostility. In the BPAQ-SF, participants are asked to rate the degree to which a statement applies to them using a five-point, Likert-type scale, ranging from “not at all like me” (1) to “completely like me” (5). Webster et al. (2014) reported very similar high convergent and discriminant validity and test–retest reliability for BPAQ-SF and the original BPAQ.
The original BPAQ is often used in clinical practice and represents a shortened version of the Buss–Durkee Hostility Inventory (Buss and Durkee 1957), a longtime standard for assessing anger and aggression. Besides the 4 dimensions of aggression captured by the BPAQ and BPAQ-SF the Buss–Durkee Hostility Inventory also includes six items measuring indirect aggression. For the purposes of the present study, we took all six items from the indirect aggression subscale of the Buss–Durkee Hostility Inventory and added them to the end of the BPAQ-SF, in order to gather participants’ responses to statements of indirectly aggressive behaviors. This addendum aimed to bolster the validity of the BPAQ-SF, as well as its ability to measure indirect forms of aggression.

The Reactive–Proactive Questionnaire (RPQ)

This 23-item instrument (created by Raine et al. 2006) was originally designed to measure levels of reactive and proactive aggression in children and adolescents for the purpose of discerning when a participant may be provoked into aggressive behavior (reactive), as compared to actively generating aggressive situations (proactive). In recent years, the RPQ has been validated in several studies with adults (e.g., Brugman et al. 2017; Cima et al. 2013). The RPQ includes 11 reactive items (e.g., “Yelled at others when they have annoyed you”; “Gotten angry when frustrated”) and 12 proactive items (e.g., “Had fights with others to show who was on top”; “Vandalized something for fun”). All RPQ items have a three-point response format (0 = never, 1 = sometimes, 2 = often), yet the RPQ items do not reference a specific time period for the behavior. The questions simply ask how often respondents have engaged in particular behaviors.

Polygraph System

The Computerized Polygraph System (CPSpro Fusion; Scientific Assessment Technologies 2011) was used in the present study. The CPSpro Fusion includes an electronic polygraph with up to 13 physiological recording channels. For our study, the following four channels were used:

- Two respiration channels (thoracic and abdominal) using UFI Pneumotrace transducers.
- One skin conductance channel using disposable silver–silver chloride electrodes.
- One relative blood pressure (BP) channel using an inflatable cuff.

The CPSpro Fusion hardware was connected via a USB 2.0 HID port to a Dell Inspiron 15 laptop computer (Intel Core i3 processor and 4.0 GB of RAM) running a Windows 8 operating system. The hardware comprises 24-bit analog-to-digital converters, samples at up to 384 Hz, and stores each physiological signal 60 times per second.

The CPSpro Fusion software was used to collect continuous physiological data and to perform subsequent offline analyses. Initial calibration period was 10 s long with individual channel sensitivities automatically adjusted by the software. While running a chart, the recorded data were displayed in real time on the screen. Time marks and artifact notations were recorded automatically on the charts. The skin conductance level, pressure in the relative BP cuff, and subject heart rate were continuously displayed in digital form on the screen. All polygraph data, questions, annotations, and time marks were stored on the hard disk for later retrieval and analyses.

Data analysis in CPSpro Fusion allows automated scoring using a relative response magnitudes method (RRM), which computes values and displays bar graphs that show the relative strengths of physiological reactions to individual questions. The RRM values represent a transformation of the raw data values to a common metric that ranges from 0 to 100 locating all of the test stimuli from the largest, assigned a value.
of 100 to the smallest that is assigned a value of 0. CPSPro Fusion algorithms have been validated in a number of recent peer-reviewed publications (e.g. Honts and Reavy 2015; Honts and Alloway 2007; Honts et al. 2004).

The following explains the characteristics of physiological reactions quantified by CPSPro Fusion in this study. The times and levels of high and low points in the response curves provided the information needed to quantify all of the physiological variables (Kircher and Raskin 2011) except for respiration length, which required a separate algorithm (Timm 1982). Skin conductance and relative BP algorithms were based on response amplitude. As described by Raskin and Kircher (2014), an electrodermal response curve was obtained from 60-Hz samples of data collected over a 20-s period following question onset. Within that interval, low points in the response curve were identified as changes from negative or zero slope to positive slope, while high points referred to changes from positive slope to zero or negative slope. The software automatically computed differences between each low point and every succeeding high point. The greatest identified difference was defined as peak amplitude.

