Original Article

# Regulatory adaptations for delivering information: the case of confession

Daniel Sznycer [a,*], Eric Schniter [b], John Tooby [a], Leda Cosmides [a]

[a] Center for Evolutionary Psychology, University of California, Santa Barbara
[b] Economic Science Institute, Chapman University

A B S T R A C T

Prior to, or concurrent with, the encoding of concepts into speech, the individual faces decisions about whether, what, when, how, and with whom to communicate. Compared to the existing wealth of linguistic knowledge however, we know little of the mechanisms that govern the delivery and accrual of information. Here we focus on a fundamental issue of communication: The decision *whether* to deliver information. Specifically, we study spontaneous confession to a victim. Given the costs of social devaluation, offenders are hypothesized to refrain from confessing unless the expected benefits of confession (e.g. enabling the victim to remedially modify their course of action) outweigh its marginal costs—the victim's reaction, discounted by the likelihood that information about the offense has not leaked. The logic of welfare tradeoffs indicates that the victim's reaction will be less severe and, therefore, less costly to the offender, with decreases in the cost of the offense to the victim and, counter-intuitively, with *increases* in the benefit of the offense to the offender. Data from naturalistic offenses and experimental studies supported these predictions. Offenders are more willing to confess when the benefit of the offense to them is high, the cost to the victim is low, and the probability of information leakage is high. This suggests a conflict of interests between senders and receivers: Often, offenders are more willing to confess when confessions are *less* beneficial to the victims. An evolutionary–computational framework is a fruitful approach to understand the factors that regulate communication.

© 2015 Elsevier Inc. All rights reserved.

## 1. Introduction: The puzzle of confession

A popular moral view of confession is that confessing is the right thing to do following wrongdoing and that it positively reflects on the moral character of the offender. Some researchers echo this view arguing that spontaneous confessions are internally motivated, guilt-driven attempts to undo the harm done and to restore the relationship (Weiner, Graham, Peter, & Zmuidinas, 1991; see Tangney, Miller, Flicker, & Barlow, 1996). From an evolutionary perspective, however, this analysis does not provide a theory of why natural selection would have favored this behavior. Consequently, the phenomenon of spontaneous confession remains puzzling.

Undoing the harm done to someone else – or claiming to be the cause of it – is not in any obvious way fitness-promoting in the general case. Indeed, the decision to commit the self-interested, other-harming act in the first place already involved the evaluation that it was of net advantage to the perpetrator. More importantly, disclosing to a victim that you were the one who harmed her could very well damage or destroy a pre-existing cooperative relationship. This is presumably the reason why the perpetrator kept the information from the victim to begin with. Absent compensatory benefits, victims of harm will at the very least revise their valuation of the perpetrator downward.

Moreover, the victim may retaliate, making further offenses costlier to the offender. Furthermore, information about the offender's uncooperative inclinations may diffuse into the community, yielding reputational costs. So, regardless of whether the offender up-regulates his valuation of the victim after inflicting harm, why would a brain shaped by natural selection be designed to spontaneously communicate such self-damaging information? Here we subject confession to an evolutionary–functional analysis to understand its logic and to highlight the advantages of analyzing issues in human communication in evolutionary–functional terms.

### 1.1. Adaptationist approaches to communication, social valuation, and social negotiation

Confession – disclosing to someone that you have harmed them – is a subtype of communication. Systems that signal a given kind of information should only evolve if they yield net benefits to the sender, on average and under ancestral-like conditions (Maynard-Smith & Harper, 2003; Tooby & Cosmides, 1990; Williams, 1966). The adaptive problem for the signaler – sending information when it is likely to be self-beneficial – is framed by the design of information-processing machinery in the receiver. In particular, what effect will the signal have on the receiver's behavior, and what fitness consequences will this behavioral response have on the sender? It is this evolved architecture in signal receivers that shapes the adaptations for communication in senders (Krebs & Dawkins, 1984)—including whether to send a given kind of

signal at all. Since disclosure of an offense triggers victim behavior that adversely impacts the sender, the default strategy is to not send the clarifying signal. This links the puzzle of confession to the evolved emotion program of shame.

By hypothesis, shame is a neurocomputational program whose evolved function is to orchestrate best-bet responses in an individual to the adaptive problem of information potentially or actually spreading to others which would have a negative effect on how those others value and treat that individual (Sznycer, 2010). Shame acts to limit the likelihood and the costs of being socially devalued. Our evolved psychology is packed with devaluation-limiting features (De Hooge, Zeelenberg, & Breugelmans, 2011; Fessler, 2001; Gilbert, 1998; Sznycer et al., 2012). For instance, actions thought to be interpreted as reprehensible by an audience trigger shame even when people know the action caused no harm to the audience or third parties (Sznycer, Tooby, & Cosmides, 2009)—what one would expect from a system designed to tamper with adverse representations about the self in others' minds. On this view, shame prompts non-disclosure. Thus, for a confession to be made, there must be other motivational forces overriding this default.

## 1.2. Welfare tradeoff functions and their regulation

For ancestral humans, actions taken by one individual would often have had positive or negative effects on the welfare of others. Selection pressures such as kin selection (Hamilton, 1964) and the asymmetric war of attrition (Hammerstein & Parker, 1982) specify strategies that conditionally trade off the welfare of the actor against the welfare of another organism (the target). If these selection pressures acted on our ancestors, then circuits are required to compute the effects of acts on self and others, and deploy a weighting function indexing the extent to which the self will trade off their own welfare to enhance the welfare of another—a welfare tradeoff ratio between the self ($i$) and individual $j$: WTR$ij$ (Tooby, Cosmides, Sell, Lieberman, & Sznycer, 2008).

Recent studies support the hypothesis that the human mind is equipped with a welfare tradeoff architecture. Consistent with the WTR hypothesis, people allocate resources between self and others as if a target-specific threshold divided those tradeoffs that are worth making from those that are not (Delton et al., 2008; see Jones & Rachlin, 2006; Sell, Tooby, & Cosmides, 2009).

An individual is benefited when others recalibrate their welfare trade-off ratios toward her upwards, and is harmed when others recalibrate their WTRs toward her downwards. Reciprocally, it costs more to place a higher weight on someone else's welfare. So there are a variety of social strategies that evolved to minimize losses by not putting too great a weight on others, and to make gains by inducing others to place a higher weight on the actor's welfare. Welfare trade-off ratios as internal magnitudes regulating choices constituted adaptive problems for socially interacting humans, and gave rise to a suite of what have been called recalibrational emotions (Tooby et al., 2008). Anger is triggered when another person places too low a weight on the welfare of the actor (compared to what the actor implicitly computes she is entitled to). Once triggered, the function of anger is to bargain for a higher WTR through conditional threats of increasing harm or decreasing benefits to the "offender," or at worst to curtail the costs of making sacrifices for another that are not redeemed by reciprocal valuation by that person.

