Decision Making Consequences of the Paradoxical Flip

by

Houston F. Lester

A thesis submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Master of Science

Auburn, Alabama
August 9, 2010

Keywords: confidence, social comparison, paradoxical flip,
decision making

Approved by

Daniel J. Svyantek, Co-chair, Professor of Psychology
Ana Franco-Watkins, Co-chair, Assistant Professor of Psychology
Adrian L. Thomas, Associate Professor of Psychology

Abstract

The hard-easy manipulation (i.e., manipulation of item difficulty) has been used to demonstrate that participants are sometimes overconfident while believing they are worse than average. This finding is often referred to as a paradoxical flip. Prior research has examined how this reversal occurs and whether it is a real psychological phenomenon. The purpose of this study is to investigate the paradoxical flip's decision making consequences in terms of losing bets and to determine if the paradoxical flip can be abated by providing participants with additional information concerning their cohort's performance. Results indicated that people exhibiting decision behaviors consistent with the paradoxical flip do lose more bets. However, additional cohort performance information did not reduce the amount of the paradoxical flip.

Table of Contents

List of Tables

List of Figures

Introduction

Confidence is the subjective probability estimate of one's abilities or of the quality of decisions made (Keren, 1991). In a given domain, an individual has a confidence level. Calibration can be conceptualized as the fit between the individual's confidence level and his or her ability in the domain (Keren, 1991). Prior research investigating confidence and decision making has revealed that often participants are not well calibrated (i.e., their confidence levels do not correspond to performance; Larrick, Burson, & Soll, 2007). The ability to discriminate between high and low performance domains for one's self is important for decision making in many areas (e.g., job performance and economic decisions; Bolger, Pulford, & Coleman, 2008). Overconfidence and underconfidence (over- and underconfidence) are denoted by perceived confidence levels that are higher and lower than actual performance, respectively. Furthermore, depending upon the difficulty level of the tasks, participants tend to be either over- or underconfident. Typically, poor performance and overconfidence occur for hard tasks and good performance and underconfidence occur for easy tasks (Ronis & Yates, 1987).

Perceived percentile[1] (i.e., the percentage of people that a participant believes he or she performs equal to or better than on a given task), is another measure of subjective confidence. Because perceived percentile and confidence ratings are both measures of

[1] Although percentile rank is the correct term for the construct, percentile will be used because it is the term that has been used in prior research.

subjective confidence, it is logical to believe that a positive relationship should exist between them (Larrick et al., 2007). Researchers have found that these constructs are often positively related; however, when extremely difficult and easy items are used this relationship is called into question.

The hard-easy manipulation (i.e., presenting participants with items that are extremely easy and difficult) influences mean perceived percentile ratings and over- and underconfidence differently. Kruger (1999) found that when asked to rank their ability on an easy task (e.g., operating a mouse) the participants' average percentile was above fifty demonstrating the better than average effect (BTA). However, when asked to estimate their ability on a difficult task (e.g., juggling ability) participants rated themselves as worse than average (WTA; i.e., a mean perceived percentile below fifty).[2] These results are inconsistent with the findings in the confidence literature (i.e., overconfidence on hard and underconfidence on easy tasks). These results may be attributed to the theoretical difference between the social comparison (i.e., perceived percentile) and confidence constructs. These constructs have different foci. Confidence ratings require the participant to estimate his or her performance, while perceived percentile ratings require the participant to estimate his or her performance relative to the performance of the reference group. The differential hard-easy effects on confidence and perceived percentile are known as the paradoxical or counterintuitive flip (Moore & Kim, 2003).[3]

[2] Although the mean of a percentile rank distribution is not always the 50th percentile, prior research used it as the mean or median of the distribution (Kruger & Dunning, 1999).

[3] Confidence and perceived percentile can also be referred to as absolute and relative subjective performance ratings.

The goals of this study are to: (1) investigate whether the paradoxical flip is real, specifically, a result of psychological processes and not an artifact of the methodology used to assess confidence and perceived percentile, and (2) determine its decision making implications in terms of bet outcomes. If the flip leads to poor decisions (e.g., losing bets), then a third goal of the study will be to eliminate poor decision making by providing participants with their cohort's performance norms (i.e., typical performance). This study used items from different domains and difficulty levels to reach these goals. The following sections provide the conceptual foundation needed to address the study's three goals. Specifically, I explain the constructs that compose the flip (i.e., calibration and over- and under placement), discuss the methodology used to assess confidence and perceived percentile ratings, review the psychological and statistical explanations of the hard-easy effect on confidence, perceived percentile, and the paradoxical flip, and illustrate why the flip is important.

Confidence Paradigm Methods

To ensure that the paradoxical flip is due to psychological processes and not methodological differences, the methods used to assess perceived percentile and confidence must be as similar as possible. Historically, most confidence paradigms use alternate choice formats where participants are instructed to choose the correct answer choice and provide a confidence rating. Confidence ratings are provided for each item and are either in the half (i.e., 50-100%) or full range format (i.e., 0-100%). The accuracy of these confidence ratings are typically assessed by the calibration formula (i.e., $\bar{r} - \bar{c}$; average confidence rating – percent correct) which is then used to determine if the participant is either over or underconfident (Ronis & Yates, 1987). Perfect calibration

occurs when a participant's average confidence equals percent correct, while positive and negative calibration scores correspond to over- and underconfidence, respectively.

Response format controversy. When attempting to determine if participants are demonstrating a bias (e.g., over- and underconfidence or BTA/WTA), the researcher needs to be certain that the response format is not influencing the way the participant responds. As previously stated, confidence ratings are typically assessed using alternate choice questions with either the half or full range confidence format (Fischhoff, Slovic, & Lichtenstein, 1977; Larrick, et al., 2007). Some researchers believe that the half range is the most appropriate because it does not make sense to make a dyadic decision and provide a confidence rating below 50%. Does a confidence rating of 30 indicate a 70% confidence rating for the option not chosen? To avoid confusion, some researchers recommend using the half-range confidence rating format (Keren, 1991). However, the half range confidence rating format has its limitations as well. One limitation is that it is possible to answer less than half of the items correctly, which is the lower limit or floor of the half range response scale. For example, in the half range format, participants who score less than 50% of the items correctly will be overconfident simply because confidence rating options below 50% are not available (Juslin, Anders, & Henrik, 2000). Despite its limitations, the current study used the half range format. Items were pilot tested to eliminate sets of questions where participants answered less than half of the alternate choice items correctly or responded to all of the items correctly. The pilot study reduces the likelihood that scale-end will affect the current study.

Statistical explanation for the hard-easy effect on calibration. Before providing explanations of the paradoxical flip, the hard-easy effect on the components of

the paradoxical flip must be understood. The hard-easy effect on calibration is a component of the paradoxical reversal. Several explanations for this effect on calibration are based on statistical reasoning (i.e., defining item difficulty post hoc, linear dependency, and scale-end effects; Juslin et al., 2000). In addition to the response format, multiple methods (i.e., a priori and post hoc) are used to determine which items are considered hard and easy. Post hoc item definition occurs when the hard and easy items are defined by the proportion of participants who answered the items correctly. This practice is inappropriate and leads to statistical confounds that may result in the hard-easy effect.

Post hoc item definition combined with the linear dependency between proportion correct and over- and underconfidence is the underlying mechanism of Juslin et al.'s (2000) explanation of the hard-easy effect on calibration. Proportion correct and over- and underconfidence are linearly dependent because the amount of over- and underconfidence is contingent upon the proportion of items correct. Juslin et al. demonstrated that underconfidence on easy items and overconfidence on hard items can occur simply due to random sampling error when the hard and easy items are defined post hoc. Two sources of sampling error that may affect the participant's accuracy are item selection and the probability of guessing correctly (Juslin et al.). For example, participants were perfectly calibrated with confidence ratings and accuracy equal to 75%. However, when participant's sampling error was introduced it affected the proportion of items correct, resulting in accuracy scores ranging from 70 to 80%. This is a stringent test of their hypothesis because the 5% sampling error estimate is very modest, and sampling error in the confidence ratings was omitted. The probability of guessing correctly (i.e.,

50%) in a dyadic choice is much larger than 5%. If 70% of the participants answered the item correctly then it was considered hard. If 80% of the participants responded correctly, then it was considered easy. Because participants were perfectly calibrated at the population level (i.e., their confidence ratings were always 75%), post hoc item selection and accuracy's sampling error were the only causes of over- and underconfidence in the hard and easy items, respectively. This demonstration provides strong evidence that post hoc item selection is inappropriate. Conducting a pilot study allowed the difficulty of the items to be defined before they were implemented in this study.

Additional statistical explanations proposed by Juslin et al. (2000) are that regression toward the mean and scale-end effects could create the hard-easy effects on calibration. Scale-end effects occur when a participant gets 100 or less than 50% of the items correct. When all of the items are answered correctly, the participant can only be underconfident. At the other extreme, when a participant is correct less than 50% of the time he or she is always overconfident because the lowest confidence rating available is 50%. Regression toward the mean, which can contribute to the counterintuitive flip, is the most difficult statistical artifact to abate. The correlation between proportion correct and confidence is not a perfect one; thus, the plot of the proportion correct and confidence ratings will tend to have a regression effect. Furthermore, the larger the distance between the observed value and the mean, the larger the regression effect (Campbell & Kenny, 1999). The hard-easy manipulation allows regression effects to occur. Easy tasks result in scores that are much larger than the overall mean; hard tasks result in scores that are much smaller than the overall mean. Therefore, when differential hard-easy effects are

the impetus that calls a relationship into question, regression towards the mean must be analyzed as a potential explanation for the differential effects.

Improved methodology. Methodology developed by Larrick et al. (2007) differentiated itself from the traditional calibration research by abandoning the dyadic choice format and determining the difficulty of the items beforehand. A split sample technique was also developed as a means of negating linear dependence. Using two samples allows perceived percentile and confidence estimates to be compared to different accuracy scores within the same domain. The difficulty of the items was decided before the participants completed the measure by manipulating the stringency of the criteria. For example, Larrick et al. asked the participants to estimate the year in which a person won the Nobel Prize in literature. In the easy condition, the answers were considered correct if they were within 30 years of the correct year, while the criterion in the difficult condition required participants to guess within five years of the actual year. The departure from the dyadic choice format allowed the researchers to reduce the scale-end effects and ensured that the full range format is interpretable (i.e., 0 to 100% confident). The new methodology can be used without any of the previously mentioned confusion (i.e., scale-end effects, interpretation of confidence ratings less than 50, and linear dependence). Using the improved methodology, Larrick et al. explained the hard-easy effect on calibration as follows: Hard and easy tasks affect participants' confidence and accuracy. However, the hard-easy manipulation affects accuracy more than it affects confidence ratings (i.e., participants are conservative when adjusting their confidence ratings). Based on the calibration formula, this will result in overconfidence on difficult items and underconfidence on easy items. Although these methodological improvements provide

for an explanation of the hard-easy effect on calibration that is not statistical in nature, two potential drawbacks of this response format are that the participants may not provide answers for questions about which they are unsure, and variance in the answers provided could make it difficult to set the accuracy criteria.

