Perception is Relative: Sequential Contrasts in the Field *

Saurabh Bhargava
(JOB MARKET PAPER #1)
UC Berkeley
saurabh@econ.berkeley.edu

This version: December 23, 2007

Abstract

A large psychological literature suggests that individuals rely on comparative perception when making sequential decisions or assessments. Such a perceptual bias could influence behavior in settings from employee hiring and medical diagnosis to investment appraisal and product evaluation. This study presents a theoretical framework which offers predictions to differentiate perceptual errors from rational behavior and provides empirical evidence for such "contrast effects." The empirical focus is an analysis of sequential exam evaluation in a large undergraduate course with supporting evidence for perceptual errors in sentencing decisions by judges in PA courts and in dating decisions by speed daters. There is modest evidence that graders are negatively biased when evaluating exams which follow high and low streaks of extreme exams. Relative to a typical score, this effect is on the order of a 12% increase in leniency following a streak of three low scoring exams, and a 6% grade reduction following three high scoring exams. Ostensibly, learning or quota constraints for high and low grades could rationally account for this finding. However, the fact that these effects: (i) decay fully after a single period, (ii) persist despite grader experience, and (iii) are non-existent for highly transparent (multiple-choice) questions suggests an alternative explanation. Stronger evidence exists for contrast effects in judicial sentencing. Judges are 9% more likely to be lenient in sentencing of summary offenses after exposure to a criminal felony. The effects disappear for days with exposure to multiple felonies. In dating, highly attractive or unattractive prior partners produce a 13 to 17% distortion relative to the baseline decision to date. An original survey of real estate agents suggests that these effects may extend to non-random settings such as home purchases.

*This paper was heavily shaped by numerous discussions with Stefano DellaVigna, Botond Koszegi, and Matt Rabin, and I owe them a special thanks. David Card, Paul Chen, Salar Jahedi, Shachar Kariv, Prasad Krishnamurthy, Ritu Mahajan, Day Manoli, Enrico Moretti, Vikram Pathania, Uri Simohnson, Kevin Stange, Sharad Tandon, Aman Vora and participants of the Psychology & Economics Seminar at UC Berkeley also provided thoughtful comments and feedback. David Moyer, Deb Weber, Leah Woolsey, and anonymous others generously helped secure data for this project.

1 Introduction

Most economic models assume that in sequential decisions, a prior observation influences later choices only insofar as it provides new information or affects preferences. As an example, imagine an experienced judge who must evaluate a series of defendants. According to the standard model the evaluation of one defendant should be immune from the influence of a prior defendant whose case does not provide relevant information. Concretely, if a case involving a traffic violation is preceded by a case involving a violent assault, the judge should not *perceive* the traffic violation more leniently than had it instead been preceded by a more modest charge. Similarly, a doctor's diagnosis of a patient, an employer's evaluation of a job candidate, or an instructor's grade of an exam should not be influenced by immediate prior patients, candidates or exams that don't provide information which affect the decision-maker's beliefs or preferences.

However, psychologists assert that such sequential context may systematically bias perception—as well as the resulting judgments. In the lab, subjects asked to evaluate the guilt of a particularly egregious criminal tend to judge subsequent criminals with greater lenience (Pepitone and DiNubile 1976). Photographs of unattractive individuals, descriptions of highly expensive cameras, and scrutiny of hostile behavior elicit comparable generosity in subsequent evaluations of a similar nature (Kenrick and Gutierres 1980; Herr 1986; Simonson and Tversky 1992). An earlier class of studies demonstrates that subjects systematically overestimate or underestimate sensory dimensions such as the length of a line, weight of an object, or loudness of a sound after prior exposure to extreme instances of such stimuli (Hood 1950; Hovland et. al. 1958; Krantz and Campbell 1961). Figure 1 illustrates this type of error in spatial perception. The illusion that the central disk on the left is larger than the central disk on the right is caused by the surrounding disks which serve as a perceptual contrast.

While studies of social perception arguably do not rule out alternative explanations as convincingly as studies of physical perception, this pattern of comparative assessment or judgment is consistent with a sequential *contrast effect*. Contrast effects are defined here as a negative (positive) error in perception, relative to a Bayesian baseline, prompted by recent exposure to more positive (more negative) information along some dimension.[1]

This paper provides a model of sequential decision-making based on the claim that perceptions are fundamentally relative, and then documents empirical evidence for such behavior across three domains. While this research is motivated by the belief that contrast

[1] Much of this literature is not necessarily invested in labeling this behavior as a cognitive or perceptual "error." The earlier literature which examines contrast effects in the sensory perception of physical stimuli is much more explicitly concerned with whether such behavior constitutes an error than later experiments which examine different types of social perception.

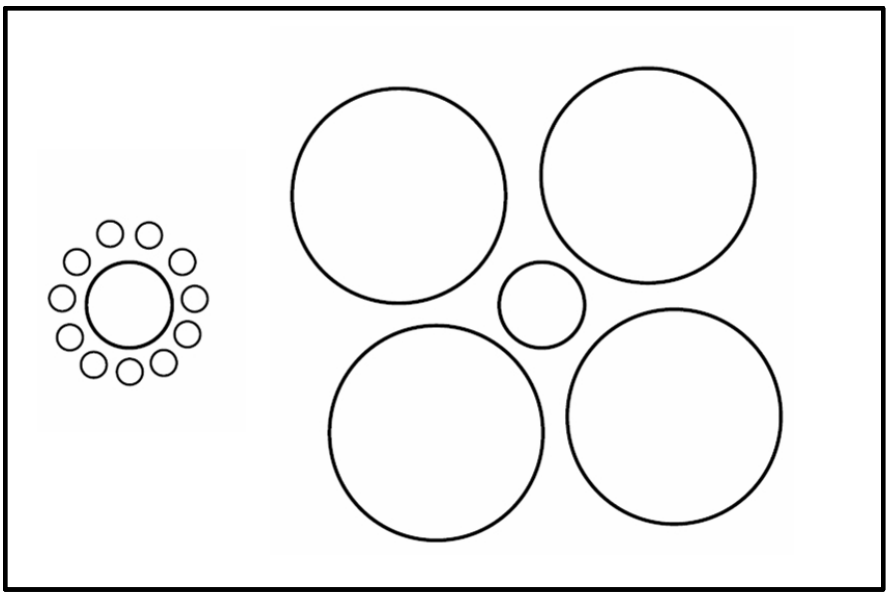

Figure 1, An Example of a Spatial Perceptual Contrast
(Based on Ramachandran & Blakeslee, *Phantoms in the Brain*)

effects influence decisions across many economically relevant settings, I principally explore such effects in the grading of exams from a large undergraduate course. I additionally summarize evidence from two other sequential decision-making domains. The first involves judicial decisions in lower Pennsylvania courts (Bhargava and Cann 2007), while a second concerns dating decisions by participants of speed dating sessions (Bhargava and Fisman 2007).[2]