In contrast, guilt is a recalibrational emotion that is triggered when one discovers that one has placed too little weight on another's welfare from one's own point of view. It is a recalibrational process that operates even in the absence of knowledge by the victim of the guilt-producing act (Smith, Webster, & Eyre, 2002; Sznycer, 2010). Guilt is triggered by acts that harm a valued target to an unexpectedly large extent. The functional product of guilt (in contrast to shame) is to increase one's WTR toward the harmed individual, to bring it to the level of the equilibrium long-run valuation the system places on the welfare of the other, and in the shorter run to mitigate or remedy the negative effects

of the harmful act, when the earlier valuation turned out not to be in the interests of the offender.

In sum, an act entailing too low a WTR toward another will elicit anger in the victim and shame, guilt, or both in the offender. Selfish acts often involve both reputational damage and unwarranted cost imposition from the offender's perspective, and so shame and guilt tend to co-activate. However, the distinct functional signatures of each emotion can be discerned. For instance, publicity exacerbates shame more than guilt (Scarnier, Schmader, & Lickel, 2009; Smith et al., 2002; Sznycer et al., 2009). And cooperative motivations have a robust link to guilt, but only a contingent link to shame (De Hooge, Breugelmans, & Zeelenberg, 2008; Tangney, Stuewig, & Martinez, 2014; Wicker, Payne, & Morgan, 1983).

## 1.3. Welfare tradeoffs, the infliction of harm, and confession

When a person inflicts harm on another to accrue a benefit for himself, this reflects a WTR the offender behaviorally expressed toward the victim. Regarding selfish transactions – of relevance to confession situations – the WTR indexes the minimum benefit the self requires before imposing a given cost on another. The higher the WTR, the higher the valuation of the other, and the less frequent the selfishness; and vice versa. An individual will be motivated to benefit at the expense of another, but only when:

$$\text{B}off_O \geq \text{C}off_V \text{WTR}_{OV} \tag{1}$$

Where WTR$_{OV}$ is the WTR of the offender ($O$) toward the victim ($V$), B$off_O$ is the offender's estimate of the benefit he or she will derive from the offense, and C$off_V$ is the offender's estimate of the cost of the offense to the victim.

Following the principle of receiver-targeted communication, we advance a model where offenders target their confessions to the social valuation psychology of victims with the aim of minimizing the devaluing and punitive reactions of victims.

This *devaluation-minimizing model of confession* applies to situations where an offender benefits at the expense of a victim and the victim is unaware of the offense or ignorant about the offender's identity—at least in the estimation of the offender. The devaluation-minimizing model is limited to voluntary, spontaneous, truthful confession to the victim. Confession by force or to third parties is beyond its scope (for other types of confession, see Kassin & Gudjonsson, 2004; Schelling, 1960). The model is also limited to cooperative or potentially cooperative relationships (see below). Under the devaluation-minimizing model, spontaneous confession is a decision with one binary choice (confess or not) and three possible outcomes, each of which is associated with a different payoff.

Outcome 1. *Offender confesses.* Confessing entails costs to the offender. By hypothesis, these lie not in the production of the signal (where the cost is trivial) but in the impact on the offender of the victim's reaction to the information (Lachmann, Számádó, & Bergstrom, 2001). This reaction is driven by anger (McCullough, 2008; Sell, 2005), and may include retaliation, demands for WTR up-regulation and restitution, victim WTR down-regulation, and/or withdrawal from the cooperative relationship. Other emotions such as sadness and disgust (Lim, 2012; Schniter & Shields, 2013) may also be mobilized in the victim.

The victim's reaction is exacerbated as the difference between the WTR expected by the victim and the WTR implied in the offense increases (Sell, 2005). Inequality (1) governs offender behavior but it can also allow the victim to infer the offender's WTR via estimates of the costs and benefits involved.

$$\text{WTR}_{OV} \leq \frac{\text{B}off_O}{\text{C}off_V} \tag{2}$$

The lower $B_{OFF_O}$ and the higher $C_{OFF_V}$, the lower the upper bound of the WTR implied in the offense. When the WTR implied is low, the victim's devaluing and punitive reaction will be intense, and the cost to the offender of confessing will be high. We designate this latter variable $C_{CONF_O}$.

Confessions may also yield benefits. A confession may indicate honesty (openness and truthfulness), which may signal to the victim that the offender feels guilt and values the victim's welfare (Van Lange & Kuhlman, 1994). A confession may also allow a victim to remedially modify their course of action (Utikal, 2012). To the extent that the offender values the victim, this remediation yields indirect benefits to the offender himself (see Tooby & Cosmides, 1996). Finally, confessions obviate the memory burden, preference falsification (Kuran, 1998), and behavioral circuitousness involved in living a lie. We term the total benefit of confession to the offender, $B_{CONF_O}$. In some cases, as when less circuitousness is achieved, confessions entail direct benefits to the confessors themselves. In other cases, as when a victim is enabled to remedy their course of action, a confession represents a direct benefit to the victim and only an indirect benefit to the offender. While a confession can have the effect of benefiting a victim and even be proximately motivated by concerns over the victim's welfare, we reiterate that a mechanism for emitting confessions is not evolutionarily viable unless it is designed to benefit *the senders themselves* (here the offenders/confessors), however indirectly. In sum, confession nets offenders $B_{CONF_O} - C_{CONF_O}$.

*Outcome 2. Offender does not confess, victim finds out.* A victim who has not received a confession may still learn about the offense. For instance, the offense may be witnessed by the victim or gossipy third parties, or it may leave evidence pointing to the offender, or the offender may out himself involuntarily. We designate the total probability of the victim finding out via information leakage, $p$. If the victim finds out, the offender is assumed both to incur $C_{CONF_O}$ and to get some of the benefits of confession (e.g. dispensing with the costs of living a lie) but not others (e.g. victim's valuation of honesty)—since here cues of offender guilt are absent. On the benefit side, offenders get $kB_{CONF_O}$, where $k$ denotes the extent to which benefits are *not* eroded when the victim finds out. Note that $kB_{CONF_O}$ is a benefit stemming from the victim finding out via sources other than a confession. $k$ will often be lower than 1; compared to a confessed-to victim, a victim who finds out will tend to yield worse payoffs to the offender. In all, absence of confession followed by discovery nets the offender $p(kB_{CONF_O} - C_{CONF_O})$.

