Perception and Externalization of Group-Based Criticism and Praise

Sarah Kerns

Wellesley College

Author Note:

Sarah Kerns, Psychology Major, Wellesley College.

I thank my teammates Sarah Gladstone, Hawah Kallon, Samantha Muller, and Tara Wilfert for their substantial work in helping to design and conduct this study. Special thanks to Dr. Robin Akert for her teaching and work editing previous versions of this manuscript, Dr. Jeremy Wilmer for his help in seeing both trees and forest, and Taylor Hood for her patient aid and administration.

Correspondence concerning this research should be sent to the author.

Social Identity theory was defined by John Turner and Henri Tajfel in the 1970s to differentiate group and individual drive behaviors in the process of social interaction. The theory held that individuals ally with groups to manipulate their personal social position, choose these groups based on social status relative to other groups; and that once affiliated, individuals incorporate group standing into their own self-image, meaning that they have a vested interest in the group's social success and elevation. (Turner & Tajfel 1979). This interest has been documented to show itself in two ways: in-group favoritism and out-group discrimination.

Tajfel's theory finds its roots in sociology, and builds upon the work of Emile Durkheim and William Sumner in particular. [Durkheim, 1969) Durkheim's research into the social and societal impacts of belonging to certain groups, generally religious and political, laid the foundation for much of the discussions of group behavior today, while Sumner's expansion of the field in the early 1900s via his introduction of the concept of social mores, or "folkways", provided the basis for the study of group distinction. Though Sumner's work was focused on the larger-scale study of how each world culture has differing mores, his exploration of these customs as definitional to their resident groups can be seen as a stepping stone toward Tajfel's exploration of group differentiation. (Sumner, 1883)

Researchers have documented group support of, and advocacy by, members toward in-group individuals through favoritism. This in-group favoritism is aimed at protecting and raising the group's social standing by raising the social status and affiliation of other members. The process has also been observed inversely: where derision of, or discrimination against, out-group members or other groups is used to raise comparative in-group reputation both on an individual and societal level (Tajfel 1970).

One area of Social Identity Theory that has gotten much attention is that of group formation. Tajfel and Turner attempted to explore the boundaries of group formation using low-commitment study groups where random groups of strangers were assigned membership by researchers. They found that without any acquaintance or reason for grouping, subjects affiliated almost immediately, attributing positive qualities to in-group members, and asserting preference for them as compared to a similarly unknown out-group. (Turner & Tajfel 1979; Durrheim, et al 2016),

Researchers hypothesize that self-esteem motivates this behavior, because a positive impression of one's group allows transference of that positivity to self, thus improving self-opinion by membership in an admirable group. Additionally, if one can cause others to view ones in- group in a positive light, they may benefit from a group-wise halo effect and be viewed positively as a member of the good group (Nufur & Alessi, 2018).

Because group identity is gauged almost exclusively by comparison, out-group dynamics and competition hold essential places in consideration of the topic. Although in-group allegiance has been demonstrated to impede inter-group functioning, Saguy and Halperin's 2013 study, ***Exposure to outgroup members criticizing their own group facilitates inter-group openness***, uses negative and neutral commentaries about in-groups to understand group dynamics under conflict, and what each group feels about the other when they hear criticism. (Saguy and Halperin, 2013).

They found that, in the case of the conflicting groups of Israelis and Palestinians, Israelis who had been primed by hearing Palestinians criticize Palestinian policies were more likely to be open to believing Palestinian narratives of the conflict--demonstrating that if a one hears a member of a conflicting outgroup voice self-group criticism, they are more likely to empathize with the out-group. Whether this is due to the subject feeling as if the critic is now less biased towards them, or that he/she is expanding their in-group to include the subject, has yet to be determined. (Saguy and Halperin, 2013).

O'Dwyer, Berkowitz and Johnson's 2002 study, ***Group and person attributes in response to criticism of the in-group*** used college students comparing schools of different status to see what types of reputation affected in-group loyalty and interpretation of criticism. Particularly interesting was the finding that criticism, (in the form of an article written about the subject's school) coming from members of a high prestige institution, was viewed by subjects as general high-prestige opinion, while comments from low-prestige sources were attributed to individual unfriendly people (O'Dwyer et. al. 2002).

In the context of a society increasingly informed by social media, we decided to explore the extent to which in-group favoritism could outweigh strong positive and negative messages publically leveled at a group, and whether the group membership of the author of such messages altered or counteracted that favoritism. There have been mixed results on studies involving self-criticism by in-groups: some studies have demonstrated evidence of in-group resiliency, while others have not (Heyman, 2013). So we set out to test this again in our study.

Based on our research participants' membership to a college in-group (Wellesley College), we chose an outgroup of similar social stature, with a specifically non-rival social status to the participants' school, and well-known to the students: MIT (O'Dwyer, et. al. 2002). Next, we created two pages of comments about Wellesley College: one complimentary, the other derogatory, which were represented as actual online commentary authored by

either the in-group (Wellesley College students) or the outgroup (MIT students) (O'Dwyer, et. al. 2002). Research participants (Wellesley College students) were primed with a video framing Wellesley in a positive light as well as personal recall exercises designed to call forward happy memories of Wellesley College, then randomly assigned to one of four groups to read comments about their school. The groups were: positive comments authored by Wellesley students, negative comments authored by Wellesley students, Positive comments authored by MIT students, or negative comments authored by MIT students.

In order to obtain a wide range of data, participants were asked to rate their opinions of both in and out groups using Quantitative measures, (Pennington, et. al. 2015) in addition to Qualitative measures aimed at Social Responses (Herak, et. al. 2018) and a resource allocation task (Durrheim, 2016) rating only the Confession author's group.

We asked to what extent out-group authorship or positivity of comments would affect opinion about public comments made about a group, and hypothesized that in-group membership and in-group authorship would be protective of good opinion. We were also interested in whether there would be a shift in estimations of agreeability or reliability based on comments or condition, and hypothesized that agreeability would rise in positive comment groups, and that reliability would fall in negative commentary groups.

METHOD

Overview

This experiment was a 2x2x2 mixed factorial design. The first independent variable was the college population being rated by research participants (Wellesley College students or MIT students). The second independent variable was valence of comments (positive /negative) made about Wellesley College in a public forum. The third independent variable authorship of these public comments (by in-group- Wellesley College or out-group- MIT). All participants were Wellesley College students, in-group primed by viewing a video compilation about the college and answering a 15 question Wellesley priming questionnaire. The first independent variable was operationalized by administering IAT tasks rating either Wellesley students or MIT students, while the second and third were operationalized using printed commentary pages (Confessions). The dependent variable (opinion about a group after exposure to social commentary) was measured in three ways: with a timed Implicit Association Test (IAT); an untimed Social Judgement questionnaire; and a simple resource-assignment task.