The theoretical framework juxtaposes a model of contrast-effects with a standard model of Bayesian learning and rational preferences. The framework produces four main predictions. A first prediction of the contrast-effects model, the "antecedent effect," holds that current evaluations are negatively related to past perceptions. This prediction alone cannot differentiate contrast-effects from rational behavior. Bayesian learning could also cause a decision-maker to negatively correlate present and past evaluations. For example, in the context of grading, an instructor unaware of the underlying distribution of student ability, should rationally rely on the quality of one exam to update priors concerning the broader population. Alternatively, an instructor bound to a limited number of "A's"—due, for instance, to institutional mandates or norms—may be sensitive to prior grades even in the face of independent and identically distributed exams. The Appendix models the possible existence of such a "quota" constraint as a finite dynamic programming problem.

[2]Speed dating refers to a structured match-making process in which men and women meet multiple partners through a series of short, sequential interactions.

A second prediction of the contrast-effects model is that the influence of a past exam on a future evaluation decays as the intervening distance increases and the contrast effect fades. For a Bayesian in the standard model, even one subject to quotas, the influence of a past exam on a current evaluation should be invariant to the ordering of the exam. For example, an inference made in period 10 should weight the nine past observations equally while a decision-maker facing a constraint should be insensitive to whether a perfect score was achieved in period 4 or period 9 so long as the distribution of exams is unchanged.

A third prediction of the model of contrasts is that the antecedent effect persists even in the face of accumulated experience. In the standard model with learning, the effect diminishes as a grader evaluates more exams. For example, the first exam should command greater influence on the second period inference than the 50th exam has on the inference in the 51st period. A final prediction is that highly transparent targets, which permit little interpretive discretion, do not elicit any perceptual errors even if the decision-maker is subject to contrasts.

A primary contribution of this paper is to identify and measure perceptual biases across distinct settings in the field. The first domain of analysis is the evaluation of exams by graduate instructors across four semesters of a 300-student introductory economics course at UC Berkeley. The research design is ideally suited for an empirical test of contrasts—the evaluation order is known and quasi-random (i.e. based on alphabetical order), outcomes are transparent, and proxies for ability exist.

I find that graders who encounter a string of three poor (strong) exams, grade the subsequent exam question more generously (punitively). Relative to the average score, the contrast effect is on the order of a 12% increase in lenience when an exam follows a streak of three low scoring exams, and a 6% score loss when an exam follows a streak of three high scoring exams. The effects are directionally consistent, but not significant, for shorter streaks of exams. These possible errors in evaluation distort scores for an estimated 3 to 10% of students and final semester grades for an estimated 1 to 2% of students.

Next, the paper summarizes evidence for relative decision-making in judicial sentencing (Bhargava and Cann 2007). We examine the effect of relative case order within a day on sentencing decisions of approximately 500 judges in PA courtrooms from 2001 to 2005. Even though cases are quasi-randomly ordered, we find evidence that lower court judges exhibit greater lenience in the disposition of summary offenses—traffic offenses such as speeding and minor non-traffic offenses such as the use of false identification—if the judge has just been exposed to a hearing for a more serious criminal offense. Relative to typical decisions, judges are 9% more lenient in the adjudication of cases which follow a felony and are 6% more lenient in cases which follow a misdemeanor. Possibly due to the heightened salience of rare events, the effect is confined to days when a judge has heard only a single

felony or misdemeanor. Indeed, judges with low prior exposure to felonies tend to exhibit greater relativity in decisions than judges with a history of prior exposure.

Finally, in speed dating sessions, where interactions are also quasi-random, we find that a subject's decision to date is negatively influenced by the attractiveness of the prior partner even after controlling for current partner attractiveness (Bhargava and Fisman 2007).[3] For male subjects, relative to baseline acceptance rates, the contrast is on the order of a 17% increase in willingness to date after exposure to a single unattractive partner and a 13% decrease in willingness to date after exposure to a single attractive partner. The effects grow in magnitude after exposure to streaks of attractive, but not unattractive, partners. We find no effect for female subjects and discuss experimental evidence consistent with this gender asymmetry.

At first glance, alternative explanations such as learning or the existence of quotas for high or low evaluations could account for the findings. However, across all three domains, the negative correlation in assessments decays sharply. As the theoretical framework predicts, this rate of decay is inconsistent with standard explanations arising from rational preferences and updating even in the presence of constraints. Further, across all domains the observed effects do not fully diminish as decision-makers accumulate experience, and in grading and dating there is no evidence of contrast effects for highly transparent targets which do not permit interpretive discretion (i.e. multiple choice questions, and highly attractive or unattractive partners). These findings lend additional support to an explanation based on perceptual contrasts. It is possible, however, that limited memory coupled with selective recall or some other heuristic could also account for the evidence. Alternative explanations are addressed at greater length in the Discussion.

If perceptual errors in sequential decisions do exist, how pervasive and important are they? The analysis finds non-trivial distortions in decisions across all three examined environments. As an example, the effects observed in the analysis of sentencing decisions are comparable in magnitude to biases which other researchers have attributed to defendant and judge race (Steffensmeier and Britt 2001). Contrast effects are also likely to afflict a wide range of sequential decisions not explored in the present analysis. These include hiring decisions, admissions decisions of colleges, appraisal of investment opportunities, consumer evaluation of product price and quality, and medical diagnoses of patients. In some of these settings, even rare evaluative errors may carry high costs.

As a test for contrast-effects in an important consumer goods setting, the Discussion describes an original survey of several Minnesota offices of a large national real-estate firm. The survey indicates 71% of realtors believe that buyers perceive a home more favorably

[3] Attractiveness is measured on a 1 to 10 ordered scale by research assistants. Details of the experimental setup are described in Bhargava and Fisman (2007), as well as in Fisman et al. (2006).

after viewing a home which is overpriced, in a bad location, or otherwise unattractive. While rational explanations cannot be ruled out, these responses are consistent with the existence of contrast effects in the purchasing decisions of homes. Moreover, while the three empirical settings in this paper feature quasi-random ordering, it is conceivable that individuals or firms might strategically exploit awareness of contrasts in non-random settings for profit. Indeed, of realtors who believe in contrasts, 47% admit to manipulating order of home showings so as to increase the likelihood of a purchase. Realtors with higher (self-reported) sales productivity are marginally more likely to be aware of and to exploit this bias.