*Outcome 3. Offender does not confess, victim remains ignorant.* A victim who has not received a confession remains ignorant with probability $1 - p$. In this case, the offender is assumed to obtain neither $C_{CONF_O}$ nor $B_{CONF_O}$, and thus to remain at baseline welfare (0)—typically including the benefit of the other-harming act that motivated the offender to commit the act in the first place (Fig. 1).

By hypothesis, the mind a) represents the variables listed above, b) estimates and updates the values of those variables, and c) issues confessions based on a decision rule featuring those variables. The rule motivates confession only when confessing nets higher payoffs than not confessing:

$$B\text{conf}_O - C\text{conf}_O > p(k\ B\text{conf}_O - C\text{conf}_O) + (1-p)0 \qquad (3)$$

We note that this rule need not be deployed deliberately or consciously. As per inequality (2) and the logic of anger, the cost of the victim's reaction to the offender will be proportional to the inverse of the $WTR_{OV}$ implied in the offense (i.e. $z\frac{C\text{off}_V}{B\text{off}_O}$), where $z$ is a proportionality factor. Thus, we can rewrite (3) as:

$$B\text{conf}_O(1-pk) + z\frac{C\text{off}_V}{B\text{off}_O}(p-1) > 0 \qquad (4)$$

From (4), the following predictions follow: Willingness to confess increases as…

P1. the cost of the offense to the victim decreases (except when $p = 1$).
P2. the benefit derived from the offense by the offender increases (except when $p = 1$).
P3. the benefit of the confession to the offender increases (except when $p = k = 1$).
P4. the likelihood of information leakage increases.
P5. the benefits of confession erode when the victim finds out; e.g. honesty (except when $p = 0$).

Here we test predictions 1 through 4. To assess the devaluation-minimizing model with naturalistic offenses we conducted the following study.

## 2. Study 1

### 2.1. Method

#### 2.1.1. Participants

Participants were 89 students (49 females) from the University of California, Santa Barbara (age: $M = 20$, $SD = 2$). Participants obtained course credit for their participation.

#### 2.1.2. Procedure

This study included two parts. In the first part, participants narrated an autobiographical episode where they imposed costs on another individual (the target) and either confessed to them or not (between-subjects design). In both conditions, the criteria for episode selection
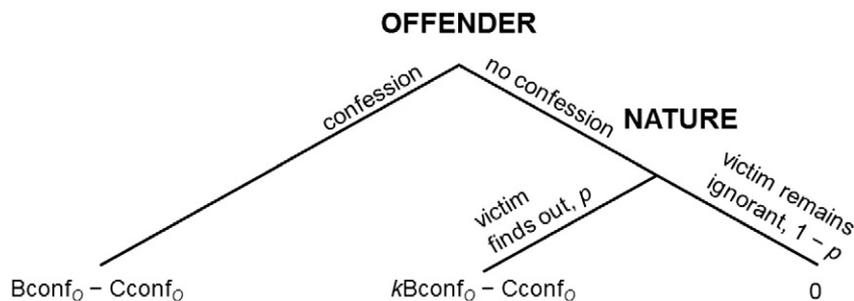


**Fig. 1.** Payoffs of the decision whether to confess. If the offender confesses, they incur the cost of confession (the victim's devaluative and punitive reaction as well as reputational damage), and receive benefits which may be direct (e.g. reduced behavioral circuitousness) or indirect (e.g. remediation of the victim's course of action). If the offender does not confess, the victim either finds out, with probability $p$, or remains ignorant, with probability $1 - p$. In the former case, the offender incurs the cost of confession and receives some of the benefits of confession (e.g. reduced behavioral circuitousness) but not others (e.g. victim's valuation of honesty). If the victim remains ignorant, the offender remains at baseline welfare—which includes the benefit of the other-harming act.

were as follows: First, the target incurred material, emotional, or reputational costs as a result of the participant's action or inaction. Second, the cost was imposed on a single individual, the target. Third, at least initially, the target did not know the identity of the offender or even that they incurred a cost. Fourth, the situation was remembered well. In addition, in *confession*, the participant eventually chose to go to the target and confess, whereas in *no-confession*, the participant did *not* confess. After narrating the episode, participants were asked why they chose to confess or not confess. To check whether the episodes met the selection criteria, at the end of the study participants were asked whether each criterion was "true" or "not true" of the episode.

In the second part, participants answered a structured questionnaire about the episodes. The questionnaire included four scales reflecting the variables that modulate confession under the devaluation-minimizing model, with items measured on 7-point Likert scales. We were unable to find existing measures with the requisite specificity, so we created the following scales.

(a) *Cost of the offense to the victim* (e.g. "How much damage did your actions cause to [target]?"; 11 items; Cronbach's α: 0.93).

(b) *Benefit of the offense to the offender* (e.g. "Leaving aside the cost incurred by [target], how much did you gain from what you did?"; 11 items; Cronbach's α: 0.83).

(c) *Likelihood of information leakage* (some items were phrased in condition-specific manner: e.g. "If you had not confessed to [target], how likely would [target] have been to eventually learn about your actions from gossip?" (*confession*), "How likely would [target] be to eventually learn about your actions from gossip?" (no-*confession*); 9 items; Cronbach's α: 0.91)

(d) *Benefit of the confession to the offender* This scale included items about both honesty [H] and remediability [R] (some items were phrased in condition-specific manner: e.g. "If [target] had found out what you had done from sources other than your confession, how annoyed at you would [target] have been about the fact that you didn't tell them?" [H] (*confession*), "My confession would allow [target] to change [target]'s plans for the better" [R] (no-*confession*); 11 items; Cronbach's α: 0.81). We note that in this and the following studies the variable *benefit of the confession* is, as per the devaluation-minimizing model, meant to index benefits *to the offender*, however indirect. However, this variable was operationalized as direct benefits of the confession to the victim (remediability) and as victim's valuation of the offender (honesty) (see Supplementary Information, available on the journal's website at www.ehbonline.org).

In the final probe, six participants, three in each condition, stated that their episodes failed to meet one or more selection criteria. The analyses below exclude these participants, although their inclusion does not significantly alter the results. The episodes reveal the participants understood the criteria for episode selection, including the somewhat abstract notion of "cost."