Participants

Study subjects were 61 female students of Wellesley College, 25 were first year students, 17 were sophomores, 7 were juniors, and 4 were seniors. The mean participant had completed 0.81 years of college (M= 0.81, SD=0.89). 59 participants identified as female, 1 as non-binary; and 10% as Black/African, 10% as Latin/Hispanic, 27% as White, and 51% as Asian. Participants were run in small groups of 1-6, and each session lasted approximately 40 minutes. Four participants' data had to be dropped (and replaced by new participants) due to test interruption by fire drill, two participants' data was dropped due to failure on the manipulation check questions.

Participants volunteered to take part in the study, and received research participation credit toward their psychology course requirement.

Procedure

Participants were told a cover story explaining the study as an investigation of cognitive processes regarding Wellesley College, or other Boston area schools, as affected by exposure to social media. Participants were told that past studies had found social media capable of influencing emotional states and cognition. They were informed that they would be reading Confessions pages and responding to a questionnaire, a cognitive speed task, and a cognitive game related to their reading.

Participants were assured that their data would be compiled anonymously using randomly assigned ID numbers and were instructed to omit their name from all documents. Each group was informed that participation was voluntary, that their session would take approximately 45minutes, and that subjects would earn a research participation credit. Finally, subjects were informed that they would be provided further study information, including our hypotheses, at the conclusion of the study. Each participant signed an informed consent form.

Priming in-group Identity/Videos

Each study participant was primed for in-group identity by viewing the same 12-minute compilation video. This video was built to encourage identification of the subjects as Wellesley College students, and consisted of 4 parts. The first was excerpted from student-produced Wellesley College news satire program, BoobTube, and contained many in-jokes and common campus themes. The second was excerpted from a professionally produced black and white video interview with Wellesley Alumna, and former Secretary of State, Hillary Clinton. During the interview, Secretary Clinton described the benefits she accrued from

attending Wellesley College. Third was a professionally produced video interviewing eager first year students of Wellesley College and their families on move-in day. They discussed their hopes regarding Wellesley College and how it would help them to succeed after college. And finally, fourth, a professionally produced Wellesley College advertisement called "This is the Place", concentrating on Wellesley as a place where students find their calling. Each session of participants watched the video as a group and participants were instructed not to talk during the presentation.

Priming in-group Identity/Questionnaire

After watching the Wellesley College in-group priming video, all participants responded to fifteen short answer questions about Wellesley College. The questions, delineated below, served to further prime in-group membership as members of the Wellesley College community.

1. *Are you a Wellesley College student? (Yes/No)*
2. *What is your major?*
3. *What psychology class are you currently in?*
4. *What is your class color? (Red/Green /Yellow /Purple)*
5. *Do you live in on campus? If so, what is your dorm?*
6. *What is your favorite dorm to eat in on campus?*
7. *Are you in any orgs on campus? If so, which ones?*
8. *What is your favorite Wellesley tradition and why*?
9. *Have you ever attended a Wellesley sport event?*
10. *Pub, Hoop, or El Table? (Pub/Hoop/El Table/None)*
11. *Who is your favorite Wellesley professor and why?*
12. *How many times have you walked around Lake Waban? (0/1/2/3/4/5/6/7/8/9/10/>10)*
13. *How many of the "50 Things to Do Before You Graduate" have you accomplished? (0/ 1-10/ 11-20/ 21-30/ 31-40/ 41-50)*
14. *What is your favorite location on campus and why?*
15. *What is your favorite Wellesley memory?*

Administering and testing Independent variables:

College Being Rated

The first independent variable was the college population that research participants were rating. Each participant was administered both an IAT task to rate "Wellesley students" and a second IAT task to rate "MIT students".

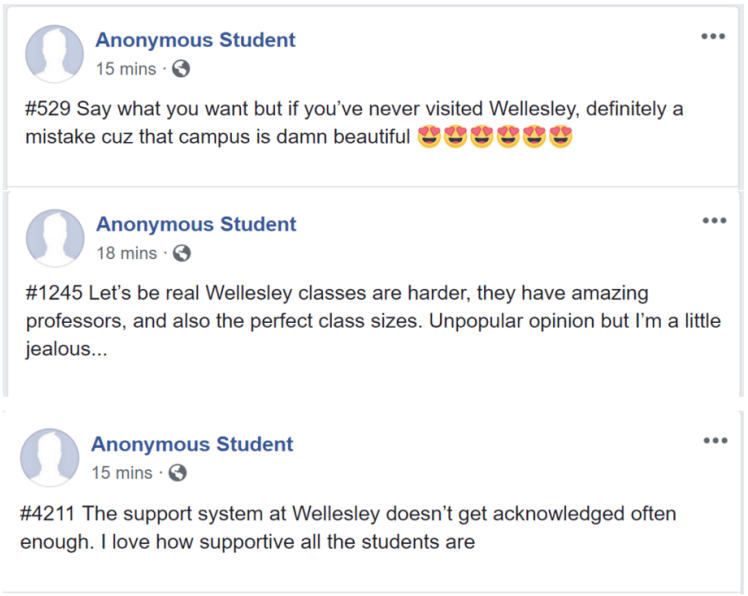
Valence of Confessions-

The second independent variable: was valence of commentary about Wellesley students (positive=complimentary or negative=derogatory) and was operationalized using researcher-created "Confessions" pages. *Confessions forums* were defined for participants as anonymous online groups where individuals share honest feelings, beliefs, or worries with their community. Each participant was given a printed page containing either 12 complimentary or 12 derogatory comments about Wellesley students which they were told had been taken from actual college Confessions forums.