This research is in the spirit of other papers which have discussed how comparative assessments shape behavioral outcomes (Kahneman and Miller 1986; Koszegi and Rabin 2007). More broadly, it is linked to the literature which examines the influence of context (Tversky and Simonson 1993; Ariely et. al. 2003; Kamenica 2007) as well as biases in perception on economic decisions. The latter includes research on quasi-Bayesian updating (Tversky and Kahneman 1971; Rabin and Schrag 1999; Rabin 2002), as well as categorical biases (Quattrone and Jones 1980; Fryer and Jackson 2007). It is also a part of a small empirical literature which investigates the role of sequential context in economic decision-making (Simonson and Tversky 1992; Simonsohn 2006; Simonsohn and Loweinstein 2006).

The remainder of this paper proceeds as follows. The next section describes the background of the theoretical and empirical research on contrast effects. Section III outlines a theoretical framework which generates competing predictions for the standard and contrast-effects models. The following section presents empirical results in the domain of exam evaluation, and then summarizes empirical evidence in two additional settings of judicial decisions and romantic dating. Section V outlines alternative explanations, synthesizes the empirical results and discusses applicability to other domains. The final section discusses policy implications, identifies directions of future study, and concludes.

2 Background

During the last several years, economists have displayed increasing willingness to incorporate findings from other disciplines including psychology, neuroscience, social cognition, and sociology into models of decision-making. One element of this movement has sought to clarify the effect of cognitive biases on behavior. A cognitive bias not yet subject to extensive empirical inquiry, despite rich experimental and some theoretical treatment, is the role of relative assessments in judgment.

A central tenet of social cognition is that judgments are fundamentally comparative. The innate role of relativity in judgments is underscored by research suggesting that such

relativity improves the efficiency of cognitive processing, as well as studies arguing that it cannot be removed through deliberative effort (Friedrich et. al. 1972; Shapiro and Spence 2005; Keil et. al. 2006).

In the case of evaluations or judgments made in sequence, prior exposure to an exemplar (i.e. a specific, extreme example of a judged category), tends to serve as a reference point from which subsequent assessments of more moderate stimuli, along the same attribute dimension, are contrasted away.[4] Researchers have labeled this systematic enhancement or diminishing of an assessment due to exposure to a prior exemplar a "contrast effect."

The original class of experimental studies documenting sequential contrast effects were psychophysical in nature and demonstrates that subjects systematically overestimate or underestimate sensory dimensions (e.g. weight, length, and color) of targets after exposure to extreme examples of such targets along the same sensory dimensions (Hood 1950; Hovland et. al. 1958; Krantz and Campbell 1961). For example, in a study conducted in 1961, researchers find that 160 subjects asked to sequentially judge the length, in inches, of lines from a randomly assigned group of either shorter lines (ranging in length from 6 to 20 inches) or lengthier lines (from 20 to 36 inches) judge a reference line of 20 inches as 35% ($p < .01$) longer when in the former as compared to the latter group (Krantz and Campbell 1961). This group of studies indicates that subjects are prone to errors due to sequential context.

More recently, experimenters have found evidence for contrast effects in sequential assessments of social, as opposed to physical, stimuli. Studies of social judgment have uncovered behavior consistent with contrast effects in the inference of attitudes, assessments of emotions and happiness, judgments of fairness and criminal guilt, assessments of physical attractiveness, evaluations of athletic performances, and perceptions of the self (Damisch 2006 et. al.).

In one widely cited study, two confederates walk into male college dormitories either during or immediately prior to the airing of a popular television series, *Charlie's Angels*, featuring three attractive females. The 81 male students are then asked to rate the attractiveness of a photograph of an average looking female on an ordered scale from 1 to 7. Subjects watching the television program rate the photograph 17% ($p < .03$) lower than subjects who had been watching another program earlier in the night. The findings are robust to comparison to control groups who had not been watching television at the time the show aired, who had not been watching television in the period prior to airing, as well as subjects who had been watching a different television program during the period of airing (Kenrick and Gutierres 1980). Table 1 compares contrast effect sizes found in the

[4] The reverse of such a phenomenon is labeled an "assimilation effect." Numerous studies have attempted to identify the distinguishing features of the situations and stimuli which prompt assimilations as opposed to contrasts in judgment.

lab and the field for the domains considered in the present research.

Studies of social assessments, however, tend not to implicate the existence of a cognitive *error* as convincingly as much of the psychophysical research. In the above examples, subjects were not incentivized to provide accurate responses. Further, while Krantz and Campbell instruct subjects to report measurements with a well known metric (i.e. inches), subjects in social perception research are typically asked to use unfamiliar numerical scales. In studies which rely on uncommonly defined rating scales, the observed effects may be due not to contrasts in perception, but rather to semantic contrasts in interpretation of the unit of measure.

Table 1

COMPARISON OF CONTRAST EFFECTS IN THE LAB AND THE FIELD

		CONTRAST EFFECT ESTIMATE (Distortion in Outcome Variable Due to Contrast)	
DOMAIN	PRIME	EXPERIMENTAL RESULT	FIELD RESULT
Attractiveness	Television Show	N = 81, 0.17 (Kenrick and Gutierres, Study 1, 1980)	X
	Photographs of Models	N = 48, 0.25 (Kenrick and Gutierres, Study 2, 1980)	X
	Partners in Speed Dating	X	N = 3592, .13 to .17 (Bhargava and Fisman 2007)
Exam Evaluation	Written Exam	X	N = 5172, .06 to .12 (Present Analysis)
Judicial Sentencing	Criminal Case	N = 96, 0.38 (Pepitone and DiNubile, 1976)	N = 1.2 million, .09 (Bhargava and Cann 2007)

Notes: Experimental effect sizes are calculated as the change in the outcome variable in the contrast case relative to controls. In Kenrick and Gutierres, Study 1, the prime (or trigger for the sequential contrast) is the television show, Charlie's Angels, while the outcome variable is a rating of attractiveness on an ordered scale from 1 to 7 of a female photograph. The prime in Study 2 is a female model in an advertisement and the outcome variable is a rating on the same ordered scale of a yearbook photo of a female. The effect in the Pepitone and DiNubile study is the change in the recommended prison sentence for a hypothetical defendent found guilty of assault, preceded by a description of a homicide, as compared to the recommended sentence for a hypothetical defendent found guilty of assault preceded by a description of another assault. Subjects are male across all of the studies. The field results are distortions relative to baseline decisions (i.e. decision to date in dating, average score in grading, and lenience in judicial sentencing). The primes in each of the field examples are single exemplars with the exception of the grading domain where effects are prompted by sequences of three high or low exams.