Given the recency effect (Murdock & Bennet, 1962) and its potential for generating unequal amounts of bias between conditions, participants were asked to report time elapsed since the offense and how well they recall it. Neither varied between conditions (*P*s: .88 and .57, respectively).

### 2.2. Results

A principal axis factor analysis of the items revealed a factor structure similar to the anticipated one. Although the analysis yielded ten factors with eigenvalues above 1, only the first four were interpretable and accounted for at least 9% of the variance. These factors mapped onto the four variables discussed above. The analyses below involve the full set of items. Importantly however, reanalysis with the scales derived via factor analysis does not alter the pattern of results.

*Was the cost of the offense to the victim lower in* confession *than in* no-confession? No. Victim cost was not significantly different between *confession* ($M = 4.29, SD = 1.60$) and *no-confession* ($M = 4.32, SD = 1.16$) ($t_{69} = -0.08, P = .94$).

*Was the benefit of the offense to the offender higher in* confession *than in* no-confession? No. Offender benefit was not significantly different between *confession* ($M = 2.98, SD = 1.13$) and *no-confession* ($M = 3.26, SD = 1.29$) ($t_{81} = -1.02, P = .31$).

For each participant, a benefit-cost ratio was computed by dividing offender benefit by victim cost. Because it normalizes benefit per unit cost – a better proxy of $WTR_{OV}$ – this index may more successfully separate confessors from non-confessors. The difference between *confession* ($M = 0.82, SD = 0.52$) and *no-confession* ($M = 0.84, SD = 0.48$) was not significant, however ($t_{81} = -0.21, P = .84$).

*Was the likelihood of information leakage higher in* confession *than in* no-confession? Yes. Likelihood of information leakage was higher in *confession* ($M = 3.34, SD = 1.81$) than in *no-confession* ($M = 2.22, SD = 1.07$) ($t_{60} = 3.36, P = .001, r = .40$).

*Was the benefit of the confession to the offender higher in* confession *than in* no-confession? Yes. The benefit of the confession was significantly higher in *confession* ($M = 4.80, SD = 0.99$) than in *no-confession* ($M = 4.01, SD = 0.94$) ($t_{81} = 3.72, P = .0004, r = .38$). This effect was driven by both honesty (*confession*: $M = 5.09, SD = 1.08$; *no-confession*: $M = 4.45, SD = 0.92; t_{81} = 2.91, P = .005, r = .31$) and remediability (*confession*: $M = 4.03, SD = 1.58$; *no-confession*: $M = 2.83, SD = 1.62; t_{81} = 3.38, P = .001, r = .35$).

Space considerations preclude qualitative analysis of the episodes, but we provide a few for illustrative purposes. One participant drank his housemate's soda without permission. He then confessed because "the crime was petty and I could have easily compensated for the cost he incurred." Another participant knowingly under-contributed to a group assignment at school. He did not confess since "how would they know what my full potential was anyways? Simply put, they wouldn't." Finally, one participant stole $20 from a friend and later found out that his friend couldn't afford to pay for gas and got late to work. He confessed because "they suffered more than just the loss of money but also suffered punishment at work and most likely lost status with their boss/co-workers."

### 2.3. Discussion

As predicted, relative to unconfessed offenses, confessed offenses featured higher likelihood of information leakage and higher benefits of confession—both valuation of honesty and remediability. Some of the items measuring these variables were phrased in condition-dependent way, which may render the comparison problematic. Although we tried to conceptually equalize each query, the syntactic differences were unavoidable. However, when comparing the five items that were identical across conditions (2 leakage items and 3 benefit of confession items), there were significant or marginal advantages in favor of the confession condition for all the items (*P*s: .007–.14), suggesting that the observed effects were not due to the difference in phrasing.

On the other hand, neither victim cost nor benefit of the offense to the offender varied between the confession/no confession conditions.

This study lacked experimental control of the relevant variables so the results should be interpreted cautiously. Further, the differences between conditions may reflect confessors' perceptions formed post-confession rather than or in addition to pre-confession assessments. Although we asked confessors to provide the impressions they had before the confession, the more recent post-confession impressions may have still influenced the ratings.

Regarding the null results, perhaps before confessing, offenders did perceive their offenses as involving low victim costs and high offender benefits, but after confessing, victims' arguments to the contrary may

have changed the confessors' perceptions of these variables, nullifying the original differences (see Sell, 2005). Further, even if unconfessed offenses on average do feature high cost/benefit ratios, non-confessors may perceive that disclosure in an experimental setting involves an undesirable degree of information leakage; this may bias the sampling toward the milder offenses and nullify an actual difference.

To address these methodological issues we conducted the following experimental study.

## 3. Study 2

### 3.1. Method

#### 3.1.1. Participants

Participants were 257 students (180 females) from the University of California, Santa Barbara (age: $M = 20$, $SD = 2$). Participants obtained course credit for their participation.

#### 3.1.2. Procedure

The experiment consisted of hypothetical two-part scenarios where the participant herself was described as benefitting at the expense of another same-sex individual (e.g. a classmate). In the first part of the scenario, a situation was presented where the participant takes an action that results in a benefit for them and a cost to the other. After the first part, the participant was asked "how willing would you be to go to [the other person] and confess to [them] that you [took an action which resulted in a cost to them]." Responses were registered on a 7-point Likert scale (1 = not willing at all, 7 = very willing). In the second part of the scenario, further information was given that modulates the willingness to confess under the devaluation-minimizing model. This was followed by a second measure: "Compared to your earlier answer, how willing would you be to go to [the other person] and confess to [them] that you [took an action which resulted in a cost to them]." Responses were registered on a 7-point Likert scale ($-3$ = much less willing than before, $0$ = as willing as before, $+3$ = much more willing than before).

There were three different types of scenarios, each of which included (in the second part) the following confession-modulating factors: (a) cost of the offense to the victim, (b) benefit of the offense to the offender, (c) likelihood of information leakage, and (d) benefit of the confession to the offender. Each confession-modulating factor included one of two levels: high and low. The factors were *not* factorially crossed, however. Thus, there were 24 scenario variants in total (see Supplementary Information, available on the journal's website at www.ehbonline.org).