According to Raskin and Kircher (2014) peak amplitude of the rise in diastolic points of the relative BP is referred to as a 60-Hz diastolic response curve. It is defined by linear interpolation between adjacent diastolic points from question onset for a period of 20 s. Within that interval, the same procedures used to measure peak amplitude of the electrodermal response were used to measure the peak amplitude of the rise in diastolic points of the relative BP measure.

Respiration line length (RLL) was used to determine RRM for respiration channels and was based on the studies by Timm (1982) and Kircher and Raskin (1988). Raskin and Kircher (2014) describe it as a simple line length representing “the sum of absolute differences between adjacent 60-Hz samples of respiration starting at question onset for a period of ten seconds” (p. 97). Line length was measured separately for thoracic and abdominal respiration channels, and the two values were automatically averaged by the CPSPro Fusion software.

Automated scoring algorithms were applied within the above extraction windows for different channels with the duration between any two question onsets kept at 21 s (default setting). While the inter-stimulus interval can be lengthened, its shortening would not allow the use of scoring algorithms. The RRM analysis was used to obtain weighted sums of physiological response magnitudes across the four channels for each question (Kircher and Raskin 2011). The analysis of RRMs was automatically performed by the CPSPro Fusion software through several steps. First, the raw scores for each physiological channel were transformed into z-scores. For each question presented, the mean z-score was computed separately for thoracic and abdominal respiration channels. The z-scores for each component were then range-adjusted to obtain measures of response magnitude. When the z-scores were range-adjusted, the smallest (most negative) observed z-score receives a value of 0.0, and the greatest (most positive) z-score receives a value of 1.0. This process was accomplished with the following formula:

\[
X' = X - \frac{X - X_{\text{MIN}}}{X_{\text{MAX}} - X_{\text{MIN}}} \quad \text{(Formula 1)}
\]

where,

- \(X'\) is the relative magnitude, and it ranges from 0.0 to 1.0. \(X'\) is 0.0 if it is the smallest observed z-score (\(X_{\text{MIN}}\)), and it is 1.0 if it is the largest observed z-score (\(X_{\text{MAX}}\)).
- \(X\) is the observed z-score for a particular question, chart, and channel.
- \(X_{\text{MIN}}\) is the smallest obtained z-score for a channel of physiological data.
- $X_{\text{MAX}}$ is the largest obtained $z$-score for a channel of physiological data.

The weighted sum was automatically calculated as a global measure of each participant's physiological reaction to a question. To obtain the weighted sum for a question, the relative magnitude for each of the three physiological measures was multiplied by a weighting coefficient, and the three cross-products were summed. The weighting coefficients are those described in the discriminant analysis from Kircher and Raskin (1988).

**Polygraph Templates**

The questions from the BPAQ-SF and indirect aggression subscale, along with the RPQ were divided into four polygraph templates:

1. Q1–Q12 of the BPAQ-SF were loaded as Template 1, with each question type coded as relevant (R1–R12).
2. The six questions assessing indirect aggression were loaded into Template 2 and marked as relevant (R13–R18).
3. The first 12 questions of the *Reactive–Proactive Questionnaire* were included in Template 3 and marked as relevant (R19–R30).
4. The remaining RPQ questions (Q13–Q23) were loaded into Template 4 and coded as relevant (R31–R41).

The four separate templates were necessary to allow for brief breaks to deflate the BP cuff, in order to minimize participants’ physical discomfort associated with continuous partial venous occlusion. Each template started with two irrelevant questions, in order to establish physiological baselines for the recorded channels.

**Procedure**

Upon arrival at the lab, each participant read and signed the consent form and was informed that the study would involve a series of questions about interactions with others, while physiological data was measured. Next, the participant was asked to stand up to fit thoracic and abdominal pneumotrace transducers secured around the chest (over the rib cage) and abdomen (around the navel) with Velcro straps. Each pneumotrace was stretched by approximately 1 inch (2.5 cm) after the participant exhaled. Then, the participant was asked to sit in a polygraph chair with fully adjustable arm supports. Two disposable electrodermal activity electrodes were attached to the palmar surfaces of the distal phalanges of the index and middle fingers on the participant’s right hand. Finally, an inflatable BP cuff was placed around the upper left arm of the participant, with the cuff arrow over the brachial artery. The participant was instructed to sit comfortably in the chair with both arms placed on the armrests, while the researchers started the calibration procedure in the CPSpro Fusion software. After a series of recommended inflations and deflations of the BP cuff, as well as manual adjustment of air distribution in the cuff to minimize drift during data collection, the cuff was inflated to the minimum pressure of 50 mmHg, which was maintained throughout the experimental procedure. The pressure level was chosen to minimize discomfort and numbing sensation in the arm associated with partial venous occlusion.