Perceived Percentile Methods

After describing the methods and explanations of the hard-easy effect on calibration, the same explanations need to be made for perceived percentile. Describing the hard-easy effects on calibration and perceived percentile will provide the reader with a complete understanding of how the paradoxical flip is formed. The perceived percentile paradigm requires participants to complete a task and estimate the number of people out of 100 they believed they performed equal to or better than (i.e., perceived percentile). Unlike the confidence paradigm, when providing perceived percentile ratings, participants are only required to provide a rating upon task completion. An additional difference is that BTA/WTA is a group-level construct, while over- and underconfidence is often assessed at the individual level. To ensure that the different levels of assessment do not confound results, social comparison must be assessed at the individual level as well (Moore & Small, 2007). Overplacement (OP) and underplacement (UP) represents the individual-level construct of social comparison, which is calculated by taking each individual's perceived percentile and subtracting his or her actual percentile (Larrick, et al., 2007). OP/UP occur when an individual's perceived percentile is above or below his or her actual percentile, respectively. Accurate placement occurs when the perceived and actual percentile are equal.

Several advantages of OP/UP over BTA/WTA can be seen in the following example. If a participant's perceived percentile estimate is the 60th percentile, it will contribute to the group's BTA; however, this rating does not necessarily indicate that the participant has an inflated view of his or her relative performance. Perhaps the participant's actual ranking was in the 75th percentile. A perceived percentile score of 60 with an actual percentile score of 75 would result in an OP/UP score of -15, which denotes underestimation of relative performance.

OP/UP also allows the researcher to have a better idea about the presence or absence of bias in the data. Analyses based on group means, such as BTA/WTA, can be affected by outliers (i.e., extreme values will have a larger effect on the mean). For example, data sets in which participants display an extremely large amount of OP/UP with outliers at both extremes may not display the BTA/WTA effect. When averaged together, extremely high and low percentile ratings can result in a BTA/WTA score near 50, which will obscure any individual-level bias. In this example, OP/UP provides an additional tool for the researcher to examine whether there is bias in the data.

Psychological Explanations for the Hard-Easy Effect on Social Comparison

A key component of the paradoxical flip is the hard-easy effect on accuracy; however, it is not the only construct affected by the hard-easy manipulation (Larrick et al., 2007). Two explanations of the hard-easy effect on an individual's perceived percentile are differential information (i.e., similar to reference group neglect; Moore & Small, 2007) and that participants rate their performance as typical or modal instead of comparing themselves to the mean (Roy & Liersch, 2008).

Reference group neglect. The large differences in perceived percentile caused by the difficulty of the task, typically plus or minus ten percentiles, appear to indicate that participants engage in reference group neglect (i.e., do not consider how the difficulty of the task alters the performance of others). Participants who consider how the difficulty manipulation affects their cohort's performance would not be likely to make a large change in their perceived percentile because of the difficult task.

However, the differential information explanation (Moore & Small, 2007; Moore & Cain, 2007) stated that participants appear to engage in reference group or cohort neglect because the participants do not have information about the performance of others in the cohort. Moore and Small argued that participants are not engaging in reference group neglect; there is simply little reference group performance information to use. If the participant feels that he or she performed at either extreme (e.g., in the 90 or 10 percent correct), assuming that the participant has no information regarding how others performed, then it is reasonable to believe that the other members of the reference group performed closer to the mean. This logic would result in ratings below and above the 50^{th} percentile on hard and easy tasks, respectively.

Roy and Liersch (2008) conducted a study that provided additional support for the differential information explanation. The methodology used by Roy and Liersch provided students with statistical information about the distributions of scores (i.e., where the majority of the scores lie). Typically, the distributions of scores on easy tasks are skewed to the left and the distribution of scores on difficult tasks is typically skewed to the right. Roy and Liersch claimed that when presented with depictions of positive, negative, and normal distributions that participants rate themselves as modal and not BTA/WTA.

Rating yourself as modal in difficult or easy tasks will place your perceived percentile below and above 50, respectively. The current study builds on Roy and Liersch's methodology and the differential information theory by providing normative performance information to the experimental group. Providing participants with cohort performance norms may reduce the hard-easy effect on the participants' perceived percentile ratings. As a result, participants receiving additional cohort performance information will rate themselves as closer to the percentile mean in the hard and easy conditions and display less OP/UP.

Differential Effects of the Hard-Easy Manipulation

Conceptually, participant's OP/UP and over- and underconfidence scores should be highly related because both confidence and perceived percentile ratings are estimates of subjective confidence. Moreover, BTA/WTA and over- and underconfidence have the same first step (i.e., participants provide a subjective rating of their absolute performance level). However, the hard-easy manipulation revealed that confidence and perceived percentile are not always positively related (Moore & Healy, 2008). Participants are expected to answer substantially more easy items correctly than hard. In difficult tasks, participants rate themselves as WTA while simultaneously being overconfident; meanwhile, the opposite is true in easy tasks. Participants rate themselves as BTA while being underconfident in easy tasks. For example, Larrick et al. (2007) found that when asked to estimate football scores (i.e., a difficult task), participants rated themselves as WTA while being overconfident, and when naming capitals of each state (i.e., an easy task) participants rated themselves as BTA and underconfident. Despite the differential mean level effects that were revealed by the hard-easy manipulation, social comparison

and overconfidence are positively related across all difficulty levels and domains. As such, Larrick et al. found that perceived percentile significantly predicted overconfidence.

Psychological Explanations of the Paradoxical Flip

Most of the theories in the previous sections explain the paradoxical reversal as it relates to either perceived percentile or confidence ratings. However, several researchers provide explanations for the paradoxical flip (Moore & Healy, 2008; Larrick et al., 2007). Larrick et al. demonstrated that the differential hard-easy effects can occur between the two constructs when they are positively related. Larrick et al.'s theory is based upon the idea that participants are conservative when rating their confidence. Their explanation is as follows: relative to hard items, easy items result in increased perceived percentile, confidence, and accuracy; however, the change in mean accuracy is larger than the changes in mean confidence ratings. Thus, participants are underconfident on easy items. Large changes (i.e., decreases or increases) in perceived percentile ratings often result in OP/UP because changes in task difficulty will affect most participants' scores (i.e., decrease on a hard task or increase on an easy task). Thus, perceived percentile should remain relatively the same regardless of task difficulty.

Although Moore and Healy (2008) used a different methodology to assess absolute and relative subjective estimates, their theory also provides an explanation for the paradoxical flip. Moore and Healy's model, which is based on Bayesian belief updating, states: that after experiencing a task, people often have imperfect information regarding their own performance and even less information about the performance of others. Consequently, post-task estimates about their own performance are regressive and

post-task estimates of others' performance are even more regressive. On easy tasks people underestimate their own performance (i.e., underconfidence) and underestimate the performance of others even more which results in OP. This explains why someone would be underconfident and display OP simultaneously. The explanation works in a similar fashion for hard tasks. Participants overestimate their performance on hard tasks (i.e., overconfidence) and overestimate their cohort's performance even more, resulting in UP. This explains why someone would be overconfident and display UP simultaneously.

Importance of Subjective Confidence and the Paradoxical Flip

Explanations of the hard-easy effect on calibration, OP/UP, and how these two effects combine to form the paradoxical flip are theoretically interesting. However, it may be unclear why an organization would care about studying the paradoxical flip. In other words, why is the reduction of the hard-easy manipulation's effect on subjective confidence relevant to an organization? Subjective confidence ratings, whether they are absolute (i.e., confidence) or relative (i.e., perceived percentile), affect behavior (Gino & Moore, 2007). Despite conventional wisdom that high levels of confidence always result in good outcomes (e.g., if you believe you can achieve), overconfidence often leads to poor decision making (Bazerman & Neale, 1982). The existence of these consequences is affirmed by numerous theories in management and economics that assume the BTA effect (e.g., Benabou & Tirole, 2002; Dunning, Heath, & Suls, 2004) and overconfidence can influence individuals' decision making negatively (e.g., Bolger et al., 2008). The negative impacts of these biases are demonstrated by individuals buying plummeting stocks, because they believe their superior knowledge will allow them to know when stock prices will rise again or excessive entry in a market entry game (Bolger et al.,

2008). Excessive market entry is dangerous because payoffs diminish as the number of people in the market increases.

Prior research suggests that the hard-easy manipulation alters confidence ratings, which in turn affect the weighting of cohort advice (Gino & Moore, 2007). In Gino and Moore's study, participants were asked to provide an estimate of how much someone weighed and then allowed to see another participant's estimate of the same person's weight. The effect of the other participant's advice was defined by the amount that the participant changed the original estimate after receiving advice. Larger effects were denoted by larger changes in the original estimates. When performing an easy task, participants were overconfident and did not adjust their original estimate. As expected, hard tasks resulted in opposite findings (i.e., less confidence and overweighing advice). In organizations, overconfident employees might be less likely to take advice from their cohort even when it is applicable.

It is noteworthy that the Gino and Moore (2007) study did not address the weighting of an expert or authority figure's advice. Prior research suggests that information from experts or authority figures (e.g., managers) is often discounted less than information from peers (Sniezek, Schrah, & Dalal, 2004). Although employees receive advice from a manager frequently, the increasing trend of using work teams in organizations improves the chance of an employee receiving cohort advice (Landy & Conty, 2007; Murphy & Shiarella, 1997). Employees need the ability to take advice from fellow team members to arrive at a better decision. Providing participants with cohort performance information would allow them to have appropriate levels of confidence and potentially weigh advice appropriately.

By demonstrating that over- and underconfidence is relatively stable across time and even resistant to feedback, Jonsson and Allwood (2003) provide additional support for the organizational relevance of subjective confidence. Moreover, Moore and Cain (2007) found that 12 rounds of feedback including information concerning the participant's performance as well as the performance of the group was ineffective in improving decision making and influencing participants' confidence ratings. The resiliency of the bias (i.e., over- and underconfidence) provides additional support for why this bias should be investigated. Current research has been unable to reduce the amount of error in subjective confidence ratings consistently. Establishing ways to reduce these biases (i.e., over- and underconfidence and OP/UP) could result in better decision making by enabling employees to weigh advice appropriately.

Over- and Underconfidence and OP/UP as Predictors of Decisions

As previously illustrated, prior research has revealed that BTA/WTA and over- and underconfidence have negative decision making consequences. However, research investigating the decision making consequences of the counterintuitive flip is less definitive. Predicting an individual's behavior who displayed the flip is difficult; will the person behave in a manner that is indicative of his or her overconfidence or the percentile rating below the mean? Which of the constructs (i.e., BTA/WTA or over- and underconfidence) will be a better predictor of future decisions? To address this question, this study allowed participants to provide confidence and perceived percentile ratings and place a bet based on absolute or relative performance.