For example, in the first part of the "Ticket" scenario, the participant scribbles on something that looks like scrap paper but turns out to be his classmate's winning lottery ticket, thus making the ticket unredeemable. In the second part, the story is resumed in one of the following ways. *Cost of the offense to the victim*: The redemption value of the ticket is [$10] (*low*)/[$1000] (*high*). *Benefit of the offense to the offender*: The participant used the ticket to [doodle and scrawl random phrases to kill time] (*low*)/[write down important job-related information] (*high*). *Likelihood of information leakage*: While the participant was in the kitchen marking the ticket his classmate [remained in the living room the whole time] (*low*)/[walked by the kitchen on his way to the bathroom] (*high*). *Benefit of the confession to the offender* was modeled as *honesty* in some scenarios and as *remediability* in others. *Honesty*: The participant was told that the thing his classmate dislikes most is when people [have no sense of humor] (*low*)/[are dishonest] (*high*) (Ticket scenario). *Remediability*: In the first part of the "Sandwich" scenario the participant distractedly grabs a sandwich from the fridge and eats it, later realizing that what he ate was his workmate's lunch. In the second part, the workmate [hardly has any time left before finishing his only break of the day, so even if he knew that his sandwich was gone he could not go out and get something else to eat] (*low*)/[has some time left before finishing his only break of the day, so if he knew

that his sandwich was gone he could use his remaining break time to go out and get something else to eat] (*high*).

Each participant received three scenario variants presented sequentially by computer. Each scenario type was presented once without repetition of confession-modulating factor. Order of scenario type and of level of confession-modulating factor (high vs. low) was randomized. Confession-modulating factor triplet was counterbalanced. There were 32 responses per (second part) condition on average.

### 3.2. Results

*Was willingness to confess higher when victim cost was low rather than high?* Yes. An independent samples t-test indicated that, across the three scenarios, reported willingness to confess (second part) was higher in the low cost-to-victim condition ($M = +0.94$, $SD = 1.38$) than in the high cost-to-victim condition ($M = -0.20$, $SD = 1.45$) ($t_{191} = 5.58$, $P = 8 \times 10^{-8}$, $r = .37$). This was true in two scenarios ($Ps = 2 \times 10^{-8}$–.013) but not the third one ($P = .57$). See Table 1.

*Was willingness to confess higher when the benefit of the offense to the offender was high rather than low?* Yes. Willingness to confess was higher in the high benefit-to-offender condition ($M = +0.88$, $SD = 1.33$) than in the low benefit-to-offender condition ($M = -0.52$, $SD = 1.24$) ($t_{188} = -7.45$, $P = 3 \times 10^{-12}$, $r = .48$). This was true in two scenarios ($Ps\ 1 \times 10^{-13}$–.007) but not the third one ($P = .48$).

*Was willingness to confess higher when the likelihood of information leakage was high rather than low?* Yes. Willingness to confess was higher in the high leakage condition ($M = +0.77$, $SD = 1.03$) than in the low leakage condition ($M = -0.07$, $SD = 0.82$) ($t_{191} = -6.38$, $P = 1 \times 10^{-9}$, $r = .42$). This was true in all three scenarios ($Ps$: $3 \times 10^{-5}$–.007).

*Was willingness to confess higher when the benefit of the confession to the offender was high rather than low?* Yes. Willingness to confess was higher in the high benefit-of-confession condition ($M = +1.11$, $SD = 1.18$) than in the low benefit-of-confession condition ($M = +0.27$, $SD = 0.80$) ($t_{167} = -5.84$, $P = 3 \times 10^{-8}$, $r = .41$). All three scenarios supported the prediction ($Ps$: $7 \times 10^{-6}$–.016). Both remediability and honesty yielded the predicted effects.

We note that differences in willingness to confess at stage 1 were absent and therefore unable to account for the differences at stage 2.

### 3.3. Discussion

The results of study 2 (available on the journal's website at www.ehbonline.org) supported the four predictions. Willingness to confess increased with the benefit of the offense to the offender, the likelihood of information leakage, and the benefit of the confession to the offender, and decreased with the cost of the offense to the victim.

Benefit of the confession to the offender was operationalized in two ways: As the victim valuing honesty (vs. another valuable personality trait, e.g. sense of humor), and as the confession enabling the victim to remedially alter their course of action (vs. the confession not enabling such remediation). Both kinds of benefits enhanced the odds of confession.

The behavioral effects obtained with student samples sometimes directionally differ from those obtained in samples from the general population (Peterson, 2001). To determine whether these findings replicate on a sample more demographically and geographically representative of the United States population, we conducted study 3 (available on the journal's website at www.ehbonline.org).

## 4. Study 3

### 4.1. Method

#### 4.1.1. Participants

Participants were 913 individuals (353 females) with an age range of 18 to 71 ($M = 28$, $SD = 10$). They were recruited with Amazon Mechanical Turk and paid $0.10 to complete the task.

#### 4.1.2. Procedure

The design of this experiment was the same as that of study 2, except for the following. Participants were run via web rather than in the lab. Participants completed only one 2-part scenario. Responses were recorded on 5-point Likert scales (1 = not willing at all, 5 = definitely willing (part 1); −2 = much less willing than before, 0 = as willing as before, +2 = much more willing than before (part 2)). There were 38 participants per (second part) condition on average.

#### 4.2. Results

*Was willingness to confess higher when victim cost was low rather than high?* Yes. Across scenarios, reported willingness to confess (second part) was higher in the low cost-to-victim condition ($M = +0.62$, $SD = 1.15$) than in the high cost-to-victim condition ($M = −0.01$, $SD = 1.09$) ($t_{230} = 4.30$, $P = 3 \times 10^{-5}$, $r = .27$). This was the case in two scenarios ($Ps$: $2 \times 10^{-5}$–.001) but not the third one ($P = .86$). See Table 1.

*Was willingness to confess higher when the benefit of the offense to the offender was high rather than low?* Yes. Willingness to confess was higher in the high benefit-to-offender condition ($M = +0.56$, $SD = 0.99$) than in the low benefit-to-offender condition ($M = −0.61$, $SD = 1.05$) ($t_{222} = −8.59$, $P = 2 \times 10^{-15}$, $r = .50$). Two scenarios supported the prediction ($Ps$: $2 \times 10^{-16}$–$7 \times 10^{-5}$) but the third one did not ($P = .21$).

*Was willingness to confess higher when the likelihood of information leakage was high rather than low?* Yes. Willingness to confess was higher in the high leakage condition ($M = +0.56$, $SD = 0.80$) than in the low leakage condition ($M = −0.12$, $SD = 0.67$) ($t_{220} = −6.99$, $P = 3 \times 10^{-11}$, $r = .43$). All three scenarios supported the prediction ($Ps$: $2 \times 10^{-7}$–.005).