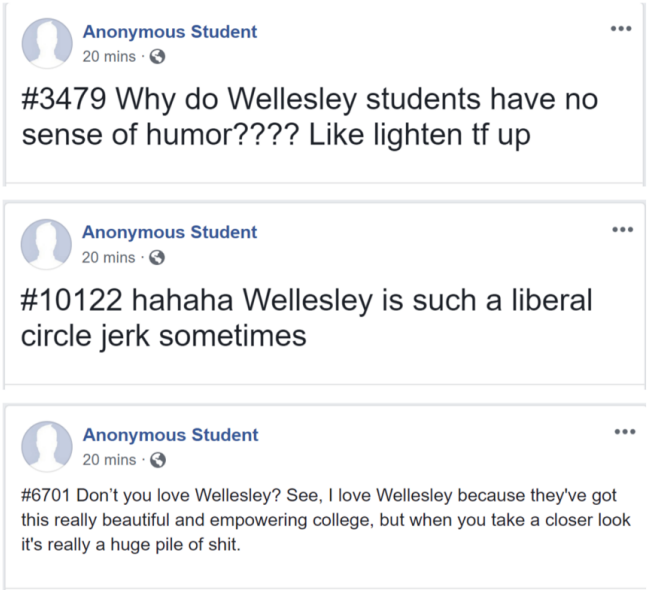
The Confessions page comments were, in fact, gathered from social media confessions pages, primarily from Wellesley College, and were lightly edited by the research team for clarity, and to meet the independent variable conditions. Participants were warned that some of the confessions contained explicit language or topics and were asked to read through them carefully.

Examples of the comments are:

Positive valence (complimentary):



Negative valence (derogatory):



Confessions' Author Group

The third independent variable was authorship of the Confession comments (Wellesley College/ MIT). The pages contained comments about Wellesley College, the author group therefore represented either in-group commentary (Wellesley students commenting about Wellesley College) or out-group commentary (MIT students commenting about Wellesley College). This variable was operationalized using college logos at the top of each Confessions page: one set of complimentary and derogatory Confessions pages had the Wellesley College logo at the top, and the comments referred to "we" and "us". The other set had the MIT college logo at the top, and the (same) comments were edited to refer to "them" and "they".

Collectively, the Author and Valence were coded as conditions A, B, C & D:

Table 1: Stimulus Conditions

Condition	Confession's Author Group	Valence
A	Wellesley student	Positive (complimentary)
B	Wellesley student	Negative (derogatory)
C	MIT student	Positive (complimentary)
D	MIT student	Negative (derogatory)

Measures

Implicit Associations Test (IAT)

After reading the Confessions page, and in order to measure the dependent variable, participants completed an IAT task on the computer. Using the Tellab program, 20 adjectives were associated (or not) with the student group being rated. Both the association, and the time required to make the association (in milliseconds) were recorded. IAT task One asked participants to associate (or not associate) the twenty adjectives with "Wellesley students", and a structurally identical second IAT task asked participants to associate (or not associate) the same 20 words with "MIT students". All participants completed both IAT tasks, as follows:

Participants were first trained to signal association of words with a given category using keyboard directional keys. For instance, "Wellesley" and "Wells" were words associated with the "Wellesley student" category. In the second IAT, "MIT" and "mit" were words associated with "MIT student" category. Participants were penalized with a red X and required to answer correctly if they failed to associate college-representative words with the college.

Next, participants were familiarized with all 20 dependent measure words in random order. The words were: *friendly, tolerant, likeable, thoughtful, supportive, principled, dependable, credible, conscientious, confrontational, rude, cold, aggressive, unpleasant, impulsive, dishonest, fickle, biased, unreasonable.* This introduction was click-through only, there were no categories to associate, and no penalty was used.

Finally, participants were asked to use keyboard directional keys to associate these 20 words with either the college ("Wellesley students" in IAT one, "MIT students" in IAT two) or with NOT. There was no red X penalty. College-specific associations used in the IAT association training section ("Wells", "mit", etc.) including their penalty Xs, were interspersed with the 20 words, as attention monitors. Response time in milliseconds was collected on all items, and error rate was collected for attention check items only, by the Tellab program.

On completion of the Wellesley College IAT, a second, and functionally identical, MIT AIT was administered to the same participant. Each IAT took approximately 5 minutes to complete.

Social Judgement Questionnaire

After completing the IAT, participants responded to a computer questionnaire which asked them to rate the Confession Authors' groups on the same twenty dependent measure adjectives used for both IATs. Ratings were given on a 7 point Likert scale, normalized for valence, where 1 = lowest positive association and 7 = highest positive association. Subjects used a computer mouse to click on a single 1-7 value for each adjective.

Money Game

The final data set was a resource assignment task where participants were asked to write an amount in response to one of the following prompts. Either: 1. "Imagine that you are given $20. You have the choice of keeping the money for yourself, or giving some to another Wellesley student. How many dollars would you give to the Wellesley student?" or 2. "Imagine that you are given $20. You have the choice of keeping the money for yourself, or giving some to an MIT student. How many dollars would you give to the MIT student?". Prompt 1 or 2 was administered to match their confession page author, with each participant offered a chance to gift money to the author group of the Confessions she had read.

Manipulation Check Questionnaire

In a computer questionnaire, participants chose answers for the following questions to assure in-group membership and delivery of the independent variables: Did you read statements written by Wellesley or MIT students? (Wellesley/MIT); Did you read positive or negative statements about Wellesley College students? (positive/negative); Are you a Wellesley College student? (yes/no)

Demographic Questionnaire:

Participants were asked to indicate their gender, ethnicity, and class year. After completing all study components, participants were debriefed and thanked for their participation, and told that they could learn about the study's results at the end of the semester.

Results

Manipulation check

The study employed three manipulation check questions. To check for in-group membership participants were asked if they attend Wellesley College, and to check for delivery of independent variables I and 2, they were asked

the Valence of the confession page they had read (positive or negative); and College community represented by their confession page (Wellesley College or MIT).

100% of sample were Wellesley College students, 97% of the sample correctly identified the valence of the confessions they had read, and 97% of the sample correctly identified the school the confessions' authors attended. The 2 research participants who did not correctly identify the Authors' Group Membership or confession valence were deleted from the sample. These results indicate a successful operationalization for the first two independent variables. 100% of participants completed an IAT rating either Wellesley or MIT students, and 95% completed and IAT rating task for both Wellesley and MIT students, indicating a successful operationalization of the third independent variable.