Prior research investigating the criterion validity of absolute and relative confidence estimates has revealed conflicting results. Festinger (1954), who coined to

term social comparison, proposed the original model. His theoretical model asserted that objective or absolute information would surmount social comparison information when both are available. Moore and Klein (2008) conducted an experiment that supported Festinger's hypothesis. This study will be described in detail because similar methods, particularly the bets, were used in this experiment. Moore and Klein provided participants with feedback (i.e., absolute and relative) about a practice test. The absolute feedback included information concerning the number items the participant answered correctly (i.e., two or eight). The relative feedback provided the participants with information concerning their percentile (i.e., 23rd percentile or 77th percentile). However, the feedback given to the participants was no indication of the participant's performance (i.e., participants were randomly assigned to low and high conditions of absolute and comparative feedback). The random assignment of feedback is important because it provides for the independent comparison of the decision making effects of relative and absolute feedback. If the feedback is indicative of actual performance, then the absolute and comparative feedback would likely be positively related; therefore, parsing out their effects would be more difficult. For example, participants who performed well would get feedback that they answered a large percentage of the items correctly and that their percentile was high.

An additional advantage of the randomly selected feedback is that it allows the researcher to provide some participants with incongruent relative and absolute feedback (e.g., high percentile and poor absolute performance). This manipulation will allow the effects of absolute and relative feedback to be pitted against one another. After Moore and Klein (2008) gave the participants feedback, participants completed the actual test

and were asked to place bets based on their absolute and relative performance. In their study, to win the bet based on absolute performance, the participant must respond correctly to 50% of the items. To win the bet based on relative performance the participant had to perform equal to or better than the 50th percentile. The bets that were used in this study are similar to those provided by Moore and Klein.

Moore and Klein's (2008) results supported Festinger's (1954) theory that absolute performance information will outweigh the relative cohort performance information (i.e., more people made their bet based on absolute performance). Absolute feedback was also found to have larger effect sizes for betting behavior, confidence in winning the bet, affective outcomes, and state self-esteem. Contrarily, Klein (1997) found that subjective evaluations of performance were only influenced by comparative feedback. The present study provided participants with cohort performance norm information and allowed participants to choose between bets based on absolute and relative performance. This design allowed the study to answer the convoluted question of which construct leads to behavior, and will provide a representation of an individual's true confidence level.

Overview of Current Study

The goals of this study are to (1) replicate the paradoxical flip; (2) investigate the flip's decision making consequences on bet outcomes; and (3) determine if providing participants with cohort performance norms before task completion can modify the flip. These goals were assessed by methodology that is a combination of techniques used by Larrick et al. (2007; i.e., defining the difficulty of the items a priori) and Moore and Klein (2008; i.e., asking participants to make bets based on absolute or relative performance).

Larrick et al.'s methodology negates linear dependence and scale-end effects, which allowed the current study to avoid the statistical/methodological confounds discussed by Juslin et al. (2000). This methodology allows the current study to focus on the lack of information about the cohort's performance as the possible cause of the paradoxical flip. Cohort performance norms were manipulated between subjects and are an extension of methods performed by Moore and Cain (2007). Moore and Cain manipulated cohort performance information to exhibit that perceived percentile estimates demonstrate differential regression instead of an egocentric bias. When given cohort performance information, participants' perceived percentile ratings were no longer affected by the difficulty of the task. The current study employed the same manipulation to determine how it affects the paradoxical flip (i.e., differential hard-easy effects on confidence and perceived percentile ratings). The manipulation of cohort performance information is hypothesized to affect the paradoxical flip, because it has already been shown to affect one component of the flip (i.e., perceived percentile). Additionally, the cohort performance information contains the number of items answered correctly by the participants who piloted the items. This additional information should improve the participants' calibration.

Hypothesis 1: The cohort performance information will allow participants in the experimental condition to display (a) better calibration, (b) less OP/UP, and (c) less of the paradoxical flip.

Hypothesis 2: Participants who display higher levels of the paradoxical flip will make fewer correct decisions in terms of winning bets.

Method

Participants

One hundred and nineteen Auburn University undergraduates enrolled in a statistics class were recruited to participate in the experiment. The participants received extra credit towards course fulfillment for their participation.

Design

Cohort performance norm information was manipulated between subjects. The performance (i.e., number of items correct) for the 25th, 50th, and 75th percentiles of a group of Auburn students who piloted the items was provided to participants in the experimental condition. All participants completed the hard, medium, and easy items in each of three trivia domains (i.e., entertainment, science/nature, and sports). Thus, this study was a 2 (training: cohort information and control) x 3 (difficulty: hard, medium, and easy) x 3 (domain: entertainment, science, and sports) mixed design. The dependent variables of interest are: accuracy, confidence, calibration (i.e., over- and underconfidence), perceived percentile, OP/UP, BTA/WTA, paradoxical flip, bets chosen, and bet outcomes.

Many of the dependent variables are also used as independent variables in the current study. OP/UP, calibration, paradoxical flip, and condition were used as independent variables. Different scoring methods were used to analyze the constructs associated with absolute and relative confidence estimates. The operational definition of each of the variables is provided.

Accuracy. Accuracy is defined as the percent of items answered correctly in each domain/difficulty level combination.

Confidence. Confidence is defined as the mean confidence rating for items in each domain/difficulty level combination.

Calibration. Calibration is defined as the fit between the participant's confidence and accuracy (i.e., $\bar{r}$ - $\bar{c}$; average confidence rating – percent correct). Over- and underconfidence was calculated by subtracting the percent of items answered correctly from the participant's mean confidence rating.

Perceived Percentile. Participants were asked to estimate the number of people out of 100 they believe they performed better than or equal to in each domain/difficulty level combination. This rating is their perceived percentile.

OP/UP. OP/UP is the individual level social comparison construct defined as the participant's perceived percentile minus his or her actual percentile. OP occurred when the participants received a positive score, and UP occurred when the participants received a negative score. To facilitate the interpretation of regression coefficients, the absolute value of the OP/UP will also be used.

BTA/WTA. OP/UP was the central variable in the relative or social comparison analyses performed in this study, because it is also used to form the BTA/WTA variable. BTA/WTA is the group level construct that is defined as the overall mean of the OP/UP variable. Positive values indicate that the group believes they are BTA, and negative values indicate that the group believes they are WTA.

Paradoxical Flip. The paradoxical flip is a combination of the calibration and OP/UP variables. The amount of the paradoxical flip displayed was calculated by

taking the absolute value of each participant's calibration score minus his or her OP/UP score (i.e., |C-(OP/UP)|). OP and underconfidence or UP and overconfidence will result in larger paradoxical flip scores. Higher scores represent more paradoxical flip.

Bets Chosen. Participants were asked to choose between an absolute and a relative bet.

Bet Outcomes. Participants' bets were evaluated to determine if each was a "win" or a "loss".

Materials

Statistical training. The participants had recently been exposed to the material on perceived percentiles and frequency distributions in their statistics class. Thus, using this sample reduces the probability of differential statistical understanding between the experimental and control groups. Additionally, all participants completed statistical training to ensure that they understood the statistical information that was necessary to complete the task. Manipulation checks were performed to ensure that the statistical training was effective. Appendix A contains the statistical training as well as the questions that the participants were required to answer correctly to complete the training.

General knowledge items. General knowledge items with hard, medium, and easy questions from each domain (i.e., sports, history, and science/nature) were included. The items within each domain were randomly selected from Trivial Pursuit 25th edition and pilot tested to determine difficulty level. All items are two alternative forced choice questions. Participants completed 10 items within each

domain/difficulty level combination resulting in a total of 90 items. Example items are presented in Appendix B.

Perceptions of Absolute and Relative Performance. Perceptions of absolute performance were measured by participants' ratings of the confidence for each question. Participants were asked to provide a confidence rating that ranged from 50 to 100% confident. One hundred percent confidence represents absolute confidence and 50% confidence represents a guess. Perceptions of relative performance were measured by participant's estimates of their perceived percentile. Participants were asked to estimate the number of people out of 100 they performed better than or equal to in each domain/difficulty level combination.

Bets. Participants were asked to place a bet based on either relative or absolute performance. To win the absolute bet participants must respond correctly to at least 70% of the items. The relative bets required the participants to perform better than or equal to 70% of their cohort. Participants were instructed to choose the bet they believe they would win. See Appendix C for complete information about the bets. Bets were compared to actual performance to determine if the bet was a "win" or a "loss".

Procedure

Participants were randomly assigned to one of two groups. The experimental group received cohort performance norm information and the control group did not. All participants were taught the statistical concepts that were necessary to complete the task. Additionally, all participants were taught that the majority of scores typically lie around the mean.

The additional cohort performance information can only be effective in reducing the flip if the participants understand the statistical information as well as transfer the knowledge to different domains. Prior research has revealed that statistical training involving the law of large numbers can be transferred across different domains (Fong & Nisbett, 1991). More importantly, statistical training influences the way participants interpret "real life" events involving uncertainty (e.g., subjective confidence estimates; Fong, Krantz, & Nisbett, 1986). Kosonen and Winne (1995) provided additional evidence for the transfer of statistical training by demonstrating that university, high school, and middle school students could transfer the same statistical training as provided in Fong and Nesbitt (i.e., the law of large numbers) to different domains.

The order of the domains as well as the order of the hard, medium, and easy items was randomized. All participants completed the nine randomly presented blocks of 10 questions within each sub-domain. Difficulty (i.e., hard, medium, and easy) and domain (i.e., sports, history, and science/nature) were fully crossed; thus, resulting in 90 items from nine sub-domains. Participants provided an answer to each question and rated their confidence from 50-100%. To eliminate the linear dependency of OP/UP and over- and underconfidence on accuracy, the split sample technique recommended by Larrick et al. (2007) was performed. Using a split half approach, the 10 items within each sub-domain were divided into two groups of five. For example, each participant completed a total of 10 items from the hard domain of sports questions. Accuracy, average confidence, and over- and underconfidence were computed for the first five questions. After the completion of the second five questions, participants provided their perceived percentile

rating for these questions from 0 to 100. Average perceived percentile and OP/UP were computed for the last five questions.

Results

Manipulation Check

A manipulation check was performed to determine whether participants responded accurately more often on the easy items than on the medium or hard items replicating the hard-easy effect. The manipulation check is necessary because the paradoxical flip only occurs when the hard-easy manipulation is effective. A 3 (item difficulty: hard, medium, and easy) x 3 (domain: entertainment, science, sports) repeated measures analysis of variance (ANOVA) was used to test this hypothesis. Mauchly's test indicated that the sphericity assumption had been violated $\chi^2(2) = 17.75$, $p<.01$; thus, the degrees of freedom were corrected using the Greenhouse-Geisser correction ($\varepsilon = .87$). Item difficulty significantly affected the number of correct responses, F (1.74, 191.25) = 1596.74, $p < .01$, $\eta^2_p = .94$. As can be noted from Figure 1, participants' accuracy decreased as the question difficulty level increased. The planned comparisons, displayed in Table 1, revealed significant differences between each of the difficulty levels. The main effect of domain was not statistically significant. However, the difficulty by domain interaction was significant, $F(3.35, 368.37) = 39.66$, $\eta^2_p = .27$. As can be noted from Figure 1, the interaction appears to be driven by accuracy in the science and sports domains. In particular, the mean accuracy for the hard items was highest in the science domain; however, accuracy was lowest in the science domain for the medium difficulty level questions. The mean accuracy scores were the lowest for the hard and easy items in

the sports domain; however, the accuracy scores for the medium difficulty level items were the highest in the sports domain.