*Was willingness to confess higher when the benefit of the confession to the offender was high rather than low?* Yes. Willingness to confess was higher in the high benefit-of-confession condition ($M = +0.83$, $SD = 0.87$) than in the low benefit-of-confession condition ($M = +0.32$, $SD = 0.80$) ($t_{225} = −4.62$, $P = 6 \times 10^{-6}$, $r = .29$). This was true in two scenarios ($Ps$: $9 \times 10^{-6}$–.0008) but not the third one ($P = .70$).

We note that differences in willingness to confess at stage 1 either were absent or were present but unable to account for the differences at stage 2.

Across studies 2 and 3, every scenario-factor combination met the predictions of the devaluation-minimizing model at least once, the exceptions being the offense cost and offense benefit manipulations of the Sandwich scenario where the victim and the offender were described as being differentially hungry. The fact that a sandwich is relatively inexpensive suggests the differential hunger may have been insufficiently diagnostic of differential WTR; this may account for the null effects.

#### 4.3. Discussion

Willingness to confess decreased with victim cost, and increased with the benefit of the offense to the offender, the likelihood of the victim finding out anyway due to information leakage, and the benefit of the confession to the offender. Both honesty and remediability enhanced the odds of confession. This replicates the results of study 2 in a sample more representative of the American population than college students (Chandler, Mueller, Paolacci, in prep) and lends further support to the devaluation-minimizing model.

### 5. General discussion

In light of the zeal with which people treat their own reputations, spontaneous confession seems like a puzzling phenomenon. But it is not: Confessions are issued when they are cheap in the currency of social devaluation—when keeping silent is costly. Otherwise, confessions are withheld. The devaluation-minimizing model of confession is a functionally-grounded model which makes precise and accurate predictions about the motivations underlying the delivery of information.

As an evolutionary perspective would suggest, confessions are governed by the payoffs *to the confessor*. The interests of senders and receivers are sometimes antagonistic (Holtgraves, 1989, study 3; Pinker, 2007). This is often the case with confessions. For instance, a confession is most useful for the remediation of a victim's plans when information about the offense has *not* leaked—when information has leaked, a

**Table 1**
Studies 2 and 3. Willingness to confess (second part of scenario) by study, modulating factor, scenario, and level of modulating factor.

| Study | Modulating factor | Scenario | Level | | t | r |
|---|---|---|---|---|---|---|
| | | | Low | High | | |
| 2 | Cost of offense to victim | Sweater | +0.46 (0.95) | −0.31 (1.30) | 2.57* | .32 |
| | | Sandwich | +0.38 (1.27) | +0.20 (1.40) | 0.57 | .07 |
| | | Ticket | +1.86 (1.27) | −0.50 (1.66) | 6.41*** | .63 |
| | Benefit of offense to offender | Sweater | −1.39 (1.34) | +1.85 (1.33) | −9.53*** | .78 |
| | | Sandwich | +0.13 (0.82) | +0.27 (0.72) | −0.72 | .09 |
| | | Ticket | −0.37 (1.06) | +0.43 (1.25) | −2.81** | .33 |
| | Information leakage | Sweater | −0.42 (0.84) | +0.73 (1.06) | −4.72*** | .60 |
| | | Sandwich | +0.03 (0.49) | +1.00 (1.13) | −4.04*** | .58 |
| | | Ticket | +0.03 (1.00) | +0.68 (0.91) | −2.80** | .32 |
| | Benefit of confession to offender | Sweater [H] | +0.14 (0.65) | +1.36 (1.27) | −4.98*** | .56 |
| | | Sandwich [R] | +0.18 (0.88) | +0.83 (1.21) | −2.50* | .33 |
| | | Ticket [H] | +0.50 (0.78) | +1.10 (0.99) | −2.60* | .32 |
| 3 | Cost of offense to victim | Sweater | +0.70 (0.96) | −0.11 (1.15) | 3.41** | .36 |
| | | Sandwich | +0.12 (1.12) | +0.17 (1.13) | −0.17 | .02 |
| | | Ticket | +1.08 (1.19) | −0.08 (0.98) | 4.59*** | .47 |
| | Benefit of offense to offender | Sweater | −1.11 (0.95) | +1.17 (0.88) | −10.65*** | .78 |
| | | Sandwich | −0.10 (0.82) | +0.14 (0.82) | −1.26 | .15 |
| | | Ticket | −0.64 (1.13) | +0.39 (0.97) | −4.23*** | .45 |
| | Information leakage | Sweater | −0.13 (0.58) | +0.76 (0.75) | −5.82*** | .57 |
| | | Sandwich | −0.20 (0.64) | +0.45 (0.92) | −3.57*** | .40 |
| | | Ticket | −0.03 (0.80) | +0.47 (0.69) | −2.91** | .32 |
| | Benefit of confession to offender | Sweater [H] | +0.26 (0.69) | +1.09 (0.77) | −4.81*** | .50 |
| | | Sandwich [R] | +0.15 (0.79) | +0.79 (0.87) | −3.50*** | .36 |
| | | Ticket [H] | +0.56 (0.88) | +0.65 (0.92) | −0.39 | .05 |

The *Low* and *High* columns show means, with standard deviations in parentheses. Asterisks indicate the significance of the t statistic (*$p < .05$, **$p < .01$, ***$p < .001$). R: benefit of confession modeled as remediability (vs. no remediability). H: benefit of confession modeled as victim valuing honesty (vs. another valuable personality trait).

confession affords little additional remediability to a victim. Also, a confession provides a victim opportunities for larger welfare enhancements when the offense implies a low $WTR_{OV}$—the more injurious the offense, the lower the unnecessary investment in a devalued relationship *if one gets the revelatory information (and acts on it)*. And yet, both when there is little information leakage and when the offense implies a low $WTR_{OV}$, offenders' motivations to confess tend to be depressed, not enhanced. This underscores a manipulative dimension to confessions (see Krebs & Dawkins, 1984) and calls into question the view of spontaneous confession as primarily driven by "personal integrity" or "overriding guilt" (Weiner et al., 1991).