Reliability analysis of dependent measure adjectives

Reliability analyses were conducted on four sub-groupings of the twenty dependent measure adjectives, each group composed of 10 adjectives. All four analyses indicated a high degree of reliability: Cronbach's r = .93 for the Positive adjectives (*Friendly, Tolerant, Likeable, Thoughtful, Supportive, Trustworthy, Principled, Dependable, Credible, Conscientious*); Cronbach's r = .90 for the Negative adjectives (*Confrontational, Rude, Cold, Aggressive, Unpleasant, Impulsive, Dishonest, Fickle, Biased, Unreasonable*); Cronbach's r = .95 for the Agreeable adjectives (*Friendly, Unpleasant, Tolerant, Confrontational, Likeable, Cold, Thoughtful, Rude, Supportive, Aggressive*). Cronbach's r = .86 for the Reliable adjectives (*Trustworthy, Dishonest, Principled, Biased, Dependable, Fickle, Credible, Unreasonable, Conscientious, Impulsive*)

Implicit Association Test:

The first measure used to evaluate the dependent variable of Opinion about a group after exposure to social commentary, was an IAT. The data gathered were per-word count of adjective use, and time required to associate the adjective with its subject (in milliseconds). The adjective tallies evaluated herein are negative associations, i.e. use of an adjective as criticism, therefore, a higher mean indicates higher number of times the word was used to criticize.

College Rated (IV 1)

The analysis of variance for repeated measure of 20 adjectives across all groups demonstrate a statistically significant main effect for the College Rated (independent variable 1). ($F(1,599) = 17.8$, $p<.001$). Participants rating Wellesley College students gave significantly fewer criticisms (M=.22, SD=.41) than those

rating MIT students (M=.31, SD=.46), as hypothesized. Additionally, there was a statistically significant main effect for Confession page stimulus condition, (F (3,1797) = 19.56, p<.001) meaning that the Confessions page stimulus condition had an effect on level of criticism. This will be elucidated more clearly in the comparison of Confession page conditions per College Rated, below. Finally, there was a statistically significant interaction between College Rated and Confession page condition (F (3,1797) =15.19, p<.001). Showing that IAT criticisms differ depending on what college is being rated, but that the level of that those difference varies based on condition (see Figure 1). We observe Authors getting more credit (or blame) for their Confessions pages when their own school is being Rated. We have named this 'The Blame Effect'.

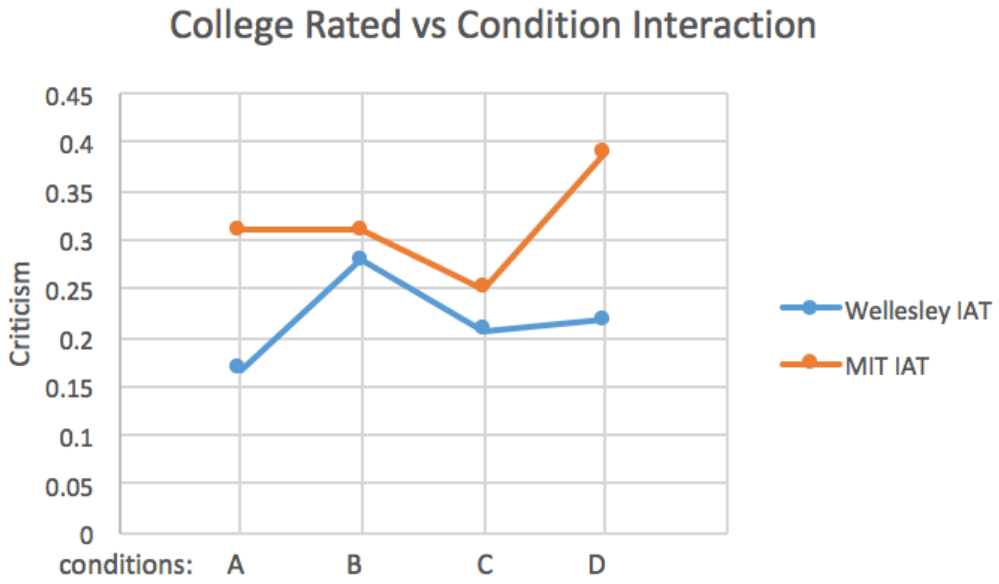


Figure 1: College Rated vs Condition Interaction, Authors get more credit (or blame) if their school is being Rated, while Authors from the other school, get a standard score, relatively independent of Valence.

Author of Confessions page, and Valence of Confessions (IVs 2&3)

When evaluating the IATs for each college individually, it is evident that Wellesley College students showed a statistically significant main effect for Valence of Confession (F (1,599) = 85.57, p<.001) with authors of complimentary comments garnering fewer negative ratings (M=.19, SD=.39) than those of derogatory comments (M=.25, SD=.43) and while there was no main effect for Authorship of Confessions (see Table 4), there was a statistically significant interaction between Valence and Authorship of Confessions pages (F (1,599) = 48.37, p<.001), specifically in the case of Wellesley College Authors, who were less criticized when giving complimentary confessions (M=.17, SD=.37), and more criticized when giving

derogatory confessions (M=.28, SD=.45) than either MIT complimentary authors (M=.21, SD=.41) or MIT derogatory authors (M=.22, SD=.41) when rated on the Wellesley College IAT.(see figure 2a)

IATs rating MIT students showed similar results, demonstrating a statistically significant main effect for Valence of Confession (F (1,599) = 58.86, p<.001) with Authors of complimentary Confession pages garnering fewer criticisms (M=.28, SD=.45) than those of derogatory Confessions (M=.35, SD=.48), there was no main effect for Authorship of Confessions (see Table 2), but a statistically significant interaction between Valence and Authorship of Confessions pages (F (1,599) = 46.29, p<.001), with MIT authors being less criticized when giving complimentary Confessions (M=.25, SD=.43), and more criticized when giving derogatory Confessions (M=.39, SD=.49) than either complimentary or derogatory (M=.31, SD=.46) Wellesley College authors who were almost identical in the criticism that they received using the MIT IAT. (see figure 2b)

