Kugler, Matthew B.; Goethals, George R.

Working Paper

Social comparison of abilities at an elite college: Feeling outclassed with 1350 SATs

WPEHE Discussion Paper, No. 70

Provided in Cooperation with:
Williams Project on the Economics of Higher Education, Williams College

Suggested Citation: Kugler, Matthew B.; Goethals, George R. (2006) : Social comparison of abilities at an elite college: Feeling outclassed with 1350 SATs, WPEHE Discussion Paper, No. 70, Williams College, Williams Project on the Economics of Higher Education (WPEHE), Williamstown, MA

This Version is available at:
https://hdl.handle.net/10419/58240

Discussion Paper No. 70

Williams Project on the Economics of Higher Education
23 Whitman Street
Mears West
Williams College
Williamstown, MA 01267

http://www.williams.edu/wpehe

Social Comparison of Abilities at an Elite College:

Feeling Outclassed with 1350 SATs

Matthew B. Kugler and George R. Goethals
Princeton University University of Richmond

The authors gratefully acknowledge support for this research provided by the Andrew Mellon Foundation through grants to the Williams Project on the Economics of Higher Education. The research was conducted while the second author was a faculty member in the Department of Psychology at Williams College, and the first author was a student at Williams. We thank Jared Carbone, Ann Greenwood, Chin Ho, Jennifer Huang, MiHye Kim, Melissa Murphy, Amanda Niu, and Gillian Weitz for their help in collecting and analyzing the data. And we thank Cassie King, Jerry Suls, Ladd Wheeler, Gordon C. Winston, and Joanne Wood for their helpful comments on the manuscript.

ABSTRACT

Two studies explored the experience and performance of students at Williams College in three-person groups that were homogeneous or heterogeneous in rated academic ability. In accord with hypotheses from Festinger's (1954) social comparison theory, students in academically homogeneous groups had more positive experiences and performed better on measures of written and video-taped performance. These results differ somewhat from recent studies of peer effects among roommates and from a line of recent social comparison research regarding the effect of exposure to superior others on one's own performance. In addition, students in single-sex groups had higher scores on several self-report and performance measures. Qualifying this finding were additional results showing that women did better in single-sex, while men did better in mixed-sex groups. The overall results were framed in terms of social comparison dynamics.

At all educational levels, students with varying academic ability are placed together in a classroom environment. In some cases, this mixing is minimized through tracking. In others, the integration is encouraged based on the beliefs that less able students can learn better from being educated with more able students and that any harm to the educational progress of the more able is small in comparison. There are even data suggesting that more able students might benefit from teaching less able students, as an older sibling would a younger (Zajonc, 1976). On the other hand, a recent consideration of the promise and reality of diversity - broadly defined as differences on "any attribute that another person may use to detect individual differences" (Williams & O'Reilly, 1998, p. 81, cited in Mannix & Neale, 2005, p. 31) - asks "what differences make a difference?" (Mannix & Neale, p. 31). When do differences interfere with effective functioning in groups? Social comparison theory, as originally formulated by Leon Festinger in 1954, quite clearly suggests that mixing individuals of heterogeneous ability levels will ultimately lead to disengagement, to the detriment of all parties. The present paper reports two studies testing the implications of social comparison theory for interactions among students of differing academic ability levels at a highly selective liberal arts college. While social comparison theory's predictions seem quite clear, there were equally clear reasons to doubt whether they would be supported in these studies. First, recent relevant research suggests that students with less ability may be inspired rather than turned off by their superior peers. Second, higher education's ethos of diversity and tolerance may be efficacious in overcoming any negative comparison effects. Third, even if disengagement is the fate of heterogeneous groups, we do not know the magnitude of the ability difference required.

We proceed as follows: First we describe our research paradigm and outline the key propositions of social comparison theory that seem directly relevant to its dynamics. Then we consider two literatures. One, largely from social psychology, concerns the impact of exposure to and interaction with superior others on typical students. Do superior others provide inspirational models, supporting high levels of aspiration, or does their superiority lead to discouragement, intimidation, cessation of comparison, and disengagement? A second literature, coming principally from economics, has examined the effects of varying ability levels of college roommates on the achievements of their peers. How are academic achievement and retention affected by different kinds of peers in one's living environment?

The research reported below explores these issues in a context of focused intellectual discussion, a portion of the educational domain that has yet to be studied and which we believe is central to the theoretical debate. Intellectual exchange inside and outside the classroom is central to the educational experience. Inevitably, all types of students at a college will have occasion to meet and interact with all other types during their formal classes and in informal discussions outside of class. It is in those places that the benefits posited by the advocates of diversity are expected to show themselves most fully and also where social comparisons of academic ability should be most prevalent. As such, it is crucial to understand the manner in which differences in academic ability and perceptions of those differences influence the quality of those exchanges.

We hope to achieve some insight into the dynamics of such intellectual interaction by studying the impact of ability differences, in some respects modest ones, among students in Williams College who participate in an unmoderated academic style

discussion of contemporary public issues. Groups of three underclass students read several *New York Times* pieces on contemporary social issues and discuss them for twenty minutes. We know each student's rated academic ability from data made available by the College admission office (described below). How will the various combinations of students react?

The great hope would be that different kinds of students would bring different perspectives to the issues and, as the number of dimensions on which a group was heterogeneous increases, so would the quality of the discussion. Higher education, particularly at elite liberal arts colleges, emphasizes the value of diversity and open-minded discourse; here, before all other places, we would be likely to find that differences aid the exchange of ideas.

We explore these issues by considering how the students in differently composed discussion groups score on three different kinds of measures. The groups are homogeneous (all three participants are in the top half of their class or all three are in the bottom half) or heterogeneous (some are top half, others are bottom half) with respect to assessed academic ability. We first examine the participants' self-reports of how much they benefited from the group exercise. These self-reports include measures of how much the participants enjoyed their experience and how much they feel they learned from it. Second, we assess both the quantity and quality of each participant's videotaped contribution to their group's discussion. Third, we evaluate the participants' written responses to a question asking what they actually learned from the discussion. In short there are self-report measures (e.g., how much did you learn from the discussion?) and

performance measures (how well did each student participate in the discussion and how much did each one show he or she learned).

As noted above, considerations from the early social comparison literature - where "pressure toward uniformity" is the driving force - lead us to expect difficulties when students of different levels of academic ability engage in a task in which that ability is relevant to success. Festinger originally formulated Social Comparison Theory (1954) to explain how a person satisfies the "drive" stipulated in his Hypothesis I "to evaluate his opinions and his attributes"[1]. While the motivations underlying the evaluation process originally described have been challenged - it now seems clear that the evaluation is far from disinterested (e.g. Suls & Wheeler, 2000a) – we believe that (and test whether) the fundamental propositions outlined fifty years ago are still sound.