Following calibration, a template was loaded and the polygraph chart was run. Templates were loaded in one of the two possible order combinations (Order 1: T1, T2, T3, T4 or Order 2: T3, T4, T1, T2) in a counterbalanced fashion. Template 4 was always loaded after Template 3 and Template 2 was always loaded after Template 1 to preserve the order of the questions in paper-and-pencil versions of the self-
report instruments. Within each polygraph session, the participant received both orders of questions (order 1 followed by order 2 or order 2 followed by order 1) so that by the end of the experimental session, each question had been asked twice, generating eight polygraph charts. Physiological response magnitudes and self-reported scores were derived as averages between the two presentations of each question.

The interstimulus interval between the questions was set at 21 s (the default setting), which is the minimum time between questions necessary to apply the scoring algorithms. After running a chart for each template (a total of four charts per subject), the participant was given a 3–4 min break, during which his/her BP cuff was deflated and he/she was free to flex the arm and move the fingers to return normal blood circulation. Chart duration varied between 4 min, 30 s (min) and 7 min, 10 s (max). The duration of the polygraph procedure (from the start of chart 1 until the end of chart 4) was approximately 32 min, with a range of 30–35 min.

To minimize response vocalizations and consequent electromyographic artifacts, response scales for each template were loaded on a computer screen in front of the participant. Thus, the participant was able to respond with a number (1–5 for the BPAQ-SF and indirect aggression subscale, and 0–2 for the RPQ) that best represented the likelihood/frequency of occurrence of each behavior. Verbal responses for each question and each instance of question occurrence were manually recorded by another researcher not involved with running polygraph charts. After administration of all templated questions, the polygraph sensors were disconnected, the researchers ensured that the participant did not experience any physical or psychological adverse effects from the procedure, and the participant was dismissed.

The researchers did not edit any of the collected data, as there were no noted obvious artifacts or upward/downward drifts in data during tracings that warranted detrending. According to the CPSpro Fusion manual, while detrending may improve the visual appearance of a tracing on the screen and printed charts, it has no effect on response magnitude analyses (Kircher and Raskin 2011). Furthermore, when CPS algorithms were independently evaluated by Slavkovic (2002) on 149 real-life specific incident cases provided by the Department of Defense Polygraph Institute, only raw data were used for statistical modeling.

**Statistical Analyses**

In combination, the three self-report instruments measured seven different types of aggressive behavior: physical aggression, verbal aggression, anger, hostility, indirect aggression, reactive aggression, and proactive aggression. For each type of aggression, mean scores were calculated across constituent questions and two administrations of each question. Mean scores were calculated for both self-report measures and relative magnitudes of physiological response. The first set of physiological data analyses included all four sampled channels (abdominal and thoracic respiration, skin conductance, and BP). These analyses were based on the RRM automatically calculated by the CPSpro Fusion software across all four channels. First, the raw scores for each physiological channel were transformed into z-scores. For each question presented, the mean z-score was computed separately for each channel. The z-scores for each component were then range-adjusted to obtain measures of response magnitude using Formula 1. The weighted sum was then automatically calculated by the software as a global measure of each participant’s physiological reaction to a question. This weighted sum was the global measure of the relative physiological response magnitude used as a dependent variable in a series of repeated measures analyses, where the seven types of aggression were levels of the within-subject variable (aggression type). Each relative physiological response magnitude was also correlated with corresponding self-report mean scores using Pearson r correlation coefficients.
The researchers also attempted to directly compare RRMss between physiological channels and self-report ratings. We obtained z-scores for each question’s self-report rating; then we calculated RRMss for each question using Formula 1. To compare mean relative self-report magnitudes with corresponding mean RRMss for physiological channels, we needed to ensure 100% equivalency in calculation of RRMss.