Only a handful of studies have examined contrast effects in the field. Simonsohn and Loewenstein (2006), find that newly transplanted residents choose to live in apartments whose rent is linked to the average housing prices of their original city of residence. In a second study, Simonsohn (2006), using Panel Study of Income Dynamics data, finds that individuals who move from one city of residence to another tend to select office commuting times tied to the typical commuting times of the city of origin. Both studies demonstrate a

negative relation between a current preference and a prior standard which appears to serve as a reference point. The key difference between this research and the present analysis is that the current paper focuses on contrast effects in sequential, high frequency, decisions. Further, while the authors attempt to control for confounds introduced by selection, the studies feature sequential choices whose order is not quasi-random.

3 A Model for Sequential Decisions

I describe a sequential decision-making environment with a simple model of signal extraction. For ease of exposition, I formalize the description in the specific context of exam evaluations where a grader assesses a series of exam questions in succession.[5] I first outline a standard model which features a rational decision-maker who cares only about the accuracy of evaluations, and forms and update his beliefs about the quality of each target with Bayes' Rule. A more plausible set of preferences which allows for constraints due to quotas for high or low evaluations is addressed in the section on model extensions and then more fully in the Appendix. The Appendix translates the case of sequential decision-making subject to quotas as a finite dynamic programming problem and outlines the comparative statics of interest.

I then introduce a contrast-effects model with a decision-maker subject to relative perception. This decision-maker misperceives signals, but is unaware of this error, and Bayesian updates as if she were not making such an error.

The model features three variables of central importance: (i) the intrinsic quality of the target, (ii) the perceived quality of the target, and (iii) the evaluation which the decision-maker assigns to the target. The critical comparative statics which emerge from the model describe the influence which a change in these variables in one period have on the perception and evaluation of a target in a subsequent period.

The transparency of these variables to the econometrician differs across the three empirical contexts. In the context of exam evaluation, one observes only the grade (evaluation) assigned to an exam. In the data on judicial decisions, the severity (intrinsic quality) of a crime is observed, as well as the judge's sentence (evaluation). Finally in the setting of romantic daters, all three variables—attractiveness (intrinsic quality), perceived attractiveness (perceived quality), and a final "yes/no" decision to date (evaluation)—are observed.

3.1 The Standard Model

Assume that each exam question to be graded is of some intrinsic quality q. This quality, q, of an exam can be decomposed into two constituent parts. The first, s, is a student-invariant

[5] Assume that the decision-maker can not reevaluate a past exam.

component attributable to the quality of the instruction (or instructor). Alternatively, s can be interpreted as the mean quality level across all targets.[6] The value of s is shared by all students and is drawn from a distribution $N(s_0, 1/\gamma)$ with precision γ. A second component of an exam's quality is a student-specific element, ψ_t, which is independent and identically distributed, and is drawn from the distribution $N(0, 1/\lambda_\psi)$ with precision λ_ψ. The idiosyncratic component can be attributed to the effort, disposition, or perhaps even luck of each student t, and is what the grader ultimately seeks to map to a grade.

$$q_t = \psi_t + s$$

It is useful to think of grading as occurring in three steps: (1) an initial assessment of quality, $q \rightarrow \widetilde{q}$, (2) an inference of the idiosyncratic portion of the exam, $\widetilde{q} \rightarrow \hat{\psi}_t = E(\psi_t \mid \tilde{q}_1, ..., \tilde{q}_t)$ based on present and past assessments of exam quality, and finally (3) an assignment of a grade, $\hat{\psi}_t \rightarrow g$. The perceived quality of the exam can be interpreted as a signal $\widetilde{q}$ of the underlying intrinsic quality. Imagine that the accuracy with which a decision-maker extracts each signal is a function of expended effort (or attention) e_t. Given infinite effort, a grader could conceivably observe an answer's true quality such that $q_t = \widetilde{q}_t$. However, assume that expenditure of effort in each period is costly, and since the main predictions of the model are not sensitive to the allocation of effort, further assume that effort is fixed.

The perceived quality of an exam answer $\widetilde{q}_t$ can then be indicated by:

$$\widetilde{q}_t = q_t + \varepsilon_t$$

where $\varepsilon_t \sim N(0, 1/\lambda_\varepsilon)$ and precision $\lambda_\varepsilon = \eta(e_t)$ describes the noise introduced to the perceived signal.

Once a grader assesses an exam's overall quality, $\widetilde{q}_t$, she can then infer the idiosyncratic component of quality, $\hat{\psi}_t$. The estimate of $\hat{\psi}_t$ crucially depends on the grader's inference, $\hat{s}_t = E_t(s \mid \tilde{q}_1, ..., \tilde{q}_t)$, of the systematic component of each score:

$$\hat{\psi}_t = E(\psi_t \mid \tilde{q}_1, ..., \tilde{q}_t) \ = \tilde{q}_t - \hat{s}_t$$

Because the grader learns about this systematic component, s, as she observes each exam,

[6] In the context of judicial decisions, one could imagine s to be the average severity of criminal infractions, within a particular category of charges. Similarly, in the context of speed dating, s could be interpreted as the average attractiveness of partners in a session.

one can express the grader's updated estimate of $\hat{s}_t$ through Bayes' rule:

$$\hat{s}_t = \frac{f(\ \tilde{q}_1, ..., \tilde{q}_t|\ s)f(s)}{\sum_{\tilde{q}\in\tilde{Q}} f(\tilde{q}_1, ..., \tilde{q}_t|\ s)f(s)}$$

where the prior distribution of s is given by $N(s_0, 1/\gamma)$. Recognizing that ψ_t is distributed normally, after grading t exams, the above expression implies that the grader estimate of $\hat{s}_t$, is a convex combination of the prior belief of the mean, s_0, and the mean quality of observed exams $\overline{\tilde{q}_t} = (\Sigma\tilde{q}_t)/t$[7]:

$$\hat{s}_t = \frac{\gamma}{\gamma + \lambda t} s_0 + \frac{\lambda t}{\gamma + \lambda t} \overline{\tilde{q}_t}$$

where $1/\lambda = 1/\lambda_\psi + 1/\lambda_\varepsilon$. It can be seen from the above expression, that, for fixed values of γ and λ, as more exams are graded, a grader's best estimate of the systematic component of each grade is increasingly weighted by the mean of observed signals, $\overline{\tilde{q}_t}$, as compared to the prior belief of the mean s_0. Further, for fixed values of γ and t, the higher the precision, λ, of the observed scores, the more relative weight is assigned to the mean of perceived signals. Finally, for fixed values of λ and t, the higher the precision, γ, of the decision-maker's prior distribution, the more relative weight is assigned to the prior belief of the mean.