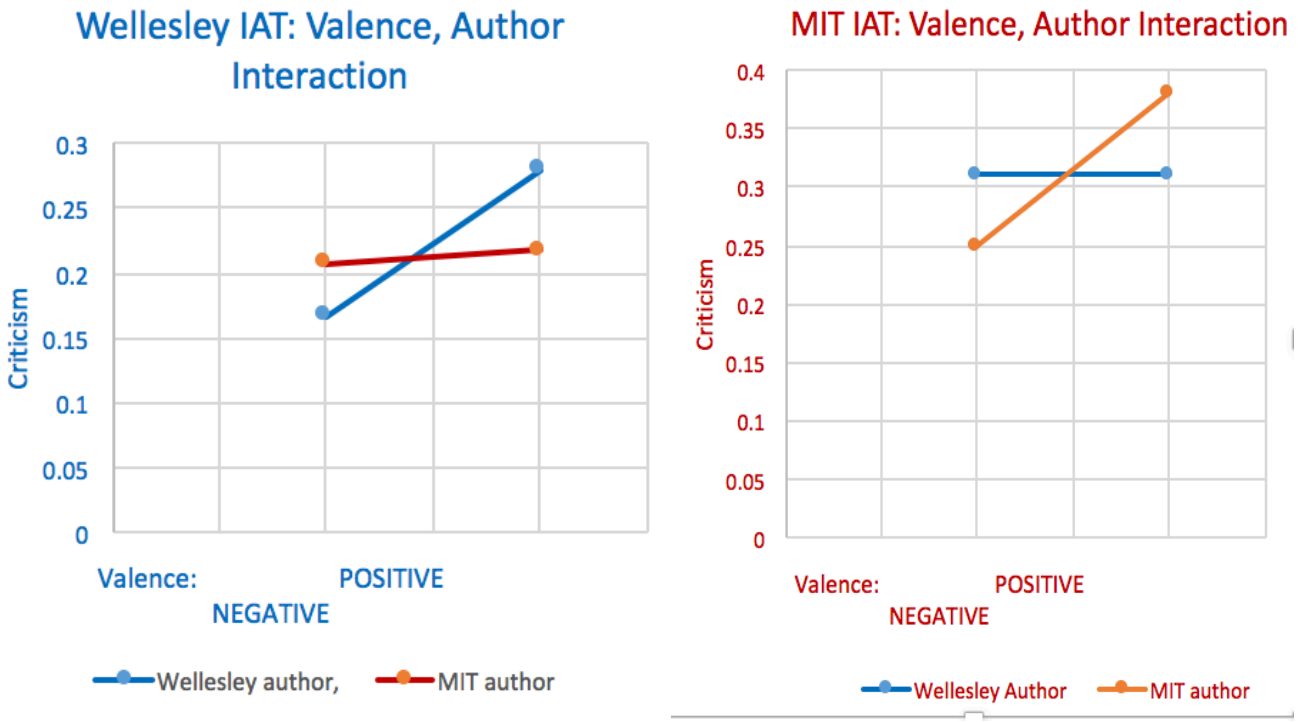


Figure 2: Valence vs Author Interaction displaying the Blame effect. Blue- Wellesley College IAT, Maroon- MIT AIT. Authors get more credit, or blame when their own school is being rated.

DV: Opinion about a group after exposure to social commentary

In evaluating the dependent variable of Opinion about a group after exposure to social commentary, two categories were used: Type of Criticism (Direct or Indirect) and Characterization (Agreeable and Reliable).

Type of Criticism (Direct criticism via negative adjectives, or Indirect criticism via positive adjectives) showed a statistically significant main effect ($F(1,2359) = 275.53$, $p<.001$) with Direct criticism more present across all cases ($M=.17$, $SD=.38$) than Indirect ($M=.095$, $SD=.29$). Additionally, there was a main effect for College which we have seen previously ($F(1,2359) = 226.90$, $p<.001$) with Wellesley College ratings containing statistically significant lower levels of criticism overall ($M=.11$ $SD=.31$), as compared to MIT ($M=.16$, $SD=.36$). There was a statistically relevant interaction between College and Type of Criticism ($F(1,2359) = 87.69$, $p<.001$) with differences between Direct and Indirect criticisms at Wellesley College being greater: Direct ($M=.16$, $SD=.37$) vs Indirect ($M=.06$, $SD=.23$) than those at MIT: direct ($M=.18$, $SD=.38$) vs indirect ($M=.13$, $SD=.34$). (see figure 3).

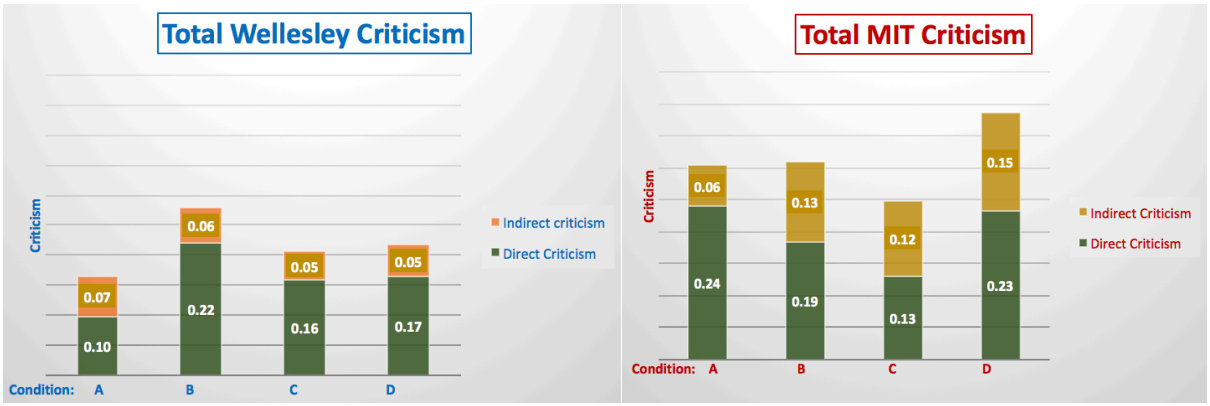


Figure 3 Criticism divided by direct and indirect, per college

Looking at this data via proportionally, by showing what proportion of total criticism is indirect (and by inference, since criticism is either direct or indirect, what proportion is direct).Displays an interesting pattern. (see Figure 4)

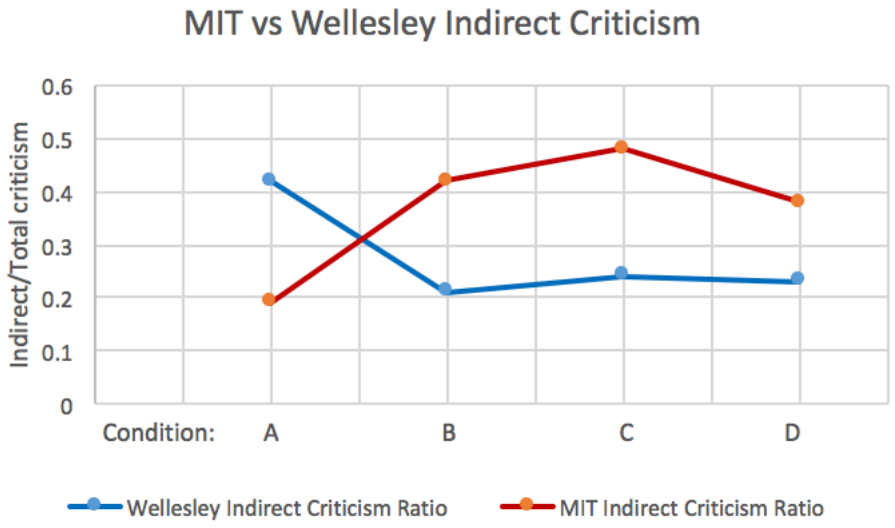


Figure 4: MIT vs Wellesley Indirect Criticism Ratio- Wellesley A, and MIT B, C & D have similar (high proportions of indirect (gentler) criticism, while Wellesley B, C &D and MIT A are similar in proportion, with a less-gentle higher proportion of direct criticism.