The automatically generated RRMss for physiological channels were based on z-scores obtained from variability in raw values for a given channel within the same individual during the presentation of a particular question (within-subject variability). On the other hand, RRMss for the self-report ratings were obtained based on z-scores calculated from variability in self-report values in the entire participant sample (between-subject variability). Thus, the researchers first recalculated RRMss for each channel by using mean raw channel values for each individual and each question administration. These means were used to obtain z-scores that were now based on the variability of mean raw channel values in the sample rather than the individual. Next, corresponding RRMss for each physiological channel were calculated for each question and question administration using Formula 1. Then the obtained channel RRMss values were averaged across the seven types of aggression.

Furthermore, although CPSpro Fusion automatically calculated the weighted sum of RRMss for physiological channels, this calculation does not represent a simple average of relative magnitudes for the three channels. Rather, the calculated weighted sum provided by CPSpro Fusion is based on the sum of relative magnitudes multiplied by corresponding weights (Kircher and Raskin 2011). The coefficients used to compute these weighted sums are proportional to the weights used to compute discriminant scores for individual questions in the probability analyses. These weights are not reported in the manual, yet the greatest weight in the algorithm was assigned to the skin conductance response. Similarly, in our sample, Repeated Measures Analyses of Variance for each channel showed significant differences among aggression types only for RRMss for skin conductance ($F_{6,90} = 2.84, p = 0.04$) and not for BP or respiratory channels. Thus, only the calculated RRMss for skin conductance were used in the statistical analysis that directly compared physiological responses with self-report ratings [a 7 (aggression type) x 2 (response type) Repeated Measures Analysis of Variance].

**Power Analyses**

Using [G-Power 3.1](Faul et al. 2009), it was estimated that the sample size for this 7 x 2 within-subject design, required to observe at least a medium effect size of within factors ($F = 0.25$) with the minimum acceptable statistical power of 0.80 at alpha = 0.05, was only 12 participants. With 16 participants in our sample, the study was powered to detect small-to-medium effect sizes ($f = 0.20$) of within-subject factors.

**Results**

The results of a one-way Repeated Measures Analysis of Variance showed a significant main effect of aggression type on the relative magnitude of physiological response ($F_{6,90} = 3.51, p < 0.01$; partial $\eta^2 = 0.19$). Pairwise comparisons, using the Bonferroni adjustment, showed that the global physiological response to questions measuring indirect aggression was significantly ($p < 0.01$) greater than the physiological response to questions assessing angry behavior ($mean \, difference = 0.08; Hedges’ \, g = 1.14$). None of the other comparisons were significant; these results are summarized in Table 1 and Fig. 2.
Table 1 Descriptive statistics for self-report ratings and magnitude of physiological responses for different aggressiveness subscales

<table>
<thead>
<tr>
<th>Aggressiveness type</th>
<th>Self-report mean rating (SD)</th>
<th>Mean magnitude of physiological response (SD)</th>
<th>Pearson r</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physical</td>
<td>1.7 (0.69)</td>
<td>0.36 (0.07)</td>
<td>0.07</td>
<td>0.80</td>
</tr>
<tr>
<td>Verbal</td>
<td>2.29 (0.78)</td>
<td>0.35 (0.09)</td>
<td>0.33</td>
<td>0.21</td>
</tr>
<tr>
<td>Anger</td>
<td>1.74 (0.68)</td>
<td>0.32 (0.07)</td>
<td>−0.49</td>
<td>0.05</td>
</tr>
<tr>
<td>Hostility</td>
<td>2.20 (0.85)</td>
<td>0.35 (0.08)</td>
<td>−0.19</td>
<td>0.48</td>
</tr>
<tr>
<td>Indirect</td>
<td>2.21 (0.41)</td>
<td>0.41 (0.08)</td>
<td>−0.58</td>
<td>0.02</td>
</tr>
<tr>
<td>Reactive</td>
<td>0.70 (0.19)</td>
<td>0.35 (0.07)</td>
<td>−0.35</td>
<td>0.19</td>
</tr>
<tr>
<td>Proactive</td>
<td>0.13 (0.18)</td>
<td>0.35 (0.08)</td>
<td>−0.05</td>
<td>0.85</td>
</tr>
</tbody>
</table>

Fig. 2 Mean magnitude of physiological responses for various aggressiveness types on the Buss–Perry Aggressiveness Questionnaire and the Reactive–Proactive Questionnaire
There was also a significant negative correlation between the global physiological response on the indirect aggression scale and the corresponding mean self-report rating (Pearson $r = -0.58$, $p < 0.02$). Additionally, there was a significant negative correlation between the global physiological response on the anger subscale of the BPAQ-SF and the corresponding mean self-report rating (Pearson $r = -0.49$, $p = 0.05$). Correlation coefficients for the other types of aggression were not significant (see Table 1).