Paradoxical Flip Replication

Participants demonstrated the flip across all difficulty levels and domains. Table 2 contains the means and 95% confidence intervals of the paradoxical flip in each difficulty level. It appears that the data replicated the paradoxical flip; however, further inspection reveals that the results only partially replicate the paradoxical flip. Typically the paradoxical flip is defined as displaying overconfidence and UP on hard items and underconfidence and OP on easy items. However, this is not the case. In general, participants in this study displayed overconfidence and UP across difficulty levels. Participants did not demonstrate underconfidence and OP on the easy items. Figures 2 displays that the participants are consistently overconfident (i.e., positive scores) and Figure 3 shows that participants are consistently UP (i.e., negative scores) regardless of the question difficulty level.

Hypothesis 1

Hypothesis 1 is comprised of three parts and stated that the cohort performance information will allow participants in the experimental condition to display (a) better calibration, (b) less OP/UP, and (c) less of the paradoxical flip. Specifically, participants displaying the flip would demonstrate overconfidence and UP on hard tasks, and underconfidence and OP on easy tasks. To test hypothesis 1, three separate analyses were run (i.e., one analysis with calibration as the dependent variable (DV), one with OP/UP as the DV, and one with the paradoxical flip as the DV. Each analysis was a 2 (training: cohort information and control) x 3 (difficulty: hard, medium, and easy) x 3 (domain:

entertainment, science, and sports) factorial repeated measures ANOVA with question difficulty and domain as the repeated within-subject variables and condition as the between-subjects variable.

Calibration. Mean calibration (i.e., average confidence - percent correct) differences existed between difficulty levels, $F(2, 218) = 170.80$, $p<.0001$, $\eta^2_p = .61$ and domains $F(2, 218) = 6.00$, $p < .01$, $\eta^2_P = .05$. These main effects are qualified by the difficulty by domain interaction, $F(3.33, 362.93) = 3.00$, $p = .03$, $\eta^2_p = .03$, Greenhouse-Geisser $\varepsilon = .83$ Figure 2 displays the means and standard errors for calibration. This interaction is driven by the sports domain. In the hard and easy difficulty level participants are the worst calibrated in the sports domain; contrarily, the participants display the best calibration for the medium sports questions. There is no difficulty by condition interaction, no domain by condition interaction, or three-way interaction. These analyses revealed that providing the experimental group with additional cohort information did not significantly improve calibration, $F(1,109) = 2.04$, $p = .62$. Hypothesis 1 was not supported for calibration.

OP/UP. Hypothesis 1(b) proposed that participants in the control (i.e., no cohort performance norm information) condition would display more OP/UP than participants in the experimental condition. Figure 3 displays the OP/UP means in each difficulty level in both conditions. There was a main effect for difficulty level on OP/UP, $F(2, 214) = 4.58$, $p = .01$, $\eta^2_P = .04$ and a difficulty by domain interaction, $F(4, 428) = 3.05$, $p = .02$, $\eta^2_P = .03$. The participants appeared to display less OP/UP in the medium domain than they did in the hard or easy domains. Additionally, the difficulty by condition interaction approached significance, $F(2, 106) = 3.01$, $p = .05$, $\eta^2_P = .03$. There is not a domain main

effect, domain by condition interaction, or three-way interaction. Providing the experimental group with additional cohort information did not significantly reduce OP/UP, $F(1,107) = .16$, $p = .69$. Although the additional cohort information did not have a significant main effect, the difficulty by condition interaction partially supports the hypothesis that participants in the control condition would display more OP/UP than the experimental condition. The interaction is driven by the experimental condition's small amount of OP/UP in the hard questions. Hypothesis 1 was partially supported for OP/UP.

Paradoxical Flip. Results indicated a main effect for difficulty and a difficulty by domain interaction for the flip. Mean differences (see Figure 4) in flip exist between difficulty levels, $F(2, 214) = 45.81$, $p < .0001$ $\eta^2_p = .30$ and in the difficulty by domain interaction, $F(3.64, 386.75) = 3.21$, $p = .02$, $\eta^2_p = .03$, Greenhouse-Geisser correction ($\varepsilon = .90$). Participants displayed more paradoxical flip in the hard conditions than in the medium or easy conditions. The difficulty by domain interaction was driven by the science domain. Participants displayed less of the paradoxical flip on the hard science questions than on the hard entertainment or sports; however, they displayed the most paradoxical flip on the medium science questions. There is no difficulty by condition interaction, domain effect, domain by condition interaction, or three-way interaction. Providing the experimental group with additional cohort information did not significantly reduce the flip, $F(1,107) = 1.84$, $p = .18$, $\eta^2_P = .02$. Hypothesis 1 was not supported for the paradoxical flip.

Bets

Hypothesis 2 proposed that participants who display high levels of the paradoxical flip would make fewer correct decisions in terms of winning bets. Although

the hypothesis did not concern what bets each of the groups chose, a closer look at the cross tabulations of the frequency of each bet made by condition and bet outcome (see Tables 3 to 6) provides the reader with a better understanding of the bet analyses. These cross tabulations provide the reader with a greater understanding of why the participants won or lost the bets.

This hypothesis was investigated in two ways. First analyses were run to determine if there were differences between the groups in terms of choosing the absolute and relative bets. These analyses were run based on the idea that the group receiving additional cohort information would choose more of the relative bets. Although the cohort group picked slightly more relative bets, there is not a significant difference between the groups in terms of the number of relative bets chosen, $F(1, 108) = 1.37$, $p = .25$, $\eta^2_p = .01$. Additionally, there are no difficulty, domain, difficulty by domain, domain by condition, or three-way interaction effects. However, there is a significant difficulty by condition interaction, $F(1.35, 146.27) = 3.86$, Greenhouse-Geisser correction ($\varepsilon = .68$), $\eta^2_p = .03$. The additional cohort information group picked the relative bet more often on the easy items; thus, losing more of the easy difficulty level bets.

After determining if the experimental condition chose more relative bets, a regression analysis was conducted with the flip, absolute calibration, and OP/UP as predictors of the number of bets won out of the nine each participant completed. The overall model was significant, $F(3, 104) = 8.14$, $p < .01$ indicating that at least one of the independent variables significantly predicted the number of bets won. However, the flip was not a significant predictor of winning bets; thus, hypothesis 2 was not supported. Of the three independent variables, calibration was the only significant predictor of the

number of bets won. Table 7 contains the regression results for these analyses. Absolute calibration (i.e., the absolute value of calibration scores) is significantly negatively related to the overall number of bets won. Larger absolute calibration scores result in winning fewer bets.

After looking at the overall analysis of the number of bets won, it was necessary to break the global hypothesis into its smaller components. This allows us to see if the flip, calibration, and OP/UP scores predicted individual bet outcomes. Thus, the flip score, absolute calibration score, and OP/UP scores (i.e., absolute value of the OP/UP and the original OP/UP scores) from each domain/difficulty level combination were used to predict the outcome of the bet in that domain/difficulty level combination (e.g., hard science). One logistic regression analysis was run for each of the domain/difficulty level combination, resulting in a total of nine analyses. These analyses revealed that calibration is the best predictor of winning or losing the bet. For example, absolute calibration and OP/UP were significantly negatively related to the hard entertainment bet outcomes ($p < .05$, see Table 8). These results are interpreted as follows. A one unit increase in absolute calibration results in a 6 percent decrease in the predicted odds[4] of winning the bet. A one unit increase in OP/UP results in a 2 percent decrease in the predicted odds of winning the bet. Because the large ranges of the calibration scores (i.e., -50 to 100) and OP/UP scores (i.e., -100 to 100) interpreting a ten unit increase in the calibration and OP/UP scales may be more meaningful. The Odds Ratio for a ten unit increase in calibration results in a 44 percent decrease in the predicted odds of winning the bet, while the same increase in OP/UP results in a 16 percent decrease in the predicted odds of winning the

[4] Due to the nonlinear nature of logistic regression, β's are difficult to interpret. Thus, odds ratios were interpreted (Allison, 2007).

bet. These results indicated that calibration was a much better predictor of winning the hard entertainment bet than OP/UP. The Odds Ratios and other relevant logistic regression output for all of these analyses are presented in Table 8. Each Odds ratio has the same interpretation as the previously explained logistic regression analysis.

Discussion

Contributions of the Current Study

The study addressed three contingent questions about the paradoxical reversal: is it "real", if it is "real", does it (i.e., paradoxical flip, over- and underconfidence, and OP/UP) result in losing more bets, and can their effects be minimized by providing participants with cohort performance norm information?

The answer to the first question; "Is the paradoxical reversal real?" is a tentative yes. This study partially replicated the paradoxical flip. Typically the paradoxical flip is defined as displaying overconfidence and UP on hard items and underconfidence and OP on easy items. In general, participants in this study displayed overconfidence and UP across difficulty levels. The partial replication could be due to the methodology that this study employed. Previous research (Moore & Healy, 2008) investigating the flip asked participants to estimate the number of items that they answered correctly, which is slightly different than the calibration paradigm used in this study. Larrick et al. (2007) manipulated item difficult by changing the accuracy criterion; instead, this study pilot tested items to manipulate difficulty. These slight methodological differences are likely the impetus behind the partial replication of the paradoxical flip. Determining whether the paradoxical flip is a psychological construct and not a result of methodological problems is crucial for future research involving the paradoxical flip. If the paradoxically flip was found to be the result of methodological confounds, then it is not worthy of

further study. However, this study provided further evidence that the paradoxical flip is a psychologically driven construct and that it is worthy of further investigation.

After establishing that the paradoxical flip is a psychologically plausible and replicable construct, the study looked to answer questions about the implications of the flip. Does the paradoxical flip result in losing more bets? The components (i.e., over- and underconfidence and OP/UP) of the paradoxical flip do lead to poor betting performance individually; however, when they were combined to form the paradoxical flip (i.e., |C-(OP/UP)|) it was not significantly related to betting performance. Similar to prior research (Moore & Klein, 2008) calibration was found to be the strongest predictor of the bet outcomes. This study provides additional support to the idea that absolute performance information is a better predictor of betting and decision making behavior. The demonstration that the OP/UP and over- and underconfidence biases result in poor decision making has many practical implications. Managers should be willing to provide employees information concerning their absolute as well as relative performance to facilitate improved decision making.

Upon establishing that over- and underconfidence and OP/UP do have negative effects on betting behavior, the study looked to determine if additional cohort performance information could improve the participants decision making. Providing the experimental condition with additional cohort performance information did not have a significant main effect in reducing the amount of over- and underconfidence, OP/UP, or the paradoxical flip. However, the difficulty by condition interaction in OP/UP partially supports the hypothesis that participants in the control condition would display more OP/UP than the experimental condition. The interaction is driven by the experimental

condition's small amount of OP/UP in the hard questions. Although the additional cohort information did not consistently reduce the amount of over- and underconfidence, OP/UP, or paradoxical flip, there is a plausible explanation as to why this occurred. All participants completed the statistical training, which told them, among other things, that the majority of scores are typically around the mean. This information could have increased the salience of their reference group. The participants considered how their cohort performed on the tasks which allowed the participants to demonstrate little OP on the easy questions. Thus, the additional cohort information did not have the desired effect. Prior studies (e.g., Moore & Cain, 2008) found that over- and underconfidence were resistant to feedback, and this study attempted to determine if these biases (over- and underconfidence, OP/UP, and the paradoxical flip) could be ameliorated by providing participants with additional information about their cohort's performance. As previously discussed, cohort performance information has been found to affect different conceptualizations of relative performance estimates (Moore & Klein, 2008). However, this study demonstrated that additional cohort information was ineffective in terms of improving calibration and the amount of the paradoxical flip. It provided a partial replication of the Moore and Klein's finding that additional cohort information improves relative performance estimates.