The dependent variable of Characterization (via Agreeability or Reliability group adjectives) showed a statistically significant, main effect ($F(1,2359) = 30.58$, $p < .001$)) with criticisms using Agreeability adjectives (i.e. characterizations of disagreeability) ($M=.12$, $SD=.32$) surpassing those using Reliability adjectives (i.e. characterizations of unreliability) ($M=.10$ $SD=.31$). And a main effect for College ($F(1,2359) = 65.69$, $p<.001$) as we have seen before, with total negative Characterizations falling more heavily on MIT ($M=.32$, $SD=.16$), than Wellesley College ($M=.31$, $SD=.10$). The statistically significant interaction ($F(1,2359) = 58.45$, $p<.001$) of College Rated on Characterizations of unreliability, showed Wellesley was rated generally more Reliable ($M=.09$, $SD=.29$) than MIT. Between conditions, there were also significant effects for agreeability, based on College ($F(1,599) = 55.88$, $p<.001$) and Condition ($F(1,599) = 28.23$, $p<.001$) with a significant interaction ($F(1,599) = 21.37$, $p<.001$) (see figures 5&6). While Wellesley disagreeable scores were relatively stable, Wellesley students were rated as slightly less agreeable when Authors were Wellesley and the Blame effect seemed to be absent, while MIT scores fluctuated more, again showing the blame effect.

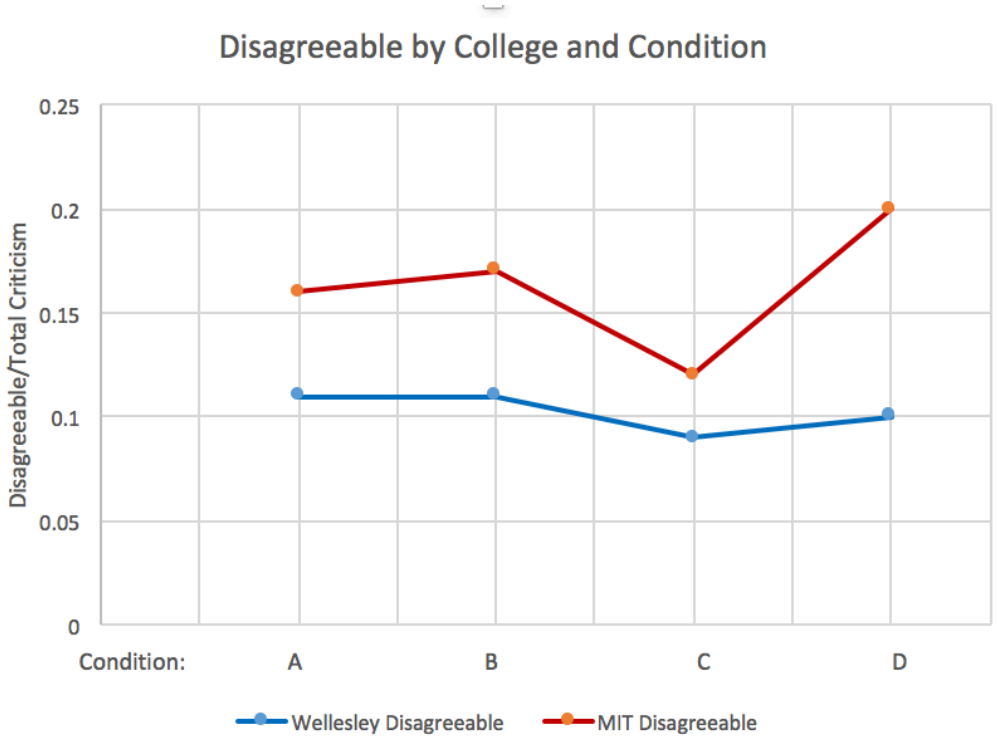


Figure 5- Disagreeable by college and Condition- outgroup shows blame effect again, while in-group remains steady with slight increase on disagreeability during self-commenting

Reliability of Wellesley students increased when self-rating, and decreased when rating the out-group; Wellesley was counted most reliable, and MIT least reliable when offering negative self-commentary.

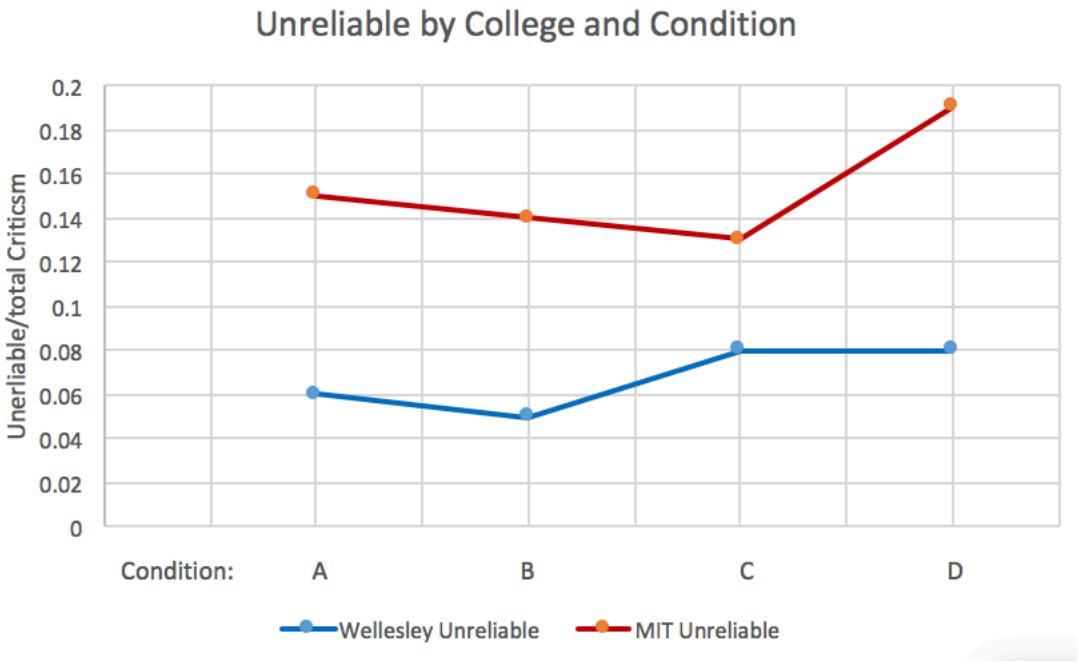


Figure 6: Unreliable by College and Condition- Wellesley counted most reliable, and MIT least reliable when offering negative self-commentary