Offenders seem to play by the rule "If I'll get caught, I might as well confess and get some credit" (Bering & Shackelford, 2004). Thus, even when confessions are spontaneous rather than prompted (cf. Weiner et al., 1991), victims face a signal detection problem: Is the confession honest (i.e. motivated by concerns over my welfare), or is it selfishly motivated? Victims are expected to watch for selfishly motivated confessions and to downgrade them accordingly. As cues of information leakage increase, the victim's valuation of the offender's disclosure is expected to decrease. In effect, confessions preceded by accusations (Weiner et al., 1991, experiments 4 & 5) or followed by detection (Utikal, 2012) – entailing information leakage in both cases – lose their mollifying power.

The devaluation-minimizing model of confession was derived from the game-theoretic payoffs of confession to the sender. These payoffs are premised on a rich psychological architecture for welfare tradeoffs that receivers and senders share. According to the welfare tradeoff framework, factors of social value are integrated into target-specific WTRs, which regulate other-impacting behavior. Further, the mind features recalibrational engines designed to adaptively update the values of WTRs (and their factors) in the minds of self and others. The anger psychology of the victim and the nature of the offender–victim relationship provide the backdrop against which the offender makes a communicative decision. This decision bears the stamp of the WTR machinery. Consistent with the hypothesis that anger deals with being on the receiving end of an insufficiently high WTR rather than with harm per se, victim anger is sensitive to offender benefits besides victim costs (Sell, 2005). Mirroring this, the offender's willingness to confess is governed by offender benefits (and victim costs) in anger-minimizing fashion. Notice that the particular effects of benefit derived by offender on victim anger and on offender confession are consistent with the WTR framework but inconsistent with other theories of social outcomes. For instance, for a given victim cost, higher offender benefits represent higher offender–victim inequity, and so equity theory would predict victim distress and anger *increasing* on offender benefit (Sell, 2005; Walster, Berscheid, & Walster, 1973). As applied to confession, equity theory would also make the wrong prediction regarding offender benefits. We note that other cues of WTR should also govern confession. For instance, accidental harm is less diagnostic of WTR than intentional (WTR-consulted, targeted) harm (Sell, 2005), so in general the former should yield more confession than the latter.

The offender's communicative decision is also constrained by the shame architecture. Shame accounts for the default where compromising information about the self, including information about harming a victim, is not only withheld but also keenly guarded. Exceptions to this default exist of course, but they are lawful. In particular, confession is facilitated when compensatory benefits offset the costs of devaluation, retaliation, and reputational loss. (It could be argued that under high information leakage shame *facilitates* confessions to meliorate the net effect of the victim's reaction on the offender[1]).

Finally, confession decisions also seem determined by guilt (Baumeister, Stillwell, & Heatherton, 1994). In this case, it is a concern over the other's welfare and/or the relationship rather than a

devaluation-minimizing goal that motivates confession—although devaluation reduction may be an outcome. Guilt can account for a subset of confession episodes that are puzzling from the perspective of shame. For instance, the study 1 episode where the participant confessed to his friend in spite of the latter incurring high costs is inconsistent with shame but is consistent with guilt and its promotion of disclosure, helping, and relationship maintenance. The effect of remediability on confession is also consistent with guilt. There appears to be distinct selfish and altruistic routes to spontaneous confession (Weiner et al., 1991). In sum, the logic of social valuation makes sense of multiple features of confession.

What is a confession? Broadly defined, it is the disclosure by an offender to his victim that the former has imposed costs on the latter. A narrower definition of confession involves, in addition to the revealing statement, a particular class of attitudes toward it: for example, that the offender regrets having taken the action or will refrain from imposing similar costs in the future (Schlenker & Darby, 1981). These recalibrational attitudes are orchestrated by the emotion of guilt and possibly shame (Schniter & Shields, 2013; Sznycer, 2010). The scope of the devaluation-minimizing model is the narrow class of confessions. To see how this model fails beyond its boundary conditions, consider that in rivalrous or inimical relations, offenders sometimes seek to victoriously disclose their offenses to their targets and third parties and "rub it in" (Chacon & Dye, 2007; Kassin, 1997). Although these broad-sense confessions are lawfully motivated (e.g. by the logic of deterrence or formidability-based status), they are beyond the scope of confessions in the realm of cooperative relationships. Some of the variables that regulate confession, such as the likelihood of information leakage, are not expected to generalize across communicative domains. However, a sensitivity to payoffs is expected to regulate communication more broadly (Tooby & Cosmides, 1990).

The communication sciences have documented a great wealth of design features in natural communication. However, models of the mechanisms that regulate the delivery of information are largely missing (but see Pinker, Nowak, & Lee, 2008).The current studies support the case that an evolutionary–functional framework is a productive and promising approach to uncover the regulatory logic of communication.

## Supplementary Materials

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.evolhumbehav.2014.08.008.

## Acknowledgments

## References

Baumeister, R. F., Stillwell, A. M., & Heatherton, T. F. (1994). Guilt: An interpersonal approach. *Psychological Bulletin, 115*, 243–267.
Bering, J. M., & Shackelford, T. K. (2004). The causal role of consciousness: A conceptual addendum to human evolutionary psychology. *Review of General Psychology, 8*(4), 227–248.
Chacon, R. J., & Dye, D. H. (Eds.). (2007). *The taking and displaying of human body parts as trophies by Amerindians.* New York: Springer.
Chandler, J., Mueller, P., & Paolacci, G. (d). Methodological concerns and advanced uses of crowdsourcing in psychological research. Retrieved from. http://www.jessechandler.com/uploads/2/8/0/5/2805897/mturk_adv_methods.pdf
De Hooge, I. E., Breugelmans, S. M., & Zeelenberg, M. (2008). Not so ugly after all: When shame acts as a commitment device. *Journal of Personality and Social Psychology, 95*(4), 933–943.

---

[1] We thank an anonymous reviewer for this observation.

De Hooge, I. E., Zeelenberg, M., & Breugelmans, S. M. (2011). A functionalist account of shame-induced behaviour. *Cognition & Emotion*, *25*(5), 939–946.

Delton, A. W., Robertson, T. E., Sznycer, D., Lim, J., Cosmides, L., & Tooby, J. (2008). An evolved internal regulatory variable for making welfare tradeoffs. *Paper presented at the 20th Annual Human Behavior and Evolution Society Conference, Kyoto, Japan.*

Fessler, D. M. T. (2001). Emotions and cost/benefit assessment: The role of shame and self-esteem in risk taking. In R. Selten, & G. Gigerenzer (Eds.), Cambridge, MA: MIT University Press.

Gilbert, P. (1998). What is shame? Some core issues and controversies. In P. Gilbert, & B. Andrews (Eds.), New York: Oxford University Press.