Social comparison theory holds that, to the extent that objective means are not available to help us discover the accuracy of our opinions and measures of our abilities, we turn to comparison with our peers (Hypothesis II). The most informative object of comparison is a person or group that is relatively similar and, given the option, such is what people most often select (Hypothesis III, Corollary IIIa). Importantly, social comparison theory also states that a person will be less attracted to situations where others are very divergent from them than to situations where others more closely resemble them in terms of both abilities and opinions (Derivation C). In those situations in which a discrepancy exists between their opinions or abilities and those of the person or group to whom they are comparing, people are motivated to close the gap by changing either themselves, or their comparison persons (Derivation D1-D2). When the gap does not close, people tend to cease comparing themselves with those in the group who are

very different from themselves (Derivation D3). Hostility or derogation accompanies the cessation of comparison with others to the extent that continued comparison with those persons implies unpleasant consequences (Hypothesis VI). In sum, people are unhappy in group settings where others are dissimilar, and they will take steps to reduce the dissimilarity, including rejecting dissimilar others.

These portions of social comparison theory have a direct impact on the present question: Since people are less attracted to situations in which others are less similar, does it follow that those people in academically homogeneous groups enjoy the discussion more and are more likely to engage? An affirmative answer assumes, of course, that the magnitude of the academic differences is great enough to matter. Social comparison theory suggests that our chosen environment would be highly sensitive to such differences. Any factors that increase the strength of the drive to evaluate some particular ability or opinion will increase the "pressure toward uniformity" (Derivation E) and an increase in the importance of an ability or an opinion, or an increase in its relevance to the immediate behavior, will have such an effect on the drive to evaluate (Corollary to Derivation E). In our experiments, the task at hand is an academic one, as is the principle difference between the two types of participants (high and low ability). Thus differences are more salient and problematic.

Also emphasizing the difference between these types of student is the "unidirectional drive upward" that is present in the case of abilities (Hypothesis IV). Academic ability is one of the central features of college life, and most students are motivated to get better. Thus participants will be especially reluctant to be pulled down by their peers. While some of the low ability participants may attempt to pull themselves

up (a similar effect was seen in Dreyer, 1954), changing an ability is hard. At best, altering an ability requires time, energy, and effort, something that far more difficult than performing the same task on an opinion (Hypothesis V). These hypotheses suggest that high ability students will be unwilling and low ability students will be unable to reach a happy uniformity.

With uniformity being difficult to reach, participants in heterogeneous groups are faced with a problem. People prefer dealing with similar peers and, when they find that the peers they are comparing to are not similar, they tend to stop. Festinger's theory considered the possibility of hostility in the aftermath of cessation of comparison, saying that it would not generally occur in the case of abilities (Corollary VIA), but that it might if continued comparison with those persons implies unpleasant consequences (Hypothesis VI). Unfortunately, unpleasant consequences flow from comparisons in heterogeneous groups. For the lows, comparisons lead to the conclusion that they are not as smart as the highs. Recognition of superior ability is more likely the motivation for sabotage (Hoffman, Festinger, & Lawrence, 1954) than productive discussion. No one likes being outclassed. As mentioned above, however, the low's desire to cease comparison is in conflict with the desire to better their performance, meaning there is a small chance that they will violate expectations and thrive. The highs have a more predictable pattern. They have more ability than the lows and, presumably, are better equipped to participate in the discussion. During comparison, the highs realize that they cannot expect interaction at their preferred level. This causes frustration. While this may not lead to hostility per se, Schachter (1950) relates the cessation of comparison to the cessation of communication and interaction. If both types of participants find comparison unpleasant,

the general negative feeling should serve to inhibit the discussion and reduce the productivity of the whole experience for all parties.

While the theoretical implications are quite clear, there still remains the question of whether the academic differences among better and worse students at a private liberal arts college are noticeable in a relatively brief discussion and, if so, whether they matter. All college students are more homogeneous in academic ability than the general population and the differences between the top and bottom half of Williams students in our sample is especially small, only 150 points on combined math and verbal SAT scores. As we shall see, this difference did matter, but at the outset we were far from certain that it would. We believe that the small magnitude of this difference between conditions provides for a very conservative test of the theory, as greater diversity could be expected to magnify whatever differences we find.

As noted earlier, the recent literature on the effects of social comparison with superior others on motivation and educational achievement - surveyed by Wheeler and Suls, (2005) - raises questions about our interpretation of the classic theory. But the recent literature is itself somewhat ambiguous. Lockwood's and Kunda's research (Lockwood & Kunda, 1997; Lockwood, 2002) shows that exposure to a high achieving role model, a superstar, can increase self-esteem and motivation, but only if the superstar is not a direct peer. They considered the impact of exposure to a successful upperclassman's profile on both beginning and advanced students. The former were inspired because they believed they could achieve similar success, the latter knew otherwise (1997). Interestingly, it was much harder to persuade beginning students that the experience of a failed advanced student was predictive of their probable outcomes

(Lockwood, 2002). Related to these findings is research showing that people intentionally compare themselves with superior targets (Collins, 2000; Wheeler, 1966; Suls & Tesch, 1978) and that such comparisons produce more favorable self-assessments (Pelham & Wachsmuth, 1995). In Lockwood and colleagues' research, it appears that these motivations encourage participants to assimilate a model's positive outcomes into their own expectations while protecting them from being discouraged by negative exemplars.

Blanton, Buunk, Gibbons, and Kuyper (1999) conducted a longitudinal investigation of the effects of comparison on academic performance among 9th grade students in the Netherlands. In general, students most often reported comparing their exam grades with others who had slightly higher scores. More relevant to the current research was the finding that, controlling for prior grades, upward comparison predicted higher grades both cross-sectionally and longitudinally. Huguet, Dumas, Monteil, and Genestoux (2001) found similar results, but were unable to clarify the mechanism at work; moderating variables proved elusive.

These results suggest that good things follow from comparison with dissimilar superior others, or upward comparison. In contrast, a host of studies by Marsh and colleagues (Marsh, 1987; Marsh, 1991; Marsh, Hau, & Kraven, 2004; Marsh, & Parker 1984; Marsh, Kong, & Hau, 2000) have found evidence for what has been termed the Small Fish in a Big Pond Effect (SFBPE). In a representative high school study, having high quality peers negatively affected one's academic self concept, selection of advanced coursework, and educational and occupational aspirations while the student was in high school as well as college attendance and occupational aspirations two years after high

school graduation (Marsh, 1991). Further, these findings were moderated by the decline in academic self concept. Thus, we have evidence for an intimidation factor at work. Perhaps comparison with superior others does not reliably produce good outcomes.