The grader relies on this inference of $\hat{s}$ in order to back out an estimate of the idiosyncratic component of exam quality:

$$\hat{\psi}_t = E(\psi_t \mid \tilde{q}_1, ..., \tilde{q}_t) = \tilde{q}_t - [\frac{\gamma}{\gamma + \lambda t} s_0 + \frac{\lambda t}{\gamma + \lambda t} \overline{\tilde{q}_t}]$$

Once a grader estimates the idiosyncratic grading component, she then maps this estimate to a grade, $\hat{\psi}_t \to g$. In order to understand the assignment of a grade, one must first specify the preferences of the decision-maker. Consider a decision-maker who has simple preferences described by the following utility function, again treating effort as fixed across each period:

$$U = f(\mathbf{g}, \boldsymbol{\psi}) = -\sum_{j=0}^{t} (g_j - \psi_j)^2$$

The utility function describes a grader motivated by the desire for accuracy as indicated through the minimization of $\sum_{j=0}^{t}(g_j - \psi_j)^2$. Here accuracy in a given period is the squared distance between a grade and the unobserved idiosyncratic exam quality. Utility is decreasing in the inaccuracy of grading.[8]

[7] Note that $E(\tilde{q}_t) = E(s_t)$ since $\psi \tilde{} N(0, \lambda_\psi)$. The full derivation is provided in the appendix.

[8] It is worthwhile to note that the utility as specified doesn't account for the possibility that a grader may be pulled to both minimize deviations of evaluations from idiosyncratic scores, as well as to minimize the difference in such deviations across similar students. For example, a grader who misgrades an early

The Bayesian then solves the problem: $\max_{\{g\}} -\sum_{j=0}^{t}(g_j - \psi_j)^2$ and assigns a series of grades $g_1, ..., g_t$ in order to maximize total utility. The result of the maximization is to set $g_j^* = \hat{\psi}_j$ for all j. Proposition 0 describes the relationship between contemporaneous evaluations, perceived quality, and intrinsic quality. Derivations of all proofs can be found in the appendix.

Proposition 0 (Contemporaneous Effects) *For all t, (i) an increase in the intrinsic quality of an exam leads to an increase in the perceived quality for that same exam: $\partial\tilde{q}_t/\partial q_t = 1$, (ii) an increase in the intrinsic quality of an exam leads to a higher evaluation for that same exam: $\partial g_t/\partial q_t > 0$, and (iii) an increase in the perceived quality of an exam leads to a higher evaluation for that same exam: $\partial g_t/\partial\tilde{q}_t > 0$.*

Part (i) of this proposition confirms the basic intuition that an increase in the underlying quality of an exam in a given period translates into an increase in the perceived quality of the same exam. Part (ii) of the proposition notes that an increase in intrinsic quality should lead to a higher grade for the exam. This occurs because an increase in the observed score leads to an increase in perceived score, which in turn leads to an increase in the estimated idiosyncractic score for that period. This increase more than offsets the reduction in estimated score induced through the rise in estimated $\hat{s}_t$. A higher estimate of the idiosyncratic score maps to a higher evaluation. The final claim of the proposition follows from the reasoning underlying the claim in part (ii).

The broader motivation of the model is to highlight situations in which past perceptions and evaluations influence subsequent evaluations. In the case of the standard model, this situation arises as a consequence of a grader updating her estimates of the systematic score $\hat{s}$. Proposition 1 and Corollary 1 describe the relationship between past exams and current evaluations (i.e. the antecedent effect).

Proposition 1 (Bayesian Antecedence - Quality) *(i) An increase in the intrinsic quality of an exam leads to a lower evaluation of the subsequent exam: $\partial g_{t-k}/\partial q_{t-k-1} < 0$. The magnitude of this negative influence is monotonically decreasing and converges to 0 as the number of evaluated exams approaches ∞: $\partial g_{t-l}/\partial q_{t-l-1} < \partial g_{t-k}/\partial q_{t-k-1} < 0$ for $0 < k < l$ and $\lim_{t\to\infty} \partial g_t/\partial q_{t-1} = 0$,* and (ii) *the influence of the intrinsic quality of an exam on the evaluation of a subsequent exam is not a function of the distance between the exams: $\partial g_t/\partial q_{t-k} = \partial g_t/\partial q_{t-l}$ for all k, l.*

exam, due to ignorance of s, may refrain from accurately grading a later exam of identical quality in order to assign the students similar grades. This is one consequence of the grader's assumed inability to go back and regrade. A utility function which better captures this intuition for a decision maker i and exam j is given by: $U = -\sum_{i=0}^{t}\sum_{j=0}^{t}[(g_i - \psi_i) - (g_j - \psi_j)]^2$. The key comparative statics should hold under such a utility function.

The intuition for Proposition 1 is straightforward. Part (i) suggests that, for the rational decision-maker, the evaluation of an exam is sensitive to the quality of the prior exam due to the influence which the prior exam has on the estimation of the systematic score. A rise in a past score raises the estimate of $\hat{s}_t$ which then depresses the estimate of the idiosyncratic score $\hat{\psi}_t$. While the estimation of the systematic score is shaped by perceived quality, any change in intrinsic quality can be translated to an equivalent change in perceived quality. Because the estimation of $\hat{s}_t$ is based on Bayes' Rule, as the number of graded exams increases and the decision-maker learns the shape of the underlying distribution, the relative weight on any given exam score monotonically declines. Part (ii) of the proposition recognizes that since the estimation of $\hat{s}_t$ relies on a precision-weighted average of the prior of s_0 and the mean of observed scores $\overline{\tilde{q}_t}$, each contributing past perception $\tilde{q}_j$ $(j < t)$ has an identical influence on $\hat{s}_t$, as well as the estimate of $\hat{\psi}_t$. Importantly, because $\partial \tilde{q}_t / \partial q_t = 1$, the comparative statics with respect to $\tilde{q}_t$ also hold for q_t.

Comparative statics with respect to intrinsic quality are of limited predictive usefulness for contexts where such quality is not observed. In sequential contexts in which only evaluations are observed, one might prefer comparative statics with respect to observable alternatives. Accordingly, Corollary 1 outlines the effects which changes in prior evaluations have on subsequent evaluations.