The results of a two-way 7 (aggression type) × 2 (response type) Repeated Measures Analysis of Variance showed a significant main effect of aggression type ($F_{6,90} = 3.74, p < 0.01$). Pairwise comparisons using the Bonferroni adjustment showed that the overall RRM (across response types) was significantly higher for indirect aggression ($M = 0.453$, $SE = 0.036$) than for proactive aggression ($M = 0.234$, $SE = 0.039$, $p < 0.01$, $Hedges' g = 1.15$). None of the other pairwise comparisons were significant. The main effect of response type was also significant ($F_{1,15} = 131.50; p < 0.01; partial \eta^2 = 0.89$) generally showing a significantly greater RRM for skin conductance ($M = 0.40$, $SE = 0.04$) than self-report ratings across the seven aggression categories ($M = 0.32$, $SE = 0.03$).

Additionally, there was a significant aggression type x response type interaction ($F_{6,90} = 3.174, p < 0.01; partial \eta^2 = 0.18$). This interaction was broken down using the Dunn–Bonferroni post-hoc test of planned contrasts. As can be seen from Fig. 3, tests of the RRMs for skin conductance were significantly greater than the relative self-report magnitudes on the following four subscales: physical aggression ($mean \text{ difference} = 0.13$, $Hedges' g = 0.55$), anger ($mean \text{ difference} = 0.20$, $Hedges' g = 0.91$), indirect aggression ($mean \text{ difference} = 0.11$, $Hedges' g = 0.65$) and proactive aggression ($mean \text{ difference} = -0.20; Hedges' g = 1.0$). Using a one-tailed test of significance, a family wise error equal to 0.05, 60 degrees of freedom for the error term, and four contrasts, our $t_{\text{crit}}$ was 2.29. All of the four planned comparisons were significant at alpha = 0.05: $t_{\text{physical}} = 2.83$; $t_{\text{anger}} = 3.32$; $t_{\text{indirect}} = 2.4$; $t_{\text{proactive}} = 3.33$.

**Fig. 3** Relative response magnitudes for different aggressiveness types as a function of response type (self-report vs. skin conductance)
Discussion

The relative magnitude of physiological response on the indirect aggression subscale was the highest when compared to the other measured forms of aggression, and was negatively correlated with the self-report ratings of indirect aggression provided by the participants. Direct comparisons of RRMs for self-report ratings on the aggressiveness scales with corresponding RRMs for skin conductance confirmed the original correlational findings. Specifically, RRMs for skin conductance were significantly greater than the RRMs for self-report ratings on the anger subscale of the BPAQ-SF, as well as for indirect aggression. Additionally, RRMs for skin conductance were significantly greater than corresponding self-report ratings for the physical aggression, anger and indirect aggression subscales of the BPAQ-SF, as well as the proactive aggression subscale of the RPQ.

The discrepancy between physiological response and self-report ratings on these four subscales would seem to support the SDRB. Direct comparisons with the self-report ratings were based on the EDA channel. This process was warranted, because when response magnitudes were compared among different aggression types for each individual channel, only skin conductance (not BP, nor thoracic or abdominal respiration) showed significant differences across aggression types.

Electrodermal activity has long been regarded the most sensitive and reliable of the three channels of the polygraph (Kircher and Raskin 2002; Orne et al. 1972). Unlike the cardiovascular system that receives noradrenergic and cholinergic inputs from both branches of the autonomic nervous system, the electrodermal activity is exclusively under sympathetic control. Electrodermal activity does not depend on fluctuations of sympathetic hormones such as norepinephrine and epinephrine as the activity in the sweat-producing eccrine glands is controlled by the levels of postganglionic acetylcholine (Stern et al. 2001). Respiratory activity may be even more problematic in this regard as it is affected by both the autonomic and the central nervous system, with the latter allowing an individual to easily bring respiratory activity under voluntary control (Synnott et al. 2015).