We believe it is helpful to emphasize the differences in paradigms being employed. Lockwood and colleagues were conducting studies that were both experimental and related to the effects of *exposure to* superior and inferior others. Blanton et al and Huguet et al conduct studies that are not experimental in nature and are also looking primarily at exposure type effects, though this is less tightly controlled than in Lockwood's work. By looking at school quality as a variable, Marsh's work is distinct in emphasizing academic *interaction with* target others, though this line of research is also not experimental in nature. We present an interaction study that is experimental, and thus permits random assignment to high and low quality peers.

These psychological studies stand beside a series of recent papers by economists on the effects of mixing roommates of varying ability levels in the college environment. Studies of this sort were conducted at Dartmouth (Sacerdote, 2001), Williams (Zimmerman, 2003) and Berea Colleges (Stinebrickner & Stinebrickner, 2003). Stinebrickner and Stinebrickner (2003) found some evidence among female, but not male, students that their college grades are affected by their roommate's high school grades and their roommate's income level. They concluded that "low income students may be helped in a non-trivial fashion by being paired with higher income peers without the higher income peers incurring substantial costs" (p. 20). This finding suggests a "net gain to diversification" (p. 20). Sacerdote's (2001) study found that having a roommate in the top 25% on academic indices lifts one's own grades, and no gender differences

were reported. Zimmerman (2003) also found peer effects, with a 100 point increase in roommate's verbal SAT being associated with a small but statistically significant increase in one's own grades.

Both Stinebrickner and Stinebrickner and Zimmerman also show that roommate peer effects can be negative. The Stinebrickners find that retention can be negatively affected by low income peers. Zimmerman finds that low SAT peers can negatively affect the academic performance of students in the middle SAT range. One unique element of the Zimmerman study was that it also examined the effect of first year entries, living groups of 20-30 first year students. Students with low verbal SAT scores in entries where the average verbal SAT score was also low showed markedly poorer performance. Zimmerman found no gender differences. Sacerdote also has an interesting finding in this regard. While pairing a student from the top quarter with one from the middle half dramatically lowers the performance of the "high" student, it gave very little advantage to the "middle" student. In fact, Sacerdote reports a redistribution experiment considering rooms with a middle and top student in academic ability. He finds that both top and middle students would do better if they were paired with a roommate similar to themselves rather than different.

Taken together, the theoretical and empirical literature reviewed above provides grounds for worrying that when students are asked to interact with peers of different academic abilities, there will be a strong tendency to disengage on the part of the high ability participants. It is likely, but not quite so certain, that the lows will show a similar effect. This would create a pattern of results by which adding smarter students to a group of poorer students would lower the engagement and performance of the group. We

investigated these possibilities in two experiments. The first was small scale, considering gender and academic ability in the context of the three-student discussion paradigm noted above. Upon analyzing these data, we discovered effects that we believed warranted fuller exploration. We, therefore, ran Experiment 2, with far more groups and some additional dependent measures. In both studies we assessed engagement both by asking students how much they benefited from the discussion and also by measuring the quality of their videotaped participation in the group discussion and the quality of their written reports of what they learned.

Experiment 1

Method

Participants

One hundred thirty-eight Williams College first-year students and sophomores volunteered to participate in this study in the 1999-2000 academic year. They were paid $15.00 or received one-hour of extra-credit in an Introductory Psychology course. The study was called "College Students and Public Affairs."

Procedure

Participants were scheduled in groups of three, such that all three participants were in the same class (freshman or sophomore). Participants were greeted by an experimenter who explained briefly that the study entailed reading three articles from the *New York Times*, discussing those articles as a group, and answering questions about what they had read and discussed. The participants sat at a round table with a microphone in the center. The experimenter explained that they would be observed through a one-way

mirror and that their discussion would be videotaped using the microphone and a ceiling-mounted camera.

Participants were given twenty-minutes to read three articles, twenty-minutes for discussion, and twenty-minutes to answer a questionnaire asking about what they had learned from reading and discussing the articles. After giving instructions the experimenter left the room, and subsequently returned twice, first to ask the participants to begin the discussion and then to ask them to stop the discussion and complete the questionnaires. The room lighting was arranged so that when the experimenter was out of the room she was still partially visible in the adjoining room through the one-way mirror.

Academic Ratings

Participants were scheduled on the basis of an "academic rating" assigned by the Office of Admission when students apply. The experimenter was blind to those ratings. Academic ratings are based on students' secondary school grades, the quality of their secondary school academic program, their SAT's, and information in recommendations that seems to reveal academic potential. The academic ratings have been used for many years and are the best predictor of student grades at Williams. While the overall academic rating predicts student grades better than any of its components, the best single predictor among the elements is Verbal SAT.

Materials

Participants read three articles published in the *New York Times* in the summer of 1999. The first discussed the increasing amount of time that Americans spend at work vs. leisure (So Much Work, So Little Time, by Steven Greenhouse). The second discussed issues in genetic engineering raised by the then recent finding that Princeton scientists

had created a genetically smarter strain of mice (Ideas & Trends: Eek!; The Hidden Traps in Fooling Mother Nature, by Nicholas Wade). The third dealt with AIDS prevention (Focusing on Prevention in Fight Against AIDS, by Lawrence K. Altman).

The questionnaire participants completed at the end of discussion addressed four general areas: 1) how often they had previously engaged with the kind of topics they had just discussed; 2) how they assessed their performance in the discussion; 3) how much they gained from the discussion, and 4) how often they might take up such topics in the future. Specifically, the questionnaire asked students use seven-point scales to rate how often they had read such articles in the past, how much they had discussed them, how effectively they believed they had participated in the discussion, how well they thought they compared with other Williams students in their understanding of the topics after discussing the articles, how much they learned from reading and/or discussing each article, how much they learned from each of the other two students and, finally, how interested they would be in reading or discussing such articles in the future. In addition, participants were given three pages numbered one through ten on which they could write ten statements about the ideas or information they learned from reading and/or discussing each of the three articles. They could use the reverse side of the pages for writing additional comments.

Coding of written responses

Trained undergraduate raters individually coded each participant's written statements of ideas and information. A quantity rating gave credit for each idea or piece of information the participant wrote. A quality score rating from one to three was given to each statement on the basis of its specificity, detail, and elaboration. A total quality score

and an average quality score per statement were calculated for each participant. Inter-rater agreement on quantity scores was virtually 100%. For quality scores it was 74% and all disagreements were averaged.

Coding of discussion videotapes

Undergraduate raters coded each statement in the videotape of each discussion. First each rater proposed a written "order of talk" that listed who spoke when on the tape. There was near 100% agreement on who was speaking and any disagreements were resolved through discussion. Then each statement was given a length rating from one to four, depending on whether the statement was less than 5 seconds, from 6 to 10 seconds, 11 to 15 seconds, or greater than 15 seconds. Inter-rater agreement for quantity ratings was 95%. Disagreements were resolved through averaging.