Corollary 1 (Bayesian Antecedence - Evaluations) *(i) An increase in the evaluation of an exam leads to a lower evaluation of the subsequent exam holding other past evaluations constant. The magnitude of this negative influence is monotonically decreasing and converges to* 0 *as the number of evaluated exams approaches* ∞*:* $\frac{\partial g_{t-l}}{\partial g_{t-l-1}}|_{g_{t-i-1,i>l}} < \frac{\partial g_{t-k}}{\partial g_{t-k-1}}|_{g_{t-i-1,i>k}} < 0$ *for* $k < l$*, and* $\lim_{t\to\infty} \frac{\partial g_t}{\partial g_{t-1}}|_{g_{t-i,i>1}} = 0$*, and (ii) the influence of the evaluation of an exam on the evaluation of a subsequent exam, holding other past evaluations constant, is not a function of the distance between the exams:* $\frac{\partial g_t}{\partial g_{t-k}}|_{g_{t-i,i>k}} = \frac{\partial g_t}{\partial g_{t-l}}|_{g_{t-i,i>l}}$ *for all* k*,* l*.*

Note that the analogy between Proposition 1 and Corollary 1 critically relies on whether the change in a prior evaluation can be isolated to a change in that period's underlying exam quality. If evaluations in other past periods are held constant, then the lone channel through which one grade is able to influence a subsequent grade is if there is a shift in the perceived quality of the prior period exam. Given this "constancy condition," the intuition underlying Corollary 1, mirrors the intuition discussed in Proposition 1.

3.2 The Contrast-Effects Model

The standard model features a decision-maker whose perception is free of systematic error and who is subject to simple preferences over accuracy. Such a decision-maker may exhibit

antecedent effects, but only in evaluations of early, and not later, exams.

Next I turn to model of contrast effects in perception. A decision-maker in such a model follows the same three stages of evaluation outlined above. However, errors in perception, due to contrast effects, appear in the assessment stage, $q \rightarrow \tilde{q}^c$, where $\tilde{q}^c$ is the perceived score in the presence of a contrast. Specifically, $\tilde{q}_t^c$ depends on the true signal of the contemporaneous exam, $\tilde{q}_t$, as well as a function $\ell(.)$ of imperfect perception of prior signals $\tilde{q}_{t-1}^c, \tilde{q}_{t-2}^c, \tilde{q}_{t-3}^c ...$ and the transparency of a question α:[9]

$$\tilde{q}_t^c = f(\alpha, \tilde{q}_t, \tilde{q}_{t-1}^c, \tilde{q}_{t-2}^c, ..., \tilde{q}_1^c) = q_t + \ell(\alpha, \tilde{q}_{t-1}^c, \tilde{q}_{t-2}^c, ..., \tilde{q}_1^c) + \varepsilon_t$$

where $-1 < \partial\ell/\partial\tilde{q}_{t-k}^c < 0$ and $\partial\ell/\partial\tilde{q}_t \geq 0$ for $k > 0$ captures an error in perception consistent with the notion of a sequential contrast effect. The function ℓ dictates that a current signal is perceived in contrast to the preceding signal. A more positive prior signal more negatively biases the perception of the subsequent signal. The transparency of a target, $\alpha \in [0, 1]$, denotes the ease with which targets can be accurately perceived where $\alpha = 1$ indicates complete transparency.

The following linear formulation for $\ell(.)$ captures this negative relationship between present and past perceptions:

$$\tilde{q}_t^c = q_t - (1 - \alpha)[\delta_1 \tilde{q}_{t-1}^c + \delta_2 \tilde{q}_{t-2}^c + \delta_3 \tilde{q}_{t-3}^c + ...] + \varepsilon_t$$

A condition of exponential decay, $\delta_k = \delta_1^k$ and $0 < \delta_k < 1$, permits the perception $\tilde{q}_t^c$ to be expressed as a negative function of only the last exam $\tilde{q}_{t-1}^c$. To understand the intuition for this, note that $0 < \delta_k < 1$ implies that a change in the quality of one exam negatively influences the perceived quality of the subsequent exam. This distortion in the subsequent perception then negatively influences the perceived quality of the next exam, and so on. Without the imposition of additional restrictions, the original agitation alternates from negative to positive influence on a future exam as it exponentially decays across periods. The condition, $\delta_k = \delta_1^k$, ensures that, excluding the last period, the direct influence of a past period on the present period offsets any indirect influence of that exam as earlier contrasts decay. Conceivably, one could impose additional restrictions so that more distant past periods might additively, and negatively, influence $\tilde{q}_t^c$. This existence condition for streaks of exams is discussed in the model extensions.

Consistent with psychological theory, the magnitude of the perceptual error is related to the degree of interpretive discretion permitted in the evaluation (Higgins 1996). A fully transparent set of evaluations, $\alpha = 1$ leads to no contrast effect, $\tilde{q}_t^c = q_t + \varepsilon_t$. A complete

[9] One could alternatively express ℓ as a function of $(\tilde{q}_{t-1}, \tilde{q}_{t-2}, \tilde{q}_{t-3} ...)$ as opposed to $(\tilde{q}_{t-1}^c, \tilde{q}_{t-2}^c, \tilde{q}_{t-3}^c, ...)$ without changing any of the key comparative statics. However, the former formulation seems less congruous with the underlying psychology.

lack of transparency, $\alpha = 0$, leads to a full contrast $\tilde{q}_t^c = q_t + \ell(.) + \varepsilon_t$.

Critically, while the decision-maker has accurate beliefs over the distribution of ψ_i and priors identical to his rational counterpart with respect to s, the decision-maker believes the perceived signal is entirely attributable to the contemporaneous exam—that is, she believes that $\widetilde{q}_t^c = \widetilde{q}_t$. The decision-maker therefore infers the systematic component of the score by calculating $\hat{s}_t^c = E_t(s \mid \tilde{q}_1^c, ..., \tilde{q}_t^c)$:

$$\hat{s}_t^c = \frac{\gamma}{\gamma + \lambda t} s_0 + \frac{\lambda t}{\gamma + \lambda t} \overline{\tilde{q}_t^c}$$

It follows that the idiosyncratic score of each student can be estimated by $\hat{\psi}_t^c = \tilde{q}_t^c - \hat{s}_t^c$.

With this estimate in hand, the decision-maker subject to perceptual errors chooses a series of grades $g_1^c, ..., g_t^c$ to solve the following problem: $\max_{\{g^c\}} - \sum_{j=0}^{t} (g_j^c - \psi_j)^2$. The result of the maximization is to set $g_j^{c*} = \hat{\psi}_j^c$ for all j. The interpretation of the contrast-effects model with respect to the relationship between intrinsic quality, perceived quality and assigned grades in the contemporaneous period is analogous that described in Propositon 0.

A number of more useful comparative statics describe the relationship between intrinsic quality and evaluations across periods. The central prediction is that a contrast effect results in a negative relationship between a prior period's quality and the present evaluation. This relationship does not disappear as the number of evaluated exams increases, and decays as the intervening distance between the periods grows. Proposition 2 formalizes the claim.