While this does make the EDA channel less prone to being affected by extraneous factors, it in no way eliminates the inherent problems associated with any attempt to infer psychological states based on peripheral nervous system activity. One such confounding variable in our study that may have increased EDA activity irrespective of the underlying psychological processes is a general arousal value associated with a particular question. For example, physiological responses to BPAQ-SF items that are about overt aggression/violence (e.g., “Given enough provocation, I may hit another person” or “How often have you used physical force to get others to do what you want?”), may have produced greater arousal than items pertaining to indirect aggression (e.g., “When people are bossy, I take my time doing what they want, just to show them”; “When someone really irritates me, I might give him or her the silent treatment”).

In fact, this issue plagued early polygraph protocols based on the relevant–irrelevant (R–I) question structure. The biggest criticism of such protocols was that because the intent of the questions was transparent, both non-deceptive and deceptive subjects were more likely to be aroused by relevant than by irrelevant questions because they would recognize the importance of such questions and would feel threatened by them (e.g. Lykken 1981; Horowitz et al. 1997). Consistent with this criticism, the R–I protocol produced an unacceptably high rate of false positives (Horowitz et al. 1997). Consequently, in the present study significant differences between self-report ratings and magnitudes of physiological response (EDA) for the physical aggression and anger subscales of the BPAQ-SF and proactive aggression of the RPQ may have been driven by greater perceived arousal values of the items on these scales irrespective of whether the participant had committed such acts and intended to be untruthful (i.e.,
not by socially desirable responses to these items). Since in the present study we did not collect arousal ratings for various questions from study participants, this explanation cannot be ruled out.

Conversely, the SDRB explanation seems to be the strongest for questions pertaining to indirect aggression, which might be expected to be less threatening than items on the physical aggression and anger subscales of the BPAQ-SF, and thus result in less arousal. Specifically, there were four factors supporting this explanation:

1. Indirect aggression items produced the largest physiological response across all of the sampled channels, which cannot be readily explained by their greater arousal value compared to the items on the other aggression subscales.

2. The magnitude of this global physiological response was negatively correlated with the self-report ratings of the frequency of such behaviors (i.e., the mean rating corresponded to the category “unlike me”).

3. The separation between physiological (skin conductance) and self-report responses remained significant and consistent with correlational analyses, when both types of responses were converted to the same scale and directly compared.

4. Finally, the relative magnitude of skin conductance response for items assessing indirect aggression was significantly greater than the relative magnitude of skin conductance response for direct aggression (i.e., proactive aggression scale).

Limitations

These findings perhaps relate to the participant population as much or more than the various aggression scale items. While it is possible that questions pertaining to indirect aggression may be most susceptible to SDRB, it is just as plausible that this may only be relevant with a population of young college students with no record of violent behavior and criminal history. Expression of proactive, physical aggression, anger, and hostility among college students may not be as common as day-to-day instances of indirect aggression, such as spreading rumors about people they do not like or giving the “silent treatment” to someone who irritates them. Our data on violent crimes among college students of the University, from which the sample was drawn supports this suggestion. Several other studies have also shown high prevalence of relational aggression among college students (Basow et al. 2007; Dahlen et al. 2013; Loudin et al. 2003). Relational aggression is a form of indirect aggression that refers to a set of behaviors through which the aggressor intends to harm others by adversely affecting their relationships, sense of belonging, social status, or reputation (Dahlen et al. 2013). Relational aggression includes behaviors such as gossip, spreading malicious rumors, social exclusion, among others. Thus, caution should be taken in making any generalizations beyond the current study population when analyzing our relatively novel use of the polygraph procedure to investigate the accuracy of responses to aggression questions.

Conclusions

The results of our study suggest that the use of automated state-of-the art polygraph systems may be an efficient way to identify sensitive items on self-report instruments that may produce biased responses. Sampling of autonomic responses (in conjunction with administration of questionnaires purported to measure specific behaviors, attitudes, or feelings) may help to further validate such instruments, as well as to develop reliable countermeasures to minimize responder bias within a specific population of interest. For example, those who have a record of committing illegal (criminal populations) or legal (law
enforcement or military populations) violent acts might differ from other populations in terms of their physiological responses to direct aggression items and/or the discrepancy between their physiological responses and self-report on direct aggression. Further, individuals within given populations may differ in their responses depending on individual circumstances (e.g., eligibility for parole in the prison population; length of duty or number of enemy engagements in the military population). In the absence of such knowledge about a population of interest, the interpretation and generalization of results of self-report measures should be done with caution. Finally, our findings suggest that there may be a need to re-examine some of the prior research that has detected or failed to detect differences using similar self-report instruments to measure behaviors.

References