Each statement was also given a quality rating of negative one to three, based on how effectively the statement advanced the discussion, and contributed to the intellectual quality of the discussion. Negative one scores were given to statements that halted or derailed discussion; zero was given to statements that were neutral or bland; one was given to remarks that advanced the discussion through simple statements or questions; two was given to remarks that were more thought provoking; and three was given to those rare statements that advanced the discussion productively and were exemplary in thought and expression. The two raters agreed on 73% of the quality ratings. Disagreements were resolved by averaging.

Results

There were 49 groups in the study, with 66 males and 81 females, comprising 6 all male groups, 11 all female groups, and 32 mixed groups. There were a total of 46

groups for which all participants could be classified based on academic rating. Eight groups were all top half students, 31 were mixed, and 7 were all bottom half. The unit of analysis was individual participants, such that there were 24 participants in all top half groups, 49 top half participants and 44 bottom half participants in mixed groups, and 21 participants in the all bottom half groups. Participants in the high academic rating category had a median SAT score of 1490, the participants with low ratings had a median score of 1350. We consider both self-report measures from the questionnaire responses and performance measures based on coding what the participants wrote that they learned from reading and/or discussing the articles and coding the discussion videotapes.

Differences Based on Academic Rating

Academic Rating was examined in a 2 x 2 ANOVA with Rating (low versus high) and Homogeneity (all the same versus mixed). Participants with high academic ratings said that they had gained more from the experience by learning more from the articles and discussion ($M_{high} = 4.71$, $M_{low} = 4.21$, $F = 7.03$, $p < .01$) and from each of their peers ($M_{high} = 8.65$, $M_{low} = 7.88$, $F = 4.18$, $p < .05$) than participants with low ratings. Highly rated participants also said their future interest was high, predicting that they were more likely to discuss ($M_{high} = 4.83$, $M_{low} = 4.36$, $F = 4.56$, $p < .05$) and, in a non-significant trend, read ($M_{high} = 4.68$, $M_{low} = 4.40$, $F = 3.07$, $p < .10$) more about such topics in the future.

More interestingly, there were strong homogeneity main effects on the written performance measures. Participants in academically homogeneous groups wrote statements that were both higher in total quality ($M_{hom} = 21.53$, $M_{het} = 17.25$, $F = 8.60$, $p < .01$) and more numerous ($M_{hom} = 15.24$, $M_{het} = 12.75$, $F = 8.46$, $p < .01$) than

participants in heterogeneous groups. There were no main effects for academic rating or interactions on these measures. Participants in homogeneous groups said that they compared better with their peers in their understanding of the topics after the discussion ($M_{hom} = 4.71$, $M_{het} = 4.30$, $F = 9.53$, $p = .002$). This last number was an average across the three articles. Neither the main effect for academic rating nor the homogeneity by academic rating interaction was significant. There were no significant effects on the video performance measures.

Differences Based on Gender

The same sort of two-way ANOVA was used for the sex related variables, with participant Gender (male or female) and Homogeneity (all one sex or mixed). Males consistently scored higher on measures of participant confidence, stating that they knew more about the issues before the discussion ($M_{male} = 4.20$, $M_{female} = 3.78$, $F = 5.40$, $p < .05$), compared better with respect to their peers ($M_{male} = 4.62$, $M_{female} = 4.28$, $F = 4.08$, $p < .05$), and, in a non-significant trend, participated more effectively in the discussion ($M_{male} = 4.84$, $M_{female} = 4.48$, $F = 3.28$, $p < .10$). This confidence does not stand entirely in isolation: on the total quality measure derived from the video discussion, men do better ($M_{male} = 28.34$, $M_{female} = 24.11$, $F = 6.29$, $p < .05$).

Single sex groups reported more past and future interest, saying that they had discussed such issues more often in the past ($M_{hom} = 4.59$, $M_{het} = 4.09$, $F = 6.18$, $p < .05$) and that they had greater interest in both reading ($M_{hom} = 4.78$, $M_{het} = 4.43$, $F = 4.50$, $p < .05$) and discussing ($M_{hom} = 4.84$, $M_{het} = 4.51$, $F = 3.46$, $p < .10$) similar articles in the future.

Interactions

ANOVAs also considered interactions involving gender, gender homogeneity, academic rating, and academic homogeneity. A two-way gender-by-academic-homogeneity ANOVA revealed that men suffer a greater loss of productivity in heterogeneous groups than do women on the total written quality (M_{men} = 23.36 to 15.80, M_{women} = 19.78 to 18.42, F = 4.56, P < .05) and quantity (M_{men} = 16.22 to 11.86, M_{women} = 14.40 to 13.45, F = 4.28, P < .05) scores. There are no significant interactions between academic rating and gender homogeneity.

Discussion

These results support the sobering implications of Social Comparison Theory: heterogeneous groups do relatively poorly. Our sample size, however, was small. We hoped with a larger sample to be able to explore the role of ethnic homogeneity/ heterogeneity on group performance. Also, it would have been good to know how well participants knew each other, because the homogeneity effects might be a function of preexisting friendships. A final note of concern involved the content of the articles used as a basis for discussion. Another experiment using a similar paradigm (Goethals, 2000) had found large gender effects that could be attributed to the content of the pieces used for discussion (two of the three were sports related). While we did not intend and could not determine any worrisome pattern in the pieces used for experiment 1, we were concerned that the male confidence effects might be domain specific. For these reasons, we ran experiment 2. The results for both of these experiments are discussed below.

Experiment 2

To get a larger sample size in a college with approximately 2000 undergraduates, more than 500 first and second year students were participants during a three year period,

from the fall of 2001 to the spring of 2004. This number of participants allowed us to overcome one of the main limitations of the previous study, namely the small number of academically homogeneous groups. The design and procedure were very similar to those of the previous experiment. The only significant changes were the rewording of several questionnaire items for greater clarity, the addition of questions on how well each participant knew their peers, a new selection of articles for discussion, and a more holistic scoring of the video data (see below).

Method

Participants

Five hundred and sixty four Williams College in 188 groups, the vast majority of whom were first- and second-year students, volunteered to participate in this study. They were paid $15.00 or received one-hour of extra-credit in an Introductory Psychology course. The study was once again called "College Students and Public Affairs."

Procedure

The procedure was the same as that for Experiment 1.