Additional Benefits of Providing Cohort Performance Information

In addition to the previously discussed benefits associated with reducing the paradoxical flip, providing participants with cohort performance information may have additional benefits as well. Kluger and Denisi (1996) developed the feedback intervention theory (FIT) to explain inconsistencies in feedback's effectiveness. This theory contains

five steps: behavior is regulated by perceived states of variables and goals, goals are organized hierarchically, only goals incongruent with the individual's perceived state receive attention and regulate behavior, attention is normally directed in the middle of the hierarchy, and feedback interventions can change the location of attention, which can affect behavior. Providing participants with cohort performance information will affect steps one, three, and five.

Providing employees with cohort performance information upon selection allows the employees to form realistic goals. Goals that are congruent with the organization's expectations allow the employee to know, which performance aspects need improvement before negative feedback is given by the organization. Thus, the employee can focus on deficient areas to avoid negative feedback from the organization. The FIT model asserts that when discrepancies occur between goals and perceived states some of the employee's limited cognitive resources are assigned to reaffirming his or her self concept in non task-related domains. This takes away from the task performance and creates a self-defeating cycle. However, if the participant knows early on where his or her relative performance ranks, then he or she may have time to reaffirm in another domain without sacrificing task performance.

Limitations and Future Directions

Despite the researcher's best efforts to reduce the methodological and statistical problems that plagued prior research, regression towards the mean and scale-end effects remained a factor in this study. This study intended to use a formula developed by Cronbach, Gleser, Nanda, and Rajaratnam (1972; i.e., $r_x (X - \mu_x) + \mu_x$) to correct for regression towards the mean. The symbols in the formula are as follows: r_x = reliability of

the ratings, X is the observed score and μ_x is the population mean of the ratings. Cronbach's alpha was calculated for every set of questions to estimate reliability, and the mean of the participants' confidence ratings for each question set were going to be used to estimate μ_x. However, the reliabilities of the items were so low that it was impractical to use the formula. Reliabilities close to zero would result in making all values, regardless of how extreme they were, very close to the mean.

It is important to note that these items were primarily designed to create the hard-easy effect, and that they were successful in this regard. Although correcting for regression towards the mean would have been advantageous, the study could not have been performed if the items did not create a strong hard-easy effect. An additional limitation to the study is that the hard-easy effect created the scale-end effects. The scale-end effects are not surprising because it is extremely difficult to create a strong hard-easy manipulation without using the extremes of the accuracy scale.

After highlighting the limitations of the current study, it is important to provide some suggestions for future research that will allow other researchers to avoid these concerns. Using items that have been previously validated or completing a validation study that investigates item difficulty and internal consistency would allow future researchers to correct for regression towards the mean. Researchers could also manipulate the accuracy criterion to potentially eliminate the regression towards the mean and scale-end effects. This method would likely reduce scale-end effects because it would allow the researchers to use the full scale of confidence ratings (i.e., 0 to 100 percent confident).

Conclusion

To the author's knowledge, this is the first study to investigate the consequences of the paradoxical flip. Additionally, it is the only attempt to investigate the effect of cohort performance information on confidence, calibration, and the paradoxical flip[5] when using actual cohort performance information (i.e., Moore and Klein (2008) gave participants random feedback). The strength of the paradoxical flip was tested by investigating the effects that cohort performance information has on OP/UP, calibration, and the paradoxical flip. Results revealed that the OP/UP, over- and underconfidence, and the paradoxical flip were not reduced consistently by the additional information. Future research investigating ways to reduce these biases must be performed in order to reduce their harmful decision making consequences. Reducing the amount of bias in participants' subjective confidence ratings could provide benefits in many domains (e.g., organizations, economics, and gambling; Bolger et al., 2008).

[5] Moore and Small (2007) tested the effects of cohort information on perceived percentiles.

References

Allison, P. D. (2007). *Logistic regression using the SAS system: Theory and application.* Cary, NC: SAS Institute, Inc.

Bazerman, M. H., & Neale, M. A. (1982). Improving negotiation effectiveness under final offer arbitration: The role of selection and training. *Journal of Applied Psychology 67*, 543-548.

Benabou, R., & Tirole, J. (2002). Self-confidence and personal motivation. *Quarterly Journal of Economics, 117*(3), 871–915.

Bolger, F., Pulford, B. D., & Coleman, A. M. (2008). Market entry decisions: Effects of absolute and relative confidence. *Experimental Psychology, 55*, 113-120.

Campbell, D. T., & Kenny, D. A., (1999). *A primer on regression artifacts.* New York: The Guilford Press.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972).*The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: implications for health, education, and business. *Psychological Science in the Public Interest, 5*, 69–106.

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

Fong, G. T., & Nisbett, R. E. (1991). Immediate and delayed transfer of training effects in statistical reasoning. *Journal of Experimental Psychology: General, 120*, 34-45.

Fong, G. T., Krantz, D. H., & Nisbett, R. E. (1986). The effects of statistical training on thinking about everyday problems. *Cognitive Psychology*, *18*, 253- 192.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance, 3,* 552-564.

Gino, F., & Moore, D. A. (2007). Effects of task difficulty on use of advice. *Journal of Behavioral Decision Making, 20,* 21–35.

Jonsson, A. C., & Allwood, C. M. (2003). Stability and variability in the realism of confidence judgments over time, content domain, and gender. *Personality and Individual Differences*, 34, 559-574.

Juslin, P., Anders, W., & Henrik, O., (2000). Naive empiricism and dogmatism in confidence research: A critical examination of the hard-easy effect. *Psychological Review, 107, 384 -396.*

Larrick, R. P., Burson, K. A., & Soll, J. B. (2007). Social comparison and confidence: When thinking you're better than average predicts overconfidence (and when it does not). *Organizational Behavior and Human Decision Processes, 102,* 76-94.

Klein, W. M. P. (1997). Objective standards are not enough: Affective, self-evaluative, and behavioral responses to social comparison information. *Journal of Personality and Social Psychology*, 72, 763–774.

Keren, G. (1991) Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica, 77,* 217-273.

Kluger, A. N., & DeNisi, A. (1996). Effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119,* 254-284.

Kosonen, P., & Winne, P. H. (1995). Effects of teaching statistical laws on reasoning about everyday problems. *Journal of Educational Psychology, 87,* 33–46.

Kruger, J. (1999). Lake wobegon be gone! The "below-average effect" and the egocentric nature of comparative ability judgments. *Journal of Personality and Social Psychology, 77,* 221-232.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77,* 1121-1134.

Moore, D. A., & Cain, D. M. (2007). Overconfidence and underconfidence: When and why people underestimate (and overestimate) the competition. *Organizational Behavior and Human Decision Processes, 103,* 197-213.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review, 115,* 502-517.

Moore, D. A., & Kim, T. G. (2003). Myopic social prediction and the solo comparison effect. *Journal of Personality and Social Psychology, 85,* 1121–1135.

Moore, D., & Klein, W. (2008). Use of absolute and comparative performance feedback in absolute and comparative judgments and decisions. *Organizational Behavior and Human Decision Processes, 107,* 60-74.

Moore, D. A., & Small, D. A. (2007). Error and bias in comparative judgment: On being both better and worse than we think we are. *Journal of Personality and Social Psychology, 92, 972-989.*

Landy, F. J., & Conte, J. M. (2007). *Work in the 21st century: An introduction to industrial and organizational psychology* (2nd ed.). Malden, MA: Blackwell Publishing.

Murphy, K. R., & Shiarella, A. H. (1997). Implications of the multidimensional nature of job performance for the validity of selection tests: Multivariate frameworks for studying test validity. *Personnel Psychology, 50,* 823-854.

Ronis, D. L., & Yates, F. J. (1987). Components of probability judgment accuracy: Individual consistency and effects of subject matter and assessment method. *Organizational Behavior and Human Decision Processes, 40,* 193-218.

Roy, M., & Liersch, M. (2008, November). People believe that they are prototypical, not above-average. Poster presented at the Society for Judgment and Decision Making, Chicago, IL.

Sniezek, J. A., Gunnar E. S., & Reeshad S. D. (2004). Improving judgment with prepaid expert advice. *Journal of Behavioral Decision Making,* 17, 173–90.

Table 1

Pairwise Comparisons of Difficulty Levels

Difficulty Comparison		Mean Difference	95% Confidence Interval
Hard	Medium	-34.58*	[-37.44, -31.72]
Hard	Easy	-56.93*	[-58.91, -54.94]
Medium	Easy	34.58*	[31.72, 37.44]

* indicates significant difference between comparisons with adjusted alpha = (Bonferroni correction).

Table 2

Paradoxical Flip in Each Difficulty Level

Difficulty	*M*	95% Confidence Interval
Hard	45.169	[41.67, 48.67]
Medium	30.892	[28.15, 33.63]
Easy	23.420	[18.97, 27.87]

Table 3

\

*Frequency Count for Hard Bets * Condition * Bet Outcome*

Bet Outcome			Condition		
			Cohort Information	No Information	Total
Loss	Bet Made	Absolute	84	72	156
		Relative	45	44	89
		Total	129	116	245
Win	Bet Made	Absolute	3	7	10
		Relative	32	45	77
		Total	35	52	87

Note: Participants were asked to choose between an absolute and relative bet. To win the absolute bet participants must respond correctly to at least 70 % of the items. The relative bets required the participants to perform better than or equal to 70 % of their cohort. The information in these tables is collapsed across the three domains (i.e., entertainment, science, and sports).

Because the questions are difficult, participants were more likely to win the bet if they chose the relative bet. Participants who chose the relative bet won 46% of the time (i.e., 77/166). Participants who chose the absolute bet only won 6% (i.e., 10/166) of the time. Overall, 26% of participants won their bet on the hard items.

Table 4

*Frequency Count for Medium Bets * Condition * Bet Outcome*

Bet Outcome			Condition		
			Cohort Information	No Information	Total
Loss	Bet Made	Absolute	31	31	62
		Relative	31	40	71
		Total	62	71	133
Win	Bet Made	Absolute	62	60	122
		Relative	41	37	78
		Total	103	97	200

Note: Participants were asked to choose between an absolute and relative bet. To win the absolute bet participants must respond correctly to at least 70 % of the items. The relative bets required the participants to perform better than or equal to 70 % of their cohort. The information in these tables is collapsed across the three domains (i.e., entertainment, science, and sports).

Participants who chose the relative bet won 52% of the time (i.e., 78/149). Participants who chose the absolute bet only won 66% (i.e., 122/184) of the time. Overall, 60% (i.e., 200/333) of participants won their bet on the medium items.