Hamilton, W. D. (1964). The genetical evolution of social behaviour. *Journal of Theoretical Biology*, *7*(1), 1–52.

Hammerstein, P., & Parker, G. A. (1982). The asymmetric war of attrition. *Journal of Theoretical Biology*, *96*, 647–682.

Holtgraves, T. (1989). The form and function of remedial moves: Reported use, psychological reality and perceived effectiveness. *Journal of Language and Social Psychology*, *8*(1), 1–16.

Jones, B., & Rachlin, H. (2006). Social discounting. *Psychological Science*, *17*(4), 283–286.

Kassin, S. M. (1997). The psychology of confession evidence. *American Psychologist*, *52*(3), 221.

Kassin, S. M., & Gudjonsson, G. H. (2004). The psychology of confessions. A review of the literature and issues. *Psychological Science in the Public Interest*, *5*(2), 33–67.

Krebs, J. R., & Dawkins, R. (1984). Animal signals: Mind-reading and manipulation. In J. R. Krebs, & N. B. Davies (Eds.), *Behavioural ecology: An evolutionary approach* (pp. 380–402). Oxford: Blackwell Scientific Publications.

Kuran, T. (1998). *Private truths, public lies: The social consequences of preference falsification.* Cambridge, MA: Harvard University Press.

Lachmann, M., Számadó, S., & Bergstrom, C. T. (2001). Cost and conflict in animal signals and human language. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(23), 13189–13194.

Lim, J. (2012). *Welfare tradeoff ratios and emotions: Psychological foundations of human reciprocity.* Doctoral dissertation. Santa Barbara: University of California.

Maynard-Smith, J., & Harper, D. (2003). *Animal signals.* Oxford: Oxford University Press.

McCullough, M. (2008). *Beyond revenge: The evolution of the forgiveness instinct.* San Francisco, CA: Jossey–Bass.

Murdock, B. B. J., & Bennet, B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, *64*(5), 482–488.

Peterson, R. A. (2001). On the use of college students in social science research: Insights from a second-order meta-analysis. *Journal of Consumer Research*, *28*, 450–461.

Pinker, S. (2007). The evolutionary social psychology of off-record indirect speech acts. *Intercultural Pragmatics*, *4*(4), 437–461.

Pinker, S., Nowak, M. A., & Lee, J. J. (2008). The logic of indirect speech. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(3), 833–838.

Scarnier, M., Schmader, T., & Lickel, B. (2009). Parental shame and guilt: Distinguishing emotional responses to a child's wrongdoings. *Personal Relationships*, *16*, 205–220.

Schelling, T. C. (1960). *The strategy of conflict.* Cambridge, MA: Harvard University Press.

Schlenker, B. R., & Darby, B. W. (1981). The use of apologies in social predicaments. *Social Psychology Quarterly*, 271–278.

Schniter, E., & Shields, T. (2013). Recalibrational emotions and the regulation of trust-based behaviors. In D. Gefen (Ed.), *Psychology of trust: New research.* New York: Nova Science Publishers.

Sell, A. (2005). *Regulating welfare tradeoff ratios: Three tests of an evolutionary–computational model of human anger.* Doctoral dissertation. Santa Barbara: University of California.

Sell, A., Tooby, J., & Cosmides, L. (2009). Formidability and the logic of human anger. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(35), 15073–15078.

Smith, R. H., Webster, J. M., & Eyre, H. L. (2002). The role of public exposure in moral and nonmoral shame and guilt. *Journal of Personality and Social Psychology*, *83*(1), 138–159.

Sznycer, D. (2010). *Cognitive adaptations for calibrating welfare tradeoff motivations, with special reference to the emotion of shame.* Doctoral dissertation. Santa Barbara: University of California.

Sznycer, D., Takemura, K., Delton, A. W., Sato, K., Robertson, T., Cosmides, L., et al. (2012). Cross-cultural differences and similarities in proneness to shame: An adaptationist and ecological approach. *Evolutionary Psychology*, *10*(2), 352–370.

Sznycer, D., Tooby, J., & Cosmides, L. (2009). The regulatory theory of shame. *Paper presented at the 21st Annual Human Behavior and Evolution Society Conference, Fullerton, CA.*

Tangney, J. P., Miller, R. S., Flicker, L., & Barlow, D. H. (1996). Are shame, guilt, and embarrassment distinct emotions? *Journal of Personality and Social Psychology*, *70*(6), 1256–1269.

Tangney, J. P., Stuewig, J., & Martinez, A. G. (2014). Two faces of shame: The roles of shame and guilt in predicting recidivism. *Psychological Science*, *25*, 799–805.

Tooby, J., & Cosmides, L. (1990). The past explains the present: Emotional adaptations and the structure of ancestral environments. *Ethology and Sociobiology*, *11*, 375–424.

Tooby, J., & Cosmides, L. (1996). Friendship and the Banker's Paradox: Other pathways to the evolution of adaptations for altruism. In W. G. Runciman, J. Maynard Smith, & R. I. M. Dunbar (Eds.), *Evolution of social behaviour patterns in primates and man. Proceedings of the British Academy, vol. 88.* (pp. 119–143).

Tooby, J., Cosmides, L., Sell, A., Lieberman, D., & Sznycer, D. (2008). Internal regulatory variables and the design of human motivation: A computational and evolutionary approach. In A. J. Elliot (Ed.), Mahwah, NJ: Lawrence Erlbaum Associates.

Utikal, V. (2012). A fault confessed is half redressed—Confessions and punishment. *Journal of Economic Behavior & Organization*, *81*(1), 314–327.

Van Lange, P. A., & Kuhlman, D. M. (1994). Social value orientations and impressions of partner's honesty and intelligence: A test of the might versus morality effect. *Journal of Personality and Social Psychology*, *67*(1), 126–141.

Walster, E., Berscheid, E., & Walster, G. (1973). New directions in equity research. *Journal of Personality and Social Psychology*, *25*, 151–176.

Weiner, B., Graham, S., Peter, O., & Zmuidinas, M. (1991). Public confession and forgiveness. *Journal of Personality*, *59*(2), 281–312.

Wicker, F. W., Payne, G. C., & Morgan, R. D. (1983). Participant descriptions of guilt and shame. *Motivation and Emotion*, *7*(1), 25–39.

Williams, G. C. (1966). *Adaptation and natural selection: A critique of some current evolutionary thought.* Princeton, NJ: Princeton University Press.