Materials

Participants read three pieces published in the *New York Times* in August 2001. The first was a Bob Herbert op-ed that discussed a recent hate music concert in Georgia and its free speech implications (High Decibel Hate). The second was an article concerning the US News and World Report rankings of undergraduate colleges and universities and criticisms of the current formulas ('Best' List for Colleges by U.S. News is Under Fire, by Alex Kuczynski). The third dealt with the over-prescription of drugs

such as Ritalin and the role of advertising (School's Backing of Behavior Drugs Comes Under Fire, by Kate Zernike and Melody Petersen).

As in Experiment 1 a self-report questionnaire asked students to rate (this time on ten-point scales) how often they had read or discussed similar articles in the past, how often they were likely to do so in the future, how much they knew about the topics going into the discussion, how much they learned from reading and discussing the articles, how much they had gained from the discussion, how much they enjoyed it, how well they compared to other Williams students in their understanding of the topics, how interested they would be in reading or discussing similar articles in the future, how well they contributed to the discussion, and how well they knew each of the other two students. That last variable was summed, combining the ratings for each peer. As in Experiment 1, the questionnaire also asked them to write the ideas or information they learned from reading and/or discussing each article. Each page provided space for participants to write ten written statements next to the numbers one through ten and provided room for additional comments on the reverse side as well

Coding of written responses

Trained undergraduate raters individually coded each participant's statements of ideas and information. A quantity rating gave credit for each idea or piece of information the participant stated, there were no disagreements on quantity ratings. A quality score rating from one to three was given to each statement on the basis of its centrality, detail and elaboration. Inter-rater agreement was 89% on quality ratings and disagreements were resolved by averaging.

Coding of discussion videotapes

We were able to transcribe the videotapes for the second two years of the study (N= 330). Undergraduate raters coded each participant on how well they spoke on each article as well as overall. For both measures, the scale was negative one to three, based on how effectively the person advanced the discussion and contributed important ideas. Participants who were actively detrimental to discussion received scores of negative one; participants who didn't participate in discussions or didn't offer any content in their discussions received scores of zero; participants who advanced the discussion through simple statements or questions received scores of one; participants who made thought-provoking contributions to discussions received scores of two; and the few participants who advanced the discussion productively and were exemplary in thought and expression received scores of three. If the group failed to discuss an article, these individual ratings were not penalized. The rating for that article would be null, not zero. The person's overall individual score was intended as a holistic measure, not an average. Among other things, it took into account the relative time committed to each article.

The group discussion was also rated as a whole, an important change from the first experiment. This rating did not reflect the intellectual quality of the discussion as much as the efficiency and effectiveness of the group. How much time did the group spend discussing the articles? Did they stay on task? We hoped to separate the groups that were able and willing to have a 20-minute academic discussion from those that were not. This was a 3-point scale. Agreement on individual and group video performance was nearly 100%, and disagreements were averaged.

Results

Analyses were once again done in the format of 2 x 2 ANOVAs between the relevant category and whether the participant's group was homogeneous with respect to that category. There were 25 groups in which all participants were academic highs, 27 in which all were academic lows, and 122 that were heterogeneous with respect to participant academic ability. 239 males and 324 females participated, forming 19 all male groups, 47 all female groups, and 108 groups that were heterogeneous with respect to participant gender.

Though several items were added to the questionnaire or coded differently, as described above, the same broad categories of measures were used in this experiment as were used in Experiment 1. In this experiment, however, it was possible to create several composite variables for the self report measures, making it easier to categorize them. The self reports for having read and discussed similar articles in the past were combined (α = .77) into a Past Engagement composite and the two corresponding variables related to future interest were combined (α = .85) into a Future Interest composite. Also averaged were the three measures of how much the participants thought they got out of the experience (α = .85): how much the participants believed they learned from the articles and/or discussion, how much they enjoyed the discussion, and how much they felt they gained from the discussion. These measures made up a Perceived Benefit composite. The final self report category, level of confidence in one's ability in the task, did not yield any composite variable. In fact, significant differences were seen on only one of the three "confidence" measures: the one that asked how well participants believed they compared to their peers.

Differences based on Academic Rating

Participants were classified as having high or low academic ability using a median split within each year's participant group. Cutoffs were similar for all three years. In nonsignificant trends, participants with high ratings scored higher on the Past Engagement composite ($M_{high} = 4.74$, $M_{low} = 4.57$, $F = 2.80$, $p < .10$) and also the Future Interest composite ($M_{high} = 6.72$, $M_{low} = 6.45$, $F = 3.46$, $p < .10$). These trends are similar to the finding in experiment 1 that also showed such participants reporting that they would read and discuss more frequently in the future.

Academic homogeneity was associated with several strong effects. Participants in homogeneous groups had higher scores on the Perceived Benefit composite ($M_{hom} = 6.22$, $M_{het} = 5.71$, $F = 10.50$, $p < .001$).[1] They also said that they knew their peers better prior to the study ($M_{hom} = 6.95$, $M_{het} = 5.95$, $F = 5.73$, $p < .05$). Controlling for that variable, the Perceived Benefit composite was still significant at the $p = .002$ level. Note that, as in experiment 1, there was no interaction with academic rating on this variable; a group of all academic lows was as successful as a group of all academic highs and both were less successful than heterogeneous groups.

Following these trends, we looked for similar homogeneity effects on the written and video measures. While the written measures did not reveal similar patterns, the group video ratings did. In a nearly significant trend, homogeneous groups performed better on the measure of group organization and focus than did heterogeneous groups ($M_{hom} = 2.48$, $M_{het} = 2.22$, $F = 3.32$, $p = .076$). The unit of analysis for this test was the group, not the individual. This drastically reduced the N, especially as video ratings are

[1] Because of their importance, the elements of this composite were also tested individually. They were all significant.

only available for the second two years. There were 31 homogeneous and 72 heterogeneous groups.

Unlike the homogeneity effects in experiment 1, those found here are not qualified be a gender-based interaction. In this study, both females and males are equally affected. Participants in the high academic rating category (across years) have a median SAT score of 1500 and the participants with low ratings have a median score of 1320, similar to what was seen in experiment 1.

Differences based on Gender

There were several gender-based effects, forming a clearer pattern than in experiment 1. Women had higher scores than men on the Future Interest composite ($M_{male} = 6.35$, $M_{female} = 6.71$, $F = 7.78$, $p < .01$). They also had higher scores on the Perceived Benefit composite $M_{male} = 5.64$, $M_{female} = 6.00$, $F = 13.20$, $p < .001$). On the written performance measures, they outscored men on both total ($M_{male} = 19.37$, $M_{female} = 21.11$, $F = 6.34$, $p < .05$) and average ($M_{male} = 1.58$, $M_{female} = 1.67$, $F = 9.98$, $p = .002$) statement quality. These data show that women believe they got more out of the study and, to interpret their superior written performance, were either more engaged and/or more conscientious.