Table 5

*Frequency Count for Easy Bets * Condition * Bet Outcome*

Bet Outcome			Condition		
			Cohort Information	No Information	Total
Loss	Bet Made	Absolute	0	0	0
		Relative	33	19	52
		Total	33	19	52
Win	Bet Made	Absolute	77	115	192
		Relative	55	34	89
		Total	132	149	281

Note: Participants were asked to choose between an absolute and relative bet. To win the absolute bet participants must respond correctly to at least 70 % of the items. The relative bets required the participants to perform better than or equal to 70 % of their cohort. The information in these tables is collapsed across the three domains (i.e., entertainment, science, and sports).

Participants who chose the relative bet won 63% of the time (i.e., 78/149). Participants who chose the absolute bet only won 100% (i.e., 192/192) of the time. Overall, 84% (i.e., 281/333) of participants won their bet on the easy items.

Table 6

*Frequency Count for All Bets * Condition * Bet Outcome*

Bet Outcome			Condition		
			Cohort Information	No Information	Total
Loss	Bet Made	Absolute	115	103	218
		Relative	109	103	212
		Total	224	206	430
Win	Bet Made	Absolute	142	182	324
		Relative	128	116	244
		Total	270	361	568

Note: Participants were asked to choose between an absolute and relative bet. To win the absolute bet participants must respond correctly to at least 70 % of the items. The relative bets required the participants to perform better than or equal to 70 % of their cohort. The information in these tables is collapsed across the three domains (i.e., entertainment, science, and sports).

Participants who chose the relative bet won 54% of the time (i.e., 244/456). Participants who chose the absolute bet only won 76% (i.e., 684/902) of the time. Overall, 68% (i.e., 928/1368) of participants won their bet on the all items.

Table 7

Parameter Estimates for Flip, Calibration, and OP/UP Predicting the Number of Bets Won

Independent Variable	*B*	*SE B*	*β*	*Tolerance*
Intercept	6.26	.33	0	
Calibration	-.07	.02	-.39**	.80
OP/UP	-.006	.008	-.08	.68
Paradoxical Flip	.005	.001	-.08	.69
R^2		.19		
F		8.14**		

*p < .05, ** p < .01

Table 8

Logistic Regression Results for each Domain/Difficulty Level Combination

	Goodness of Fit Test							Adjusted Odds	Max rescaled
Variable	χ^2	*df*	*p*	*B*	*SE*	Wald χ^2	*df*	Ratio	r^2
Hard Entertainment	9.46	8	.31			14.07*	2		.26
Calibration				-.06	.02	10.35*	1	.56	
OP/UP				-.02	.01	5.05*	1	.84	
Constant				-.40	.46	.77	1	NA	
Hard Science	6.69	8	.57			14.58*	2		.22
Calibration				-.03	.01	6.01*	1	.71	
OP/UP				-.02	.01	9.23*	1	.80	
Constant				-.23	.38	.36	1	NA	
Hard Sports	7.17	8	.52			18.42*	1		.33
Calibration				-.07	.02	18.41*	1	.49	
Constant				.81	.43	3.48	1	NA	

Note: Odds were adjusted to provide odds estimates for a 10 unit increase in the dependent variable. Calibration = Absolute Calibration. The goodness of fit test is the Hosmer and Lemeshow test. Failing to reject the null hypothesis suggests that the model fits the data well. All *Bs* excluding the intercepts that were included in these tables were significant. *Significant at $p < .05$.

Table 8 (continued)

Logistic Regression Results for each Domain/Difficulty Level Combination

	Goodness of Fit Test							Adjusted	Max
								Odds	rescaled
Variable	χ^2	*df*	*p*	*B*	*SE*	Wald χ^2	*df*	Ratio	r^2
Medium Entertainment	7.65	8	.47			25.46*	2		.44
Calibration				-.07	.02	16.59*	1	.48	
OP/UP				-.04	.01	18.67*	1	.67	
Constant				1.89	.42	20.35*	1	NA	
Medium Science	10.44	8	.24			22.25*	2		.38
Calibration				-.07	.02	9.14*	1	.50	
OP/UP				-.04	.01	19.86*	1	.65	
Constant				.94	.42	4.98*	1	NA	
Medium Sports	12.90	8	.11			21.33*	2		.40
Calibration				-.09	.03	11.61*	1	.42	
OP/UP				-.04	.01	14.63*	1	.70	
Constant				2.33	.51	20.45*	1	NA	

Note: Odds were adjusted to provide odds estimates for a 10 unit increase in the dependent variable. Calibration = Absolute Calibration. The goodness of fit test is the Hosmer and Lemeshow test. Failing to reject the null hypothesis suggests that the model fits the data well. All *B* excluding the intercepts that were included in these tables were significant. *Significant at $p < .05$.

Table 8 (continued)

Logistic Regression Results for each Domain/Difficulty Level Combination

	Goodness of Fit Test					Wald		Adjusted Odds	Max rescaled
Variable	χ^2	*df*	*p*	*B*	*SE*	χ^2	*df*	Ratio	r^2
Easy Entertainment	6.27	6	.39			13.64*	2		.43
Calibration				-.24	.06	13.49*	1	.10	
OP/UP				-.02	.02	.92	1	.86	
Constant				4.54	.92	24.13*	1	NA	
Easy Science	12.15	8	.14			6.11	2		.16
Calibration				-.14	.06	5.30*	1	.25	
OP/UP				-.02	.01	2.68	1	.85	
Constant				3.82	.73	27.75*	1	NA	
Easy Sports	6.72	6	.35			16.41*	1		.26
Calibration				-.20	.05	16.41*	1	.14	
Constant				3.05	.67	20.81*	1	NA	

Note: Odds were adjusted to provide odds estimates for a 10 unit increase in the dependent variable. Calibration = Absolute Calibration. The goodness of fit test is the Hosmer and Lemeshow test. Failing to reject the null hypothesis suggests that the model fits the data well. All *B* excluding the intercepts that were included in these tables were significant. *Significant at $p < .05$.

Table 8 (continued)

Logistic Regression Results for each Domain/Difficulty Level Combination

	Goodness of Fit Test					Wald		Adjusted Odds	Max rescaled r^2
Variable	χ^2	*df*	*p*	*B*	*SE*	χ^2	*df*	Ratio	
Hard Entertainment	4.95	8	.76			11.96*	2		.22
Calibration				-.06	.02	10.42*	1	.57	
Abs. OP/UP				-.02	.01	1.98	1	1.18	
Constant				-.64	.58	1.21	1	NA	
Hard Science	6.52	8	.59			6.81*	1		
Calibration				-.04	.01	7.83*	1	.68	.12
Constant				.11	.36	.09	1	NA	
Hard Sports	7.17	8	.52			18.42*	1		.33
Calibration				-.07	.02	18.41*	1	.49	
Constant				.81	.43	3.48	1	NA	

Note: Odds were adjusted to provide odds estimates for a 10 unit increase in the dependent variable. Calibration = Absolute Calibration. Abs. OP/UP = Absolute over-and under-placement. The goodness of fit test is the Hosmer and Lemeshow test. Failing to reject the null hypothesis suggests that the model fits the data well. All *B* excluding the intercepts that were included in these tables were significant. *Significant at $p < .05$.

Table 8 (continued)

Logistic Regression Results for each Domain/Difficulty Level Combination

	Goodness of Fit Test							Adjusted Odds	Max rescaled
Variable	χ^2	*df*	*p*	*B*	*SE*	Wald χ^2	*df*	Ratio	r^2
Medium Entertainment	7.41	8	.49			20.30*	2		.31
Calibration				-.07	.02	16.54*	1	.50	
OP/UP				-.04	.01	8.60*	1	.67	
Constant				2.63	.58	20.94*	1		
Medium Science	6.02	8	.64			10.33*	2		.13
Calibration				-.05	.02	6.66*	1	.63	
Abs. OP/UP				.03	.01	4.55*	1	1.32	
Constant				.27	.47	.32	1		
Medium Sports	5.51	7	.60			17.75*	1		.31
Calibration				-.12	.03	17.75*	1	.31	
Constant				2.90	.57	26.30*	1		

Note: Odds were adjusted to provide odds estimates for a 10 unit increase in the dependent variable. Calibration = Absolute Calibration. Abs. OP/UP = Absolute over-and under-placement. The goodness of fit test is the Hosmer and Lemeshow test. Failing to reject the null hypothesis suggests that the model fits the data well. All *B* excluding the intercepts that were included in these tables were significant. *Significant at $p < .05$.

Table 8 (continued)

Logistic Regression Results for each Domain/Difficulty Level Combination

	Goodness of Fit Test					Wald	Adjusted Odds	Max rescaled
Variable	χ^2	*df*	*p*	*B*	*SE*	χ^2	Ratio	r^2
Easy	4.82	3	.19					.41
Entertainment						13.22*		
Calibration				-.24	.06	13.22*	.10	
Constant				4.64	.94	24.21*		
Easy Science	3.18	5	.67			4.10*		.10
Calibration				-.11	.05	4.10*	.34	
Constant				3.60	.65	30.88*		
Easy Sports	6.72	6	.35			16.41*		.26
Calibration				-.20	.05	16.41*	.14	
Constant				3.05	.67	20.81*		

Note: Odds were adjusted to provide odds estimates for a 10 unit increase in the dependent variable. Calibration = Absolute Calibration. Abs. OP/UP = Absolute over-and under-placement. The goodness of fit test is the Hosmer and Lemeshow test. Failing to reject the null hypothesis suggests that the model fits the data well. All *B* excluding the intercepts that were included in these tables were significant. For the Wald $\chi 2$ the degrees of freedom were 1. *Significant at $p < .05$.