Social Judgement Questionnaire

The Dependent variable of Opinion about a group after exposure to social commentary was next measured by a questionnaire. The data gathered were: level of positivity associated with their Confessions page author on a 1-7 Likert scale with 1 being least positive and 7 being most positive. For the questionnaire statistics, therefore, a higher mean indicates a higher level of positivity felt towards Confessions Author.

The analysis of variance for association with the Positive adjectives indicated a significant main effect for Valence of Confessions ($F(1,56) = 66.10$, $p < .001$). Participants who had read complimentary (positive valence) comments about Wellesley College rated the Confessions' authors significantly more favorably ($M = 4.81$, $SD = 0.96$) than did those who had read derogatory (negative valence) comments ($M = 2.96$, $SD = 0.79$), as hypothesized. The main effect for Confessions Author Group and the interaction effect were not significant (see Table 2).

The analysis of variance for the Negative adjectives indicated a significant main effect for Valence of Confessions ($F(1,56) = 95.96$, $p < .001$) Participants who had read complimentary (positive valence) comments about

Wellesley College rated the Confessions' authors significantly more favorably (M = 5.07, SD = 0.82) than did those who had read derogatory (negative valence) comments (M = 3.01, SD = 0.79) as hypothesized. The main effect for Confessions Author Group and the interaction effect were not significant (see Table 2).

The analysis of variance for the Agreeable adjectives indicated a significant main effect for Valence of Confessions ($F(1,56) = 141.90$, $p < .001$). Participants who had read complimentary (positive valence) comments about Wellesley College rated the Confessions' authors significantly more favorably (M = 5.45, SD = 0.88) than did those who had read derogatory (negative valence) comments (M = 2.71, SD = 0.87), as hypothesized. The main effect for Confessions Author Group and the interaction effect were not significant (see Table 2).

The analysis of variance for the Reliable adjectives indicated a significant main effect for Valence of Confessions ($F(1,56) = 34.76$, $p < .001$). Participants who had read complimentary (positive valence) comments about Wellesley College rated the Confessions' authors significantly more favorably (M = 4.42, SD = 0.83) than did those who had read derogatory (negative valence) comments (M = 3.26, SD = 0.68), as hypothesized. The main effect for Confessions Author Group and the interaction effect were not significant (see Table 2).

Money Game

The analysis of variance of Money Game results indicated no significant effects (see Table 4). However, participants in general gave more money to themselves than to the other person (M=5.57, SD= 4.56).

Discussion

In our investigation of in and out-group dynamics, we asked to what extent out-group authorship or positivity of public comments regarding a group would affect opinion (as measured by associating 20 adjectives with or against the group). We hypothesized that in-group membership and in-group authorship would be protective of opinion across conditions. We also asked whether there would be a shift in associations with words denoting agreeability or reliability based on stimulus condition (valence and authorship), and hypothesized that agreeability would rise in positive comment groups, and that reliability would fall in negative commentary groups.

In comparing the results of the Implicit association test, Questionnaire, and resource assignment tasks, we see sensitivity decrease and possibility for confounds increase as we go further away from pure quantitative measure. The IATs provided a more sensitive measure, with the timed aspect of the test allowing less self-censoring and therefore

provided a more honest (albeit induced) opinion regarding students in their in and out groups. The nature of the measure also provided substantial data for evaluation.

The untimed semi-quantitative Social Judgement Questionnaire allowed research participants time to consider and adjust their answer for public consumption. As such, it operated as a measure of what the participants felt were acceptable opinions regarding students in their in and out groups. The resource allocation was so open that its lack of significant results speaks to the probability of multiple influencing factors other than our independent variables.

The Social Judgement Questionnaire data shows significant main effect for confession Valence (complimentary or derogatory comments) across all four word-sets (positive, negative, agreeable, reliable) and did not indicate either an effect based on Confession Author or an interaction effect. Due to the untimed nature of the test, results might be interpreted as a measure of what respondents felt was socially acceptable criticism, rather than as a measure of their actual opinion. With this consideration, we see that respondents felt more comfortable commenting based on a specific action (nice or mean comments) than on group membership.

The Implicit Association Test (IAT) data, however, was timed and one-directional. This added an urgency to answering and removed the opportunity to change answers, both factors aimed at getting an honest, rather than filtered or redacted answer. This Quantitative measure was highly sortable and using word choice, allowed for repeated measure evaluation with a high n which allowed good separation of effect. We saw that while our hypothesis that in-group favoritism as the primary force in opinions was supported by the data, and by much science before this study (Turner & Tajfel 1979; Saguy and Halperin, 2013) which showed Wellesley College getting overall fewer critical associations than MIT, there were some interesting nuances demonstrated by Valence and Authorship of comments.

We observed that Authors of Confession pages were held accountable for their nice or unfriendly comments (Valence) by study participants, with derogatory comments receiving more criticism than complimentary ones; additionally, we found that the direction of the interaction between Author and Valence was Rating-dependent. In other words, the Confession Authors from the Rated school had both the benefit and the blame for the tone of their Confessions while the Authors from the other school received a baseline in or out-group criticism independent of their commentary. For example: when rating Wellesley College, the Wellesley authors with complimentary comments got fewer negatives, and the Wellesley authors with derogatory comments got more, while the criticism of the MIT

authors was consistent across Valence. Similarly, in the MIT Rating, complimentary MIT authors got fewer criticisms, and derogatory MIT authors more, while Wellesley authors of both complimentary and critical had the same level of criticism.

We called this **the Blame Effect**, and it mirrors somewhat the findings of Dwyer's 2002 paper, where high prestige school comments were seen as coming from the institution, but low prestige school comments as coming from the individual. Our results similarly demonstrate discrimination of source as reflected in responses, when Authors from the school being Rated get both more credit and more blame for their commentary. It should be noted, though that this interaction does not supersede the overall in-group favoritism or out-group derision found in the broadest examination of the data.