Male participants thought they compared more favorably to their peers ($M_{male} = 6.89$, $M_{female} = 6.40$, $F = 8.06$, $p < .01$). This replicates experiment 1, and likely reflects feminine modesty or male immodesty. Male participants do not generally outperform females on the video quality ratings in this experiment, giving their immodesty less empirical support.

In experiment 1, we saw participants in single sex groups reporting that they were more likely to read and discuss in the future. We interpreted that as a signal, albeit a weak one, that the discussion was more successful in their cases. This was seen more directly in this study. Both men and women contributed more to the discussion on the individual video measure ($M_{hom} = 1.50$, $M_{het} = 1.31$, $F = 3.46$, $p = .06$) in single sex groups.

The other measures, however, told a more complicated story. While men had higher scores on the Future Interest composite in gender heterogeneous groups ($M_{hom} = 6.17$, $M_{het} = 6.40$, $F = 2.91$, $p < .10$), women showed the opposite trend ($M_{hom} = 6.88$, $M_{het} = 6.57$). Once again, there is a clear signal from the Perceived Benefit composite as to how much the participants believed they gained. Women in gender homogeneous groups scored higher on the composite than women in gender heterogeneous groups ($M_{hom} = 6.18$, $M_{het} = 5.85$, $F = 10.30$, $p < .001$) while men once again preferred heterogeneity ($M_{hom} = 5.13$, $M_{het} = 5.79$). Interestingly, women say they know their peers better when in homogeneous groups ($M_{hom} = 7.04$, $M_{het} = 5.87$, $F = 3.85$, $p < .05$) while men are better acquainted in heterogeneous groups ($M_{hom} = 5.86$, $M_{het} = 6.32$). As was the case with academic rating, the homogeneity effects found here persisted even at the same levels even after controlling for preexisting friendships.

One of the written performance measures also reflects the homogeneity interactions. In a non-significant trend ($F = 3.13$, $p < .10$), women write comments with higher average quality ($M_{hom} = 1.69$, $M_{het} = 1.65$,) when in homogeneous groups while men did better in heterogeneous groups ($M_{hom} = 1.54$, $M_{het} = 1.60$). In other words, both

men and women are vastly more engaged, conscientious, and happy when in groups with women, but men still converse better when just around their own kind.

Differences based on Ethnicity

Based on admission office categories, we coded participants as being white or non-white (Black, Hispanic, American Indian, Foreign, and Asian). Thirty-five percent of the students were classified as non-white on this basis. Numbers were too small to examine differences within the non-white group. There were no consistent or meaningful differences between ethnically homogeneous vs. ethnically heterogeneous groups.

Discussion

The results of the two studies reported above show in a number of ways that groups of academically homogeneous students perform better than heterogeneous groups. The studies were conducted at Williams College, a highly selective, elite liberal arts college, where all the students are highly academically competent. One might expect that heterogeneity of academic ability would not matter, or would not even be noticed, in such a short time in such an academically rarified environment, but the data show that it was noticed and that it did matter. This is shown on both self-report measures and on written and video-based performance measures. As Marsh et al (2000) noted, the school environment naturally makes ability comparisons salient.

In Experiment 1, participants in homogeneous groups wrote more and excelled on a measure of the total quality of their writing. They also felt that they had a superior understanding of the issues being discussed compared to those in heterogeneous groups. In Experiment 2, participants in homogeneous groups reported learning more from their peers, having more fun discussing the material, and gaining more from their experience

in the study. Analyses of the video measures showed that they also did a better job as a group in getting themselves organized to discuss the material and staying on task. In both experiments, the benefits of homogeneity and the costs of heterogeneity were shared across academic ability categories. Both high and low participants preferred to be with similar others.

Both studies also reported effects based on the participants' personal academic rating. In Experiment 1 those with high academic ratings felt that they learned more from participating in the study and learned more from their peers, and they indicated that they were more likely to discuss such issues in the future. However, there were no performance differences between those with high vs. low academic ratings. In Experiment 2 those with high academic ratings claimed to have read more such articles in the past and intentions to read more in the future. Again, there were no performance differences between those with high vs. low academic ratings.

These results show that while engaging the articles used in these studies appealed more to those with high academic ratings, group homogeneity also produced more engagement and enjoyment, and better performance on some measures. That is, despite the fact that the task appealed more to those with high academic ratings, satisfaction was more affected by group homogeneity. Perhaps the potential to perform well was greater for groups with more collective academic ability, but the actual performance depended not on how much ability a group possessed, but whether that ability was distributed equally or unequally. Success with the task of these experiments seems to have depended on engagement rather than ability, and engagement was superior in homogeneous groups.

These findings are entirely in accord with Festinger's statement of social comparison processes, and they suggest challenges for educators hoping to benefit students who have less academic ability by having them work together with those who have more. In theoretical terms, social comparison theory, and decades of research, make clear people's preference for comparing with similar others (Suls & Miller, 1977; Suls & Wheeler, 2000b; Suls & Wills, 1991). Past research has not as fully investigated people's preference for being in homogeneous groups, but our findings that people are more attracted to such groups and engage more effectively in them are entirely consistent with Derivations C and D3 in the original theory, which state that people are less attracted to heterogeneous groups and that they are more likely to cease comparing in them and, therefore, less likely to communicate in them. Hypothesis VI states that cessation of comparison can be accompanied by hard feelings, if continued comparison is unpleasant. We think that continued comparison with dissimilar others is unpleasant in our paradigm and that cessation of comparison and negative affect occurs in our academically heterogeneous groups.

We also explored the effects of group homogeneity/heterogeneity on the dimension of gender, specifically comparing single-sex vs. mixed-sex groups. We might expect gender homogeneity to have less effect on academic engagement than academic ability, but there are some indications that it matters. In Experiment 1 single-sex groups reported that they had discussed issues of the kind used in our studies more than mixed-sex groups. Since this could be true only by chance, it may simply reflect the fact that single-sex groups engaged more with the material than mixed-sex groups. Consistent with this possibility are the companion findings that gender-homogeneous groups

reported greater interest in reading and discussing such articles in the future. In Experiment 2, participants in single-sex groups received higher scores on the individual video measure, suggesting once again more engagement with the material in single-sex groups.

But in addition to homogeneity effects, there are other interesting effects involving gender. In Experiment 2 statistical interaction effects showed that both men and women have higher self-report and performance scores when their peers are female. That is, women do better in homogeneous groups, where there are no men, and men do better in heterogeneous groups, which include women. These effects are seen on the same composite that was sensitive to academic homogeneity, the one that averages measures of how much the participants learned from their peers, how much they gained from the discussion, and how much fun they had. The interaction also appears in the average quality of their written comments. These effects may reflect greater conscientiousness in women, as recently documented by Rubenstein's (2005) study of Israeli students' scores on Costa and McCrae's (1992) Big Five questionnaire. Both men and women likely benefit in our studies from others' conscientiousness, and women are more likely to bring that quality to their groups. The Rubenstein study also showed that women are more agreeable. That personal quality may help. Whatever the explanation, the data show that men pull down women's scores, and women pull up men's. Extrapolation suggests that single-sex education might be better for women and worse for men. Similar speculations have arisen in relation to technological education, with the suggestion that a single sex teaching environment would limit the impact of the related sex based stereotypes (Cooper & Kugler, in press).