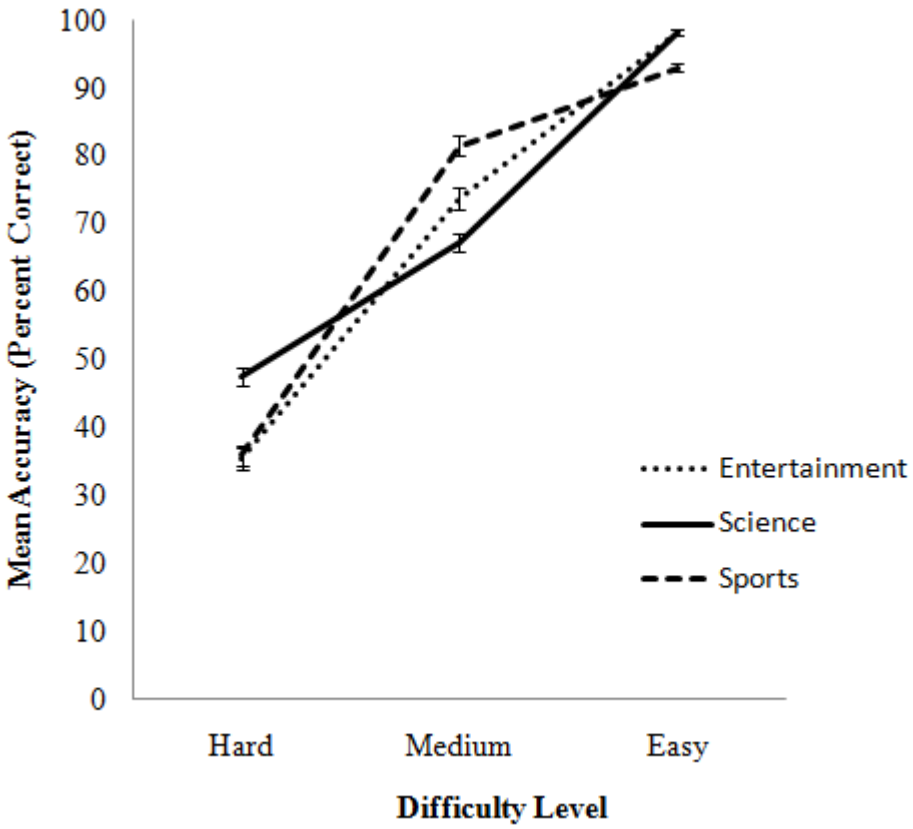


Figure 1. Mean accuracy and standard error per domain for each question difficulty level

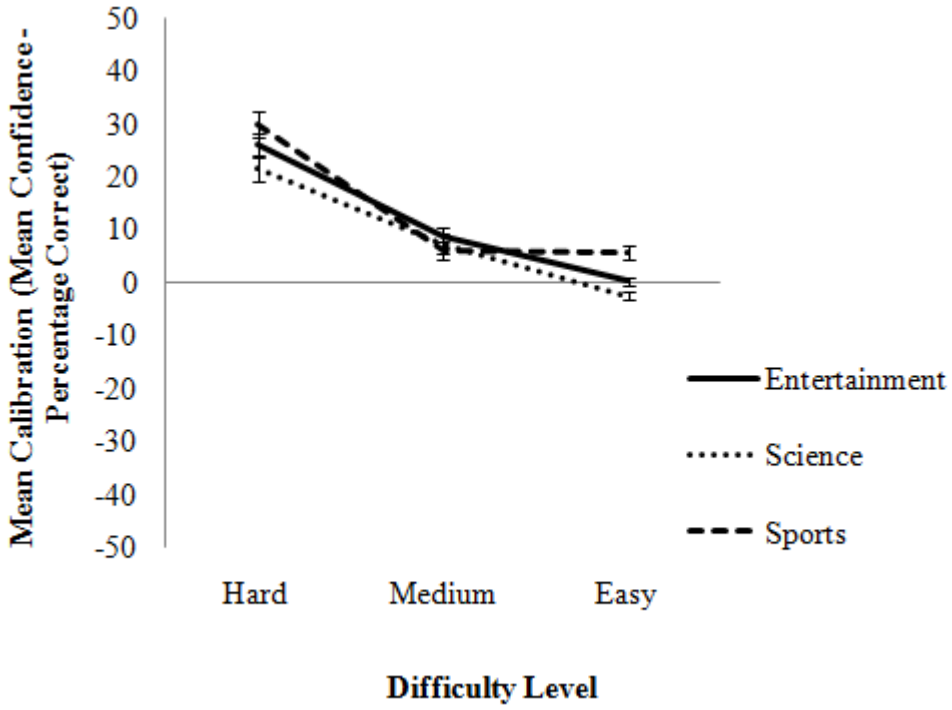


Figure 2. Mean calibration and standard error per domain for each question difficulty level.

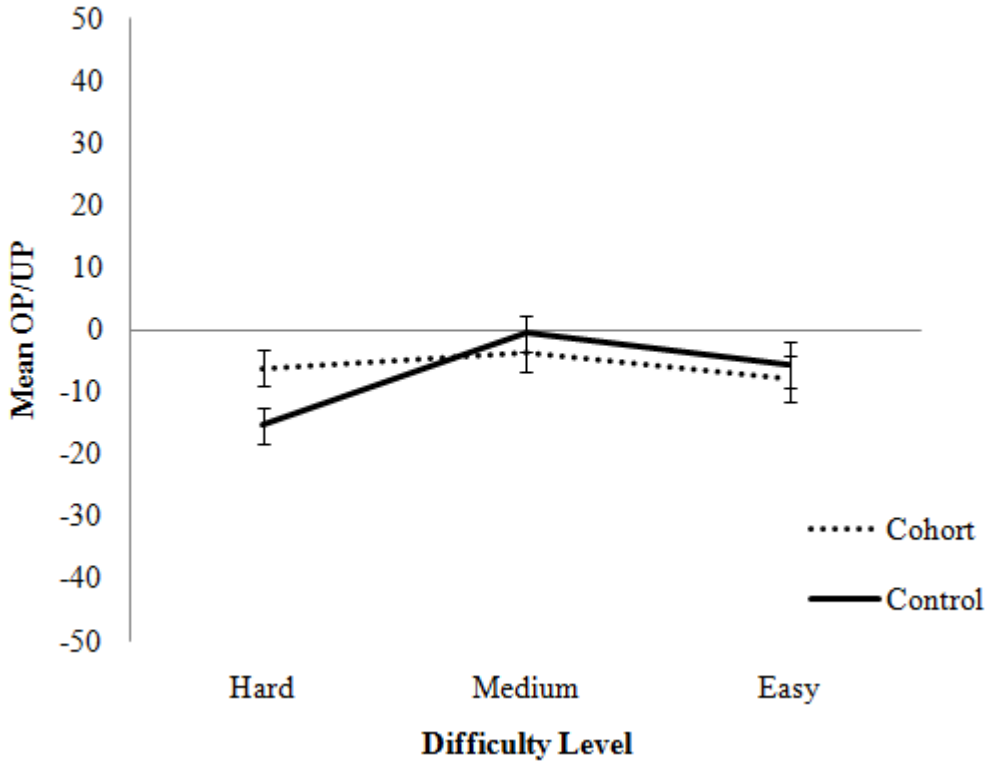


Figure 3. Mean OP/UP and standard error per domain for each difficulty level.

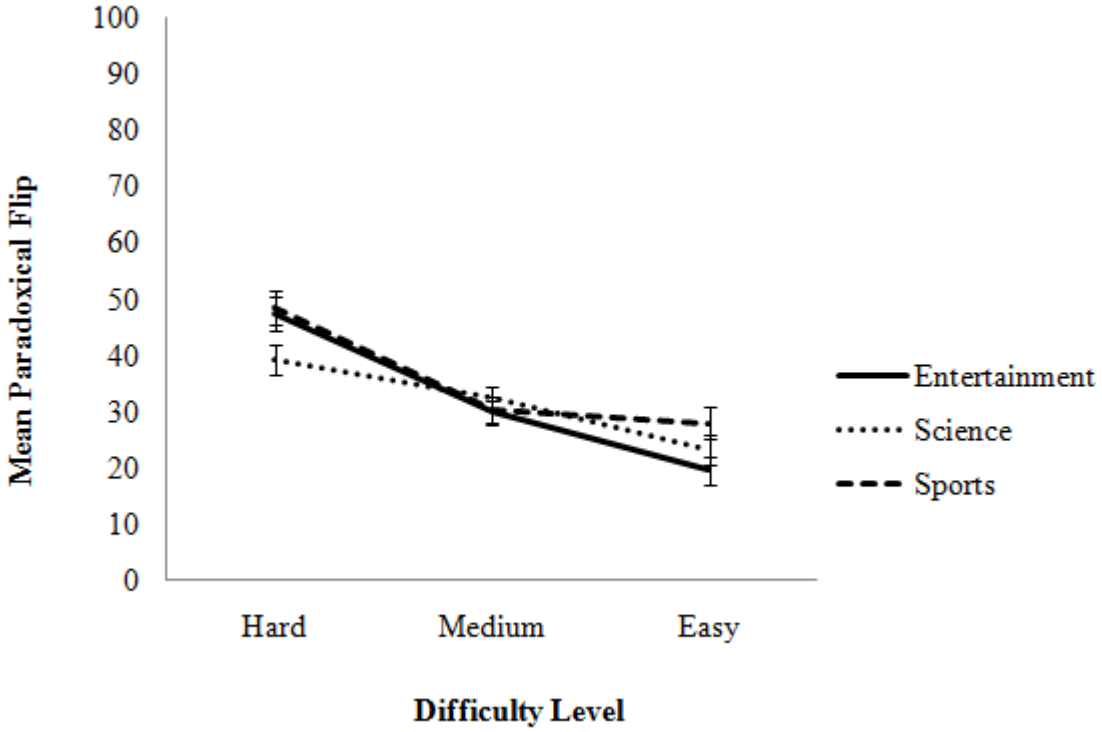


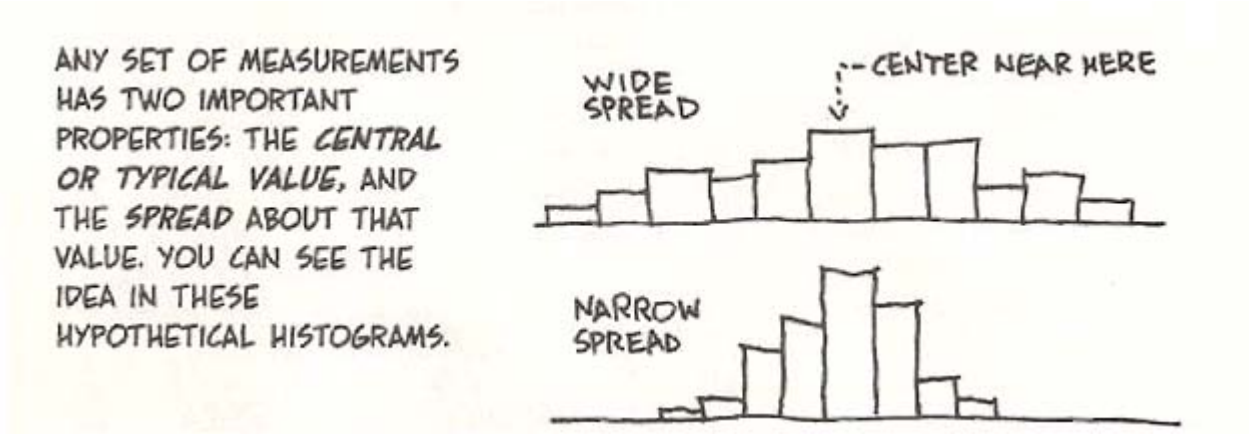
Figure 4. Mean paradoxical flip and standard error per domain for each question difficulty level.

Appendix A

Statistical Training

In this study, you will complete several tasks. We will provide you with instructions before you start each task. Please try to respond as accurately as possible. In this task, we will provide you with some information about different distributions. Your job is to learn this information and we will test your knowledge about distributions before you begin the study. Press the spacebar to begin this task.

We will now describe some characteristics of distributions. You will learn about the mean (average), the median, and percentiles of a distribution of scores. Please read this information carefully because we will test your knowledge of this information. You must respond correctly on the test to continue with the study. Press the spacebar to begin.



The bar with the arrow pointed to it occurs the most.

SUPPOSE, FOR EXAMPLE, WE ASK FIVE PEOPLE HOW MANY HOURS OF *TELEVISION* THEY WATCH IN A WEEK... AND GET THE FOLLOWING ARRAY:

OBSERVATION	1	2	3	4	5
DATA VALUE	5	7	3	38	7

THEN $x_1 = 5$, $x_2 = 7$, $x_3 = 3$, $x_4 = 38$, AND $x_5 = 7$.