Proposition 2 (Contrast Effects - Quality) *(i) For all t, and $\alpha < 1$, the perceived quality of an exam is negatively related to the intrinsic quality of the prior exam: $\partial \tilde{q}_t^c / \partial q_{t-1}^c < 0$, (ii) for all t, and $\alpha < 1$, the influence of the intrinsic quality of an exam on the subsequent evaluation is negative: $\partial g_t^c / \partial q_{t-1} < 0$, (iii) for all t, and $\alpha < 1$, the influence of the intrinsic quality of an exam on a subsequent evaluation is negatively related to the distance between the exams: $\left| \frac{\partial g_t^c}{\partial q_{t-k}} \right| \geq \left| \frac{\partial g_t^c}{\partial q_{t-l}} \right|$ for all $0 < k < l$, and (iv) the contrast effect with respect to perception and evaluation fades as exams approach full transparency: $\lim_{\alpha \to 1} \partial \tilde{q}_t^c / \partial q_{t-1}^c = 0$, and for large t, $\lim_{\alpha \to 1, t \to \infty} \partial g_t^c / \partial q_{t-1} < 0$.*

While Part (i) of the proposition follows directly from definition of the contrast effect, a simple decomposition helps to illuminate the intuition for the remaining claims. For the decision-maker subject to contrasts, assuming evaluations are not fully transparent, a past exam influences a current exam through two channels. The first is the exam's influence on the perceived quality of a subsequent exam as defined by the function ℓ. The second is the exam's contribution to learned estimates of the systematic score.

Now the intuition for Part (ii) of the proposition is straightforward. The quality of the immediately prior exam negatively influences the current perception, and leads to a downward estimate of subsequent exams due to an increase in the estimate of the systematic score $\hat{s}^c$. The two effects coincide to produce a negative relationship across periods.

Part (iii) of the proposition notes that the intrinsic quality of a proximal past exam—specifically the exam at period $(t-1)$—has greater influence on the current period evaluation than the intrinsic quality of a more distant past exam. The intuition for this claim is analogous to that outlined for Part (ii). Present and past exams are negatively related due to both learning—as well as in the case of the last period's exam—contrast effects. Since the contrast effect fully decays after one period, however, an immediately prior exam exerts a stronger influence than other past exams. The final part of the proposition speaks to the dilution of the contrast effect as the transparency of an exam increases.

As before, it is useful to recast the comparative statics above in terms of an observed variable. Corrolary 2 restates the comparative statics with respect to changes in evaluations rather than intrinsic quality.

Corollary 2 (Contrast Effects – Evaluations) *(i) For all t, the influence of an evaluation on the subsequent evaluation, holding other past evaluations constant, is negative:* $\frac{\partial g_t^c}{\partial g_{t-1}^c}|_{g_{t-i,i>1}^c} < 0$, *(ii) for all t, and $\alpha < 1$, the influence of an evaluation on a subsequent evaluation, holding other past evaluations constant, is negatively related to the distance between the exams:* $\left|\frac{\partial g_t^c}{\partial g_{t-k}^c}\right|_{g_{t-i,i>k}^c} \geq \left|\frac{\partial g_t^c}{\partial g_{t-l}^c}\right|_{g_{t-i,i>l}^c}$ *for $0 < k < l$, and (iii) for large t, the influence of an evaluation on the subsequent evaluation, holding other past evaluations constant, fades as exams approach full transparency:* $\lim_{\alpha \to 1} \frac{\partial g_t^c}{\partial g_{t-1}^c}|_{g_{t-i,i>1}^c} = 0$.

The intuition for the link between Proposition 2 and Corollary 2 follows from the same assumption of constancy in evaluations which characterized Corollary 1. When other, past evaluations are held constant, then only a perturbation in contemporaneous quality can generate a change in an evaluation.

3.3 Extensions

Quota Constraints for High or Low Evaluations. It is plausible that a decision-maker is constrained in the number of high or low evaluations which she can assign. Consider for example a grader interested in limiting the number of As so as to be able to create variation through which to distinguish students. This grader may desire to limit the number of high grades independent of any concern for accuracy. The Appendix adapts the standard model to characterize such a decision-maker with a more realistic set of preferences that reflects possible evaluative constraints. The situation is modeled as a dynamic programming

problem where a decision-maker assigns each grade sequentially by adhering to a policy rule which is a function of the estimated idiosyncratic score in each period and a state variable which is equal to the number of high grades assigned. Such preferences might readily describe decisions in broader contexts such as judicial sentencing or dating.

A first comparative static that emerges from the model is that evaluations are (weakly) negatively correlated across periods. This negative relationship arises either from the standard story of learning documented above, or from the presence of the quota constraint. In the event that such a constraint is not binding, the grader assigns a high grade only if it is sufficiently above the threshold so as to warrant the loss of freedom to assign a future high grade less the current period penalty due to loss in accuracy. The functional threshold above which a score must reach in order for the grader to assign the high grade in a given period is a positive function of the number of previously assigned high grades. A high grade in the last period consequently prompts a subsequently lower current period evaluation. Once the constraint binds, then there may be zero correlation across periods.

The key intuition produced from the dynamic programming exercise is that the prediction of decay (Prediction 3) from the contrast-effects model still serves to differentiate rational from non-standard behavior. This is true because while the decision rule of the grader may be a function of the state variable—the number of past high grades—and the estimated idiosyncratic score, it is invariant to the order in which such grades were earned. Because graders do not differentially treat proximal and more distant exams, even in the presence of quotas, the influence of a particular evaluation on a future evaluation should not be a function of the intervening distance between the periods.

An Existence Condition for Streaks. The lag structure, $\delta_k = \delta_1^k$, ensures that the perception of the current exam score is subject to a contrast effect only through the exam in the last period. However, one could specify a lag structure which would allow for additive effects across exams over several past periods. Such a pattern of parameters would permit stronger contrast effects after a sequence of high or low exams as compared to contrasts triggered by a single high or low exam. The possibility that sequences of high or low targets are important for triggering measurable contrasts is critical to the empirical analysis presented below.

To see that such a lag structure exists, first assume, without loss of generality, that $\alpha = 0$. Note that one can exploit the recursive nature of $\tilde{q}_t^c$ to rewrite $\{\tilde{q}_t^c = \tilde{q}_t - \delta_1\tilde{q}_{t-1} - (\delta_2 - \delta_1^2)\tilde{q}_{t-2} - (\delta_3 - 2\delta_1\delta_2 + \delta_1^3)\tilde{q}_{t-3} - (\delta_4 - 2\delta_1\delta_3 + 3\delta_1^2\delta_2 - \delta_2^2 - \delta_1^4)\tilde{q}_{t-4} - ...\}$. This expanded expression illustrates that so long as $\delta_1 > 0$, one can define each subsequent δ_k to produce a positive net coefficient value for each $\tilde{q}_{t-k}$. One can iteratively set $\delta_2 > \delta_2^2$, $\delta_3 > 2\delta_1\delta_2 - \delta_1^3$, and so forth for each δ_k to ensure that the net contrast effect with respect to past exams is additive across periods.