The results of our two studies as well as those of other recent studies of peer effects underline the challenge for schools of higher education that want to capitalize on the potential educational benefits of diversity. The good news is that we do not find any consistent effects related to ethnic homogeneity vs. heterogeneity. Ethnicity doesn't seem to affect the group dynamic. Academic ability and gender, however, do. Individuals in groups which are homogeneous for academic ability, even when that means that everyone is at the same low level, do better. Groups that are mixed on this dimension do more poorly, even when they contain several high quality participants. Educators should be aware of the dynamics that produce these outcomes, and consider ways of overcoming them. These results reflect dynamics that are triggered by very small comparison differences and take hold in a very short period of time. This suggests that there will be no easy answer for handling divergences in ability. The dynamics regarding gender are even more complex, but they suggest that both personality and social roles can positively or negatively affect intellectual engagement in groups. These norms present both opportunities and challenges for today's educators.

While the data patterns in these studies are reasonably clear, their generalizability is less so. Students in our studies read together for twenty minutes, talked for twenty minutes, and wrote together for twenty minutes. This is a relatively short time period. We do not know whether the same kinds of findings would emerge from lengthier interactions. Unfortunately, it actually seems likely that as more information becomes available to a person about their peers, the strength of comparison effects and any tendency to disengage would increase. The contact hypothesis, which holds that stereotypes breakdown in the face of sustained interaction, does not immediately apply

here. For repeated interactions to have a positive effect, the parties involved need to have equal status (Allport, 1954; Gaertner, Dovidio, & Bachman, 1996). In this study, the inequality of status is precisely the problem on the most relevant dimension. Still, we must be cautious in drawing conclusions about what happens in academically heterogeneous groups over longer periods of time.

Another limitation of the study is that, the differences among Williams students in academic ability are small compared to the whole range of academic ability in U.S. education. Though the differences are not dramatically smaller than those of other Tier 1 and 2 colleges and universities, other institutions of higher education and the domain most often studied in SFBPE research – high schools – may have much larger ranges. This difference has interesting implications. In terms of social comparison theory, our compacted range results in understated effects. As was said above, one of the surprises of experiment 1 was that participants even noticed the distinction between peers with high and low academic ratings. More heterogeneous institutions would likely face even worse prospects in mixing students of varying ability levels than we have reported here.

Success on the task we used was intended to be a function of interest and engagement. It could be argued that a more ability intensive task would seriously undermine the performance of the homogeneous low ability group. While we recognize that this result could occur, Social Comparison Theory would lead us to predict that the heterogeneous groups would still perform worse. SCT would posit that making the outcome of the task so directly related to the comparison dimension – ability - would cause the pressure toward uniformity to increase correspondingly. This would magnify the disengagement in heterogeneous groups still more.

In terms of policy implications, the fact that there are differences between our short experimental paradigm and longer classroom interactions, and also between Williams students and a more general sample of American undergraduates, suggest caution in drawing conclusions. We feel that the effects we found would actually be magnified under conditions of greater heterogeneity or greater task difficulty, but these predictions have yet to be tested. Regardless, it is very sobering that pressures toward uniformity and other social comparison dynamics are strong enough to make themselves felt within such a short period of time in such an academically stratified environment. Such pressures constitute formidable challenges to institutions trying to identify the best ways to facilitate students learning from each other.

REFERENCES

Allport, G.W. (1954). *The nature of prejudice*. Oxford, England: Addison-Wesley.

Altman, L. K. (1999). Focusing on Prevention In Fight Against AIDS. *The New York Times*, August 31, 1999, p. F5.

Baron, R. S. (2005). So right it's wrong: Groupthink and the ubiquitous nature of polarized group decision making. In M. P. Zanna (Ed.), *Advances in experimental social psychology*, 37 (pp. 219-253). San Diego, CA: Elsevier Academic Press.

Blanton, H., Buunk, B. P., Gibbons, F. X., & Kuyper, H. (1999). When better-than-others compare upward: Choice of comparison and comparative evaluation as independent predictors of academic performance. *Journal of Personality and Social Psychology*, 76, 420-430.

Collins, R. (2000). Among the better ones: Upward assimilation in social comparison. In J. Suls & L. Wheeler (Eds.), *Handbook of social comparison* (pp. 159-172). New York, NY: Kluwer/Plenum.

Cooper, J. & Kugler, M. B. (in press). The digital divide: The role of gender in human computer interaction. In J. A. Jacko and A. Sears (Eds.) *The human-computer interaction handbook.*

Dreyer, A. (1954). Aspiration behavior as influenced by expectation and group comparison. *Human Relations*, 7, 175-190.

Festinger, L. (1954). A theory of social comparison processes. *Human Relations*, 7, 117-140.

Gaertner, S. L., Dovidio, J. F., & Bachman, B. A (1996). Revisiting the contact hypothesis: The induction of a common ingroup identity. *International Journal of Intercultural Relations*, 20(3-4) pp. 271-290.

Goethals, G. R. (2000). Social comparison and peer effects at an elite college. *Williams Project on the Economics of Higher Education*. Discussion Paper # 55.

Greenhouse, S. (1999). ideas & trends: running on empty; so much work, so little time. *The New York Times*, Sept 5th 1999, (4) pg. 1.

Herbert, B. (1999). In America; high-decibel hate. *The New York Times*, August 20, 2001. pg. A17.

Hoffman, P. J., Festinger, L, & Lawrence, D. H. (1954). Tendencies toward group comparability in competitive bargaining. *Human Relations*, 7, 141-159.

Hogg, M. A. & van Knippenberg, D. (2003). Social identity and leadership processes in groups. In M. P. Zanna, *Advances in Experimental Social Psychology*, Vol 35 (2-43). New York: Academic Press.

Huguet, P., Dumas, F., Monteil, J.M. & Genestoux, N. (2001) Social comparison choices in the classroom: further evidence for students' upward comparison tendency and its beneficial impact on performance. European Journal of Social Psychology, 31, 557-578.

Kuczynski, A. (1999). 'Best' List For Colleges By U.S. News Is Under Fire. *The New York Times*, August 20, 2001, pg C1.