The Dependent variable adjective divisions yielded interesting outcomes as well, showing the same in-group favoritism on the whole to Wellesley College, but adding an unexpected nuance of in-group harshness. To see this, we looked at the direct and indirect criticism scores. Direct criticism is defined as a Negative word associated with the college, while Indirect criticism occurs when there was a positive word that *could* have been given to the college but was not. We note that although Indirect criticism happened less than Direct overall, the proportion of Indirect to Total criticism gave a deeper view. Within each Rating, the ratio of Indirect/Direct Criticism for Conditions B, C & D were similar, but in condition A (Wellesley Authors say nice things about Wellesley), it changed in both IAT data sets. When rating Wellesley, the proportion of indirect criticism for condition A was highest, or most gentle; when rating MIT criticism in the same A condition gives Wellesley College the *least* indirect criticism of any case. Additionally, Wellesley IAT condition A and MIT IAT conditions B, C, &D all had very similar (high) levels of indirect criticism, or were more gentle, while Wellesley IAT B, C & D and MIT A all had similar (lower) ratios of Indirect criticism. This seems to say that the in-group, while receiving less criticism on the whole, is more exacting in the criticism it delivers- with in-group being held to a more direct standard. This could also be an artifact of stimulus fatigue prompted by imperfect IAT counterbalancing, although the specificity of difference in the A condition argues for an undefined effect.

The difference between Reliable and Agreeable adjectives are likewise interesting, and seem to reinforce the stringent ratings effect we saw when examining Direct and Indirect criticisms. We see that Wellesley is considered less agreeable, but more reliable when rating Wellesley Authors. The fact that Wellesley College's most reliable state

is negative self-criticism defies the Heyman, et. Al. 2013 paper findings which posit distrust of negative commentary, and which is upheld in commentary about MIT. These findings speak to trust of the in-group (increased reliability), even if it is objectionable (decreased agreeability), not observed for the out-group.

Overall, this study confirms the very strong in-group favoritism bias seen in other literature, but it implies that authorship imparts credit or blame that persists in the face of that favoritism. We also see a strictness in the criticism faced by in-group as opposed to outgroup and a trust of in-group judgment via increased Author Reliability scores.

The study, as conducted, had some important strengths and limitations. The strengths of large data pools, and repeated measure evaluation are pronounced benefits of the IAT structure that was used, while technical difficulties and the amount of time required to get the acquired data into a useable form were drawbacks. Because counterbalancing was not sufficiently employed, with MIT IAT usually being administered immediately after the Wellesley College IAT, results could be skewed- a weakness that should be corrected in any replication. And finally, I think the study would be strengthened by carful adjustment of descriptive adjectives used for each category to increase internal validity.

Our findings, such as understanding of the blame-effect, might be employed to equalize hiring or university admission practices, and inform blind reviews or the administrative variance of group membership after an initial filtering process.

The next step in this investigation needs to be replication, with a larger number of participants, and with attention to counterbalancing. Additionally, I think the millisecond data will be interesting to investigate for evaluation of self-censoring behavior within the IAT, and individual adjective counts or adjective optimization might help to polarize results for clearer effects. For external validity, the study should be repeated at several colleges, to control for possible gender or socio-economic confounds, and then in a broader community setting to control for age and education confounds to determine if it is generalizable.

References

Brown, R. (2000). Social Identity Theory: Past Achievements, Future Problems, Current Challenges. *European Journal of Social Psychology, 30(6)*, p745-778

Block, C.J., Koch, S.M., Liberman, B.E., Merriweather,T.J., & Roberson, L. (2011). Contending with stereotype threat at work: a model of long-term responses. *The Counseling Psychologist*, *39(4),* 570–600.

Durkheim, E. (1969). "Individualism and the Intellectuals." *Political Studies* 17: 14-30.

Durrheim K., Quayle M., Tredoux C.G., Titlestad K., Tooke L. (2016). Investigating the evolution of in-group favoritism using a minimal group interaction paradigm: the effects of inter- and intragroup interdependence. *PLoS ONE 11(11):* e0165974.

Everett J., Faber N., Crockett M. (2015). Preferences and beliefs in in-group favoritism. *Behavioral Neuroscience. Feb 13;9:15.*

Herak, I., Martinie, M., Almecija, Y., & Kamiejski, R. (2018). Social Judgement of an in-group member behaving in a (non)dissonant way. *International Review of Social Psychology. 31(1):19*, 1-10.

Heyman, D. G., Fu, G., & Lee, K. (2013) Selective Skepticism: American and Chinese Children's Reasoning About Evaluative Academic Feedback. Developmental Psychology, 49, 543-553

O'Dwyer, A., Berkowitz, N.H., & Alfeld-Johnson, D. (2002). Group and person attributions in response to criticism of the in-group. *British Journal of Social Psychology, 41,* 563–588.

Pennington, C.R., Heim, D., Levy, A.R., & Larkin, D.T. (2016). Twenty years of stereotype threat research: a review of psychological mediators. *PLOS ONE*: https://doi.org/10.1371/journal.pone.0146487

Saguy, T., & Halperin, E. (2014). Exposure to outgroup members criticizing their own group facilitates intergroup openness. *Personality and Social Psychology Bulletin 40(6)*, 791–802.

Sumner, W.G., (1883). **What Social Classes Owe Each Other**.

Table 2

Summary Statistics: Nonsignificant Effects

Measured	**Main effect for Author**	**Main effect Valence of Confessions**	**Interaction effect**
Money Game	F(1,56) = .53, p = .47	F(1,56) = .53, p = .47	F(1,56) = .15, p = .70
Positive adjectives (Questionnaire)	F(1,56) = .91, p = .35	Significant, see results	F(1,56) = .21, p = .65
Negative adjectives (Questionnaire)	F(1,56) = 1.10, p = .30	Significant, see results	F(1,56) = .68, p = .41
Agreeable adjectives (Questionnaire)	F(1,56) = .003, p = .95	Significant, see results	F(1,56) = .03, p = .86
Reliable adjectives (Questionnaire)	F(1,56) = .003, p = .96	Significant, see results	F(1,56) = 1.44, p = .24
WC IAT	F(1,599) = 2.26 p = .133	Significant, see results	Significant, see results
MIT IAT	F(1,90) = .19, p = .66	Significant, see results	Significant, see results