WHAT'S THE "CENTER" OF THESE DATA? THERE ARE ACTUALLY SEVERAL DIFFERENT WAYS TO MEASURE IT. WE'LL LOOK AT JUST TWO OF THEM.

THE MEAN (OR "AVERAGE")

THE *MEAN* OR AVERAGE VALUE IS REPRESENTED BY $\bar{x}$... IT'S OBTAINED BY ADDING ALL THE DATA AND DIVIDING BY THE NUMBER OF OBSERVATIONS:

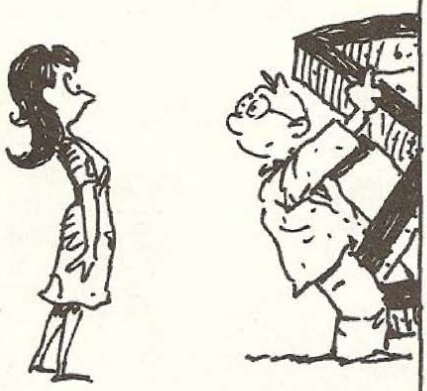
$$\bar{x} = \frac{\text{SUM OF DATA}}{n}$$

$$= \frac{x_1 + x_2 + \ldots + x_n}{n}$$

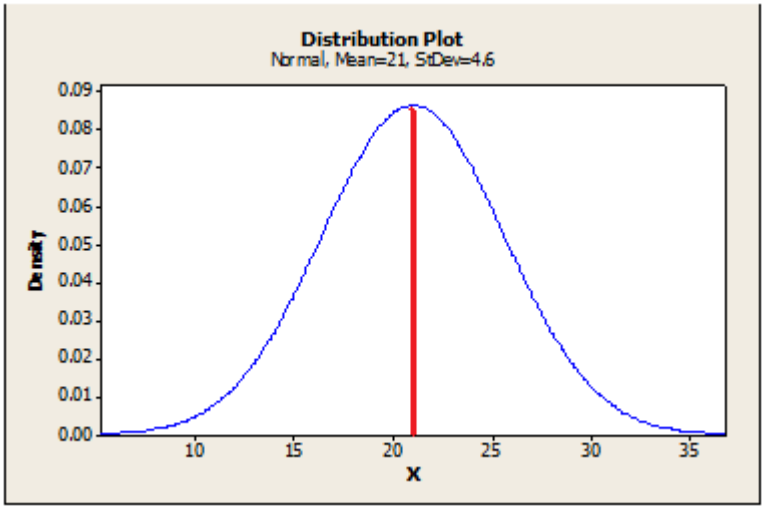
FOR OUR EXAMPLE,

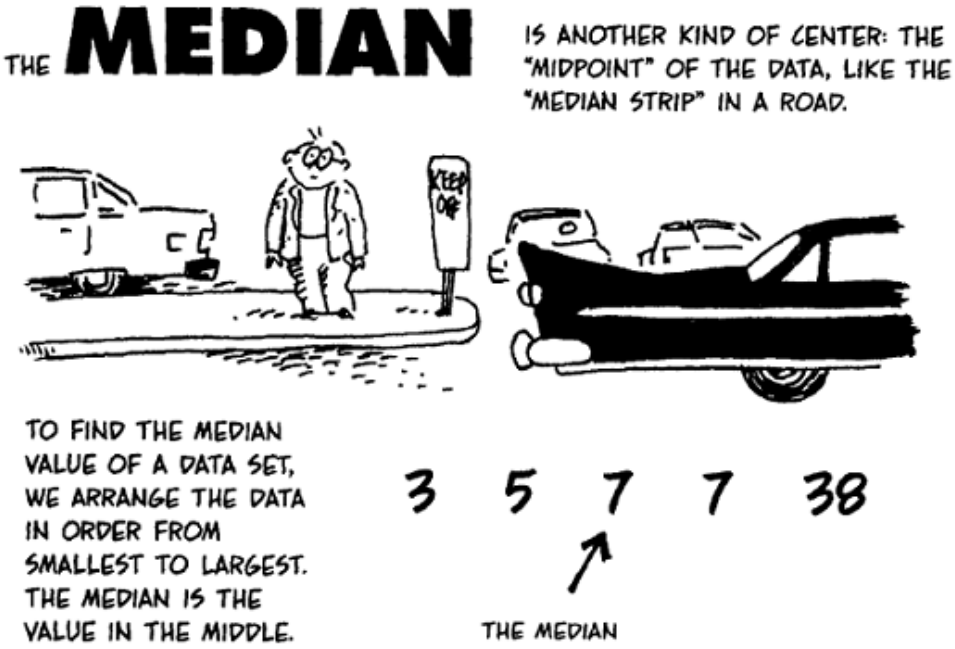
$$\bar{x} = \frac{5 + 7 + 3 + 38 + 7}{5} = \frac{60}{5}$$

$$= 12 \text{ HOURS}$$



- A distribution of scores is a graph of the data that provides information concerning the score values as well as their frequency.
    - The highest point on the graph denotes the most frequent score.
    - For example, in the picture below, the red line denotes the score that occurs most often.





Now you will learn about percentile ranks.

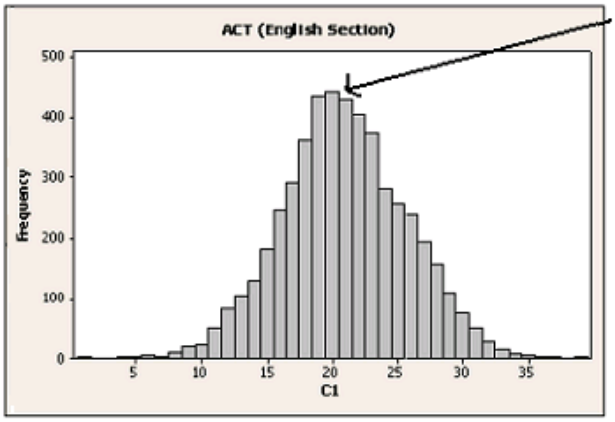
- Percentiles are used to describe the location of a single score in a distribution of scores.
    - A percentile refers to the number of people who performed equal to or worse than you on a given task.
    - Perceived percentiles range from 0 to 100 and the median percentile for a distribution of scores is 50.
    - Perceived percentile can be thought of as the number of people out of 100 you think you performed better than.

- The next slide displays a graph and chart displaying the distribution of ACT scores.

- Both the graph and the chart reveal that the majority of test takers score in the middle of the range.

On the American College Test (ACT), a standardized test that many students take before they enter college the scores range from one (the lowest score possible) to 36 (the highest score possible). Both the graph and the chart reveal that the majority of test takers score in the middle of the range.

A good way to remember that the majority of the scores lie in the middle of a distribution is to think of your friends who took the ACT. Compare the number of your friends that scored a 25 on the ACT to the number of them that scored a 36. Which one is larger?



The score made most frequently is 20 (highest point on the graph).

ACT Score Quartile Values for All Students

Quartile	English
Q3 (75th Percentile)	25
Q2 (50th Percentile)	21
Q1 (25th Percentile)	16

Q3 (75th Percentile):

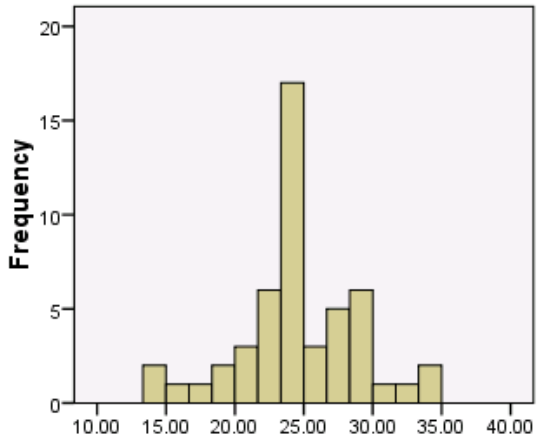
If you scored a 25 then you were better than 75% of the other students.

Q2 (50th Percentile = the median):

If you scored a 21 then you were better than 50% of the other students.

Q1 (25th Percentile):

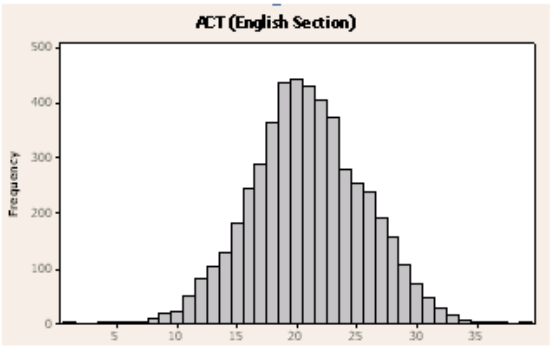
If you scored a 16 then you were better than 25% of the other students.



1. What is the most frequent value?

1) 25 2) 30 3) 20 4) 35

Select the number that corresponds to the correct answer (1,2,3, or 4)



2. In what range would you expect most people to score?

1)13-17 2)18-22 3)23-27 4)28-32

Select the number that corresponds to the correct answer (1,2,3, or 4)

To complete the training the participants had to answer the previous questions correctly.

Appendix B

Example Items

Please provide an answer that you believe to be correct. If you are not certain, please provide your best guess. Provide a confidence rating from 50 to 100 concerning the accuracy of your choice. A confidence rating of fifty represents an absolute guess and 100 represents complete confidence. After you have completed a group questions you will be asked to rate your perceived percentile (i.e., how many people out of 100 you believed you performed better than).
Examples of hard, medium, and easy from each domain:

Entertainment

Hard

7. Who played Mozart in the 1984 film "Amadeus"?
a. Jeffrey Jones
b. **Tom Hulce**

Medium

Who is the lead singer of Coldplay?
a. Stephen Tyler
b. **Chris Martin**

Easy

Sporty Spice, Baby Spice, and Posh Spice were members of what musical group?
a. **Spice Girls**
b. TLC

Sports

Hard

What 6'11" pitcher has risen an inch above Randy Johnson to be the tallest major league Baseball pitcher ever?
a. Chris Young
b. John Rauch

Medium

What team did Larry Bird play for?
a. Celtics
b. Lakers

Easy

What Chicago Bulls guard wore number 23 and led his team to 6 championships in the 1990's?
a. Michael Jordan
b. Ben Gordon

Science/Nature

Hard

What is the only breed of cat that does not have retractable claws?
a. Tiger
b. Cheetah

Medium

What vitamin is mostly required for blood coagulation?
a. Vitamin C
b. Vitamin K

Easy

An octopus has how many arms?
a. 8
b. 6

Appendix C

Bets

Participants were provided with the following information concerning the task

"The following task asks you to choose the bet that you believe you are most likely to win. The bets are based on your performance on the previous set of 10 items. Winning the bet will not result in monetary gain. Please keep in mind that you cannot choose both of the bets."

Bets

Absolute performance based bet

"*To win this bet, you will need to get more than 7 of the 10 items correct*."

Relative performance based bet

‘‘To win the bet, you will need to do better than average on the test. That is, you will need to score better than at least 70% of the other test-takers.