Lockwood, P. (2002) Could it happen to you? Predicting the impact of downward comparisons on the self. *Journal of Personality and Social Psychology*, 82, 3, 343-358.

Lockwood, P. & Kunda, Z. (1997) Superstars and me: Predicting the impact of role models on the self. *Journal of Personality and Social Psychology*, 73, 91-103.

Mannix, E. and Neale, M.A. (2005) What differences make a difference? The promise and reality of diverse teams in organizations. *Psychological science in the public interest, 6,* 31-55.

Marsh, H. (1987). The Big-Fish-Little-Pond effect on academic self-concept. *Journal of Educational Psychology*, 79, 280-295.

Marsh, H. (1991). The failure of high-ability high schools to deliver academic benefits: The importance of academic self-concept and educational aspirations. *American Educational Research Journal*, 28, 445-480.

Marsh, H., Kong, C-K, & Hau, K-T (2000). Longitudinal multilevel models of the Big-Fish-Little-Pond effect on academic self-concept: Counterbalancing contrast and reflected glory-effects in Hong Kong schools. *Journal of Personality and Social Psychology*, 78, 337-349

Marsh, H., & Parker, J. (1984). Determinants of student self-concept: Is it better to be a relatively large fish in a small pond even if you don't learn to swim as well? *Journal of Personality and Social Psychology*, 47, 213-231.

Pelham, B., & Wachsmuth, J.(1995). The waxing and waning of the social self: Assimilation and contrast in social comparison. *Journal of Personality and Social Psychology*, 69, 825-838.

Sacerdote, B. (2001). Peer Effects with Random Assignment: Results for Dartmouth Roommates. *Quarterly Journal of Economics*, 116 (2), 681.

Schachter, S. (1950). Deviation, rejection and communication. In L. Festinger & K. Back (eds), *Theory and Experiment in Social Communication* (pp. 51-82). Michigan: Research Center for Dynamics Institute for Social Research.

Sherif, M. (1936). *The psychology of social norms*. Oxford, England: Harper.

Suls, J. M. & Miller, R. L. (1977). *Social comparison processes: theoretical and empirical perspectives*. Oxford, England: Hemisphere.

Suls, J., & Tesch, F. (1978). Students' preferences for information about their test performance: A social comparison study. Journal of Applied Social Psychology, 8, 189-197.

Suls, J. & Wills, T. A. (1991). *Social comparison: Contemporary theory and research*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Suls, J. M. & Wheeler, L. (2000a). A selective history of classic and neo-social comparison theory. In J. M. Suls & L. Wheeler, *Handbook of social comparison: Theory and research*. Netherlands: Kluwer Academic Publishers.

Suls, J. M. & Wheeler, L. (2000b). *Handbook of social comparison: Theory and research*. Netherlands: Kluwer Academic Publishers.

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist*, 52 (6), 613-629.

Stinebrickner, T. R. & Stinebrickner, R. (2003). What can be learned about peer effects using college roommates? Evidence from new survey data and students from disadvantaged backgrounds. Unpublished Manuscript. Dept. of Economics, The University of Western Ontario. London, Ontario Canada. http://www.ssc.uwo.ca/economics/faculty/Stinebrickner/peers.pdf

Tajfel, H. & Turner, J. C. (2004). The social identity theory of intergroup behavior. In J. T. Jost & J. Sidanius (Eds*), Political psychology: Key readings* (pp. 276-293). New York, NY: Psychology Press.

Wade, N. (1999). Ideas & trends: Eek!; The hidden traps in fooling mother nature. *The New York Times*, September 5, 1999, (4), p 1.

Wheeler, L. (1966). Motivation as a determinant of upward comparison. Journal of Experimental Social Psychology, Suppl. 1, 27-31.

Wheeler, L. & Suls, J. (2005). Social comparison and self-evaluations of competence. In A. J. Elliot & C. S. Dweck, *Handbook of competence and motivation*. New York: Guilford Publications, pp. 566-578.

Williams, K. & O'Reilly, C. (1998). The complexity of diversity: a review of forty years of research. In B. Staw & R. Sutton (Eds.) *Research in organizational behavior* (Vol. 21, pp. 77-140) Greenwich, CT: JAI Press.

Zajonc, R. B. (1976). Family configuration and intelligence: Variations in scholastic aptitude scores parallel trends in family size and the spacing of children. *Science*, 192 (4236), 227-236.

Zernike, K. & Petersen, M. (1999). Schools' backing of behavior drugs comes under fire. *The New York Times*, August 19, 2001, pg. A1.

Zimmerman, D. (2003). Peer effects in academic outcomes: Evidence from a natural experiment. *The Review of Economics and Statistics*, 85 (1), 9-23.

[1] We thought it would be useful for the historically minded to spell out the various formal propositions in Festinger's original theory that are discussed in this paper.

Hypothesis 1. There exists, in the human organism, a drive to evaluate his opinions and his abilities.

Hypothesis 2. To the extent that objective, non-social means are not available, people evaluate their opinions and abilities by comparison respectively with the opinions and abilities of others.

Hypothesis 3. The tendency to compare oneself with some other specific person decreases as the difference between his opinion or ability and one's own increases.

Corollary III A. Given a range of possible persons for comparison, someone close to one's own ability or opinion will be chosen for comparison.

Derivation C. A person will be less attracted to situations where others are very divergent from him than to situations where others are close to him for both abilities and opinions.

Hypothesis IV. There is a unidirectional drive upward in the case of abilities which is largely absent in opinions.

Hypothesis V. There are non-social restraints which make it difficult or even impossible to change one's ability. These non-social restraints are largely absent for opinions.

Derivation D1. When a discrepancy exists with respect to opinions or abilities there will be tendencies to change one's own position so as to move closer to others in the group.

Derivation D2. When a discrepancy exists with respect to opinions or abilities there will be tendencies to change others in the group to bring them closer to oneself.

Derivation D3. When a discrepancy exists with respect to opinions or abilities there will be a tendency to cease comparing with those in the group who are very different from oneself.

Hypothesis VI. The cessation of comparison with others is accompanied by hostility or derogation to the extent that continued comparison with those persons implies unpleasant consequences.

Corollary VI A. Cessation of comparison with others will be accompanied by hostility or derogation in the case of opinions. In the case of abilities this will generally not be true.

Derivation E. Any factors which increase the strength of the drive to evaluate some particular ability or opinion will increase the "pressure toward uniformity" concerning that ability or opinion.

Corollary to Derivation E. An increase in the importance of an ability or an opinion, or an increase in its relevance to immediate behavior, will increase the pressure toward reducing discrepancies concerning that opinion or ability.

Hypothesis VIII. If persons who are very divergent from one's own opinion or ability are perceived as different from oneself on attributes consistent with the divergence, the tendency to narrow the range of comparability becomes stronger.