4 Empirical Evidence

Examples of sequential decision-making in the field are not difficult to find. However, examples which allow for clean empirical identification are more rare. This section evaluates three distinct environments which both involve decisions made in sequence, and are amenable to empirical scrutiny. A first and primary focus of the analysis tests for contrast effects in the evaluation of student exams from a large undergraduate course. Additionally, evidence is summarized from two other domains. These include non-criminal sentencing decisions of judges in Pennsylvania courts and assessments of physical attractiveness in a structured dating environment (Bhargava and Cann 2007; Bhargava and Fisman 2007).

As an organizing framework for the empirical analysis, five sets of testable empirical predictions—emerging from the theoretical model—are enumerated below. Broadly, the empirical strategy consists of demonstrating that prior targets negatively influence subsequent evaluations (Prediction 2), and then ruling out possible alternative explanations which may have emerged from a set of rational preferences and Bayesian updating (Predictions 3 and 4). Prediction 5 notes that questions which allow for little interpretive discretion should elicit no contrast effects. This serves as a placebo of sorts. The predictions concerning contemporaneous period behavior (Prediction 1) serve to corroborate the basic assumptions of the model. These predictions are outlined below:

Prediction 1: Contemporaneous Effect (S & CE).
An increase in the intrinsic quality of a target leads to a higher evaluation for the same target ($\frac{\partial g_t}{\partial q_t} > 0$; $\frac{\partial g_t^c}{\partial q_t} > 0$).

Prediction 2: Antecedent Effect (S & CE).
An increase in the intrinsic quality (or evaluation) of a target leads to a lower evaluation of the subsequent target ($\frac{\partial g_t}{\partial q_{t-1}} < 0, \frac{\partial g_t}{\partial g_{t-1}} < 0; \frac{\partial g_t^c}{\partial q_{t-1}} < 0, \frac{\partial g_t^c}{\partial g_{t-1}^c} < 0$).

Prediction 3: Contrast Effect - Decay (CE).
The influence of the intrinsic quality (or evaluation) of a target on a subsequent target is negatively related to the intervening distance between the exams.
($\left|\frac{\partial g_t^c}{\partial q_{t-k}}\right| > \left|\frac{\partial g_t^c}{\partial q_{t-l}}\right|$; $\left|\frac{\partial g_t^c}{\partial g_{t-k}^c}\right|_{g_{t-i,i\,>\,k}^c} > \left|\frac{\partial g_t^c}{\partial g_{t-l}^c}\right|_{g_{t-i,i\,>\,l}^c}$ *for* $0 < k < l$).

Prediction 4: Contrast Effect - Experience (CE).
The antecedent effect does not diminish as more exams are evaluated ($\lim_{t\to\infty} \frac{\partial g_t^c}{\partial q_{t-1}} \neq 0$, $\lim_{t\to\infty} \frac{\partial g_t^c}{\partial g_{t-1}^c}|_{g_{t-i,i\,>\,1}^c} \neq 0$).

Prediction 5: Contrast Effect - Transparency (CE).

The contrast effect disappears for questions which do not allow for interpretive discretion ($\lim_{\alpha \to 1, t \to \infty} \frac{\partial g_t^c}{\partial g_{t-1}^c}|_{g_{t-i,i>1}^c} = 0$).

Identification Strategy and Assumptions. The aim of the empirical analysis is to estimate the causal effect which past evaluations, or perceptions, have on subsequent decisions and to establish whether such an effect is consistent with the features of a contrast effect reflected by the model. The basic empirical strategy involves a dynamic linear panel data model of the following generic form:

$$y_{i,t} = \alpha + \beta x_{i,t-1} + \gamma x_{i,t} + \eta_i + \varepsilon_{i,t}$$

where i indexes a decision-maker, and t indexes the evaluation period for $t = \{1, ..., t\}$. The outcome variable $y_{i,t}$ refers to the period specific evaluation, $x_{i,t}$ is a vector of contemporaneous explanatory variables, and η_i represents fixed effects meant to account for time-constant, person specific variation. $x_{i,t-1}$ represents a past period covariate that triggers a contrast effect. As such, it can indicate a past period evaluation, in which case $x_{i,t-1} = y_{i,t-1}$. β is the parameter of interest and indicates the marginal effect, $\partial y_{i,t}/\partial x_{i,t-1}$, of the prior period covariate on the contemporaneous evaluation. $\varepsilon_{i,t}$ is an idiosyncratic error term. The presence of decision-maker fixed effects, η_i, ensures that the effect is estimated *within* rather than across individuals.

The condition which permits the identification of β is that of sequential exogeneity conditional on η_i:

$$E(\varepsilon_{i,t} |\ x_{i,1}, x_{i,2} ..., x_{i,t}, \eta_i) = 0 \ for \ t = \{1, ..., t\}$$

The contemporaneous error is uncorrelated with past or present covariates. The presence of a lagged dependent variable or covariate as an explanatory variable, as well as the assumption of correlation between the time-invariant η_i with the time-varying $x_{i,t}$, precludes a stronger assumption of exogeneity. The weaker assumption of exogeneity above is consistent with features of the empirical data. In each of the contexts, the assumption of the conditional randomness of evaluated targets is defended, and when possible, demonstrated.

With this weak assumption of exogeneity, the parameter of interest β can be estimated with a fixed effects estimator through a pooled OLS regression. This empirical strategy is implemented in the domain of exam evaluation below.

4.1 Exam Evaluation

Research Design. The primary focus of the empirical analysis involves a set of exams evaluated by graduate student instructors during four semesters of an introductory eco-

nomics course at Berkeley from 2002 to 2005. The data includes scores for each question of the final exam, composite scores for 4 to 6 problem sets (PSs) and 2 midterm exams, as well as a final grade for each of the 275 to 300 students enrolled per semester. A defining feature of the research design is that prior to evaluation, the final exams are sorted alphabetically, separated into the 6 to 9 constituent questions, and then handed off to the five student instructors who are each responsible for evaluating 1-2 stacks of ordered questions. The final exam questions consist of short answer analytic problems, brief essays, and a series of multiple choice questions (MCQs).

The data constitutes an ideal environment from which to test for contrast effects. Evaluations are observed and are made sequentially in a (presumably) known, quasi-random, alphabetical order. If any graders do grade in reverse alphabetical order, the estimated effects would be downward biased. The data also permits explicit controls for student ability through non-systematically ordered or deterministically graded scoring components such as problem sets and multiple choice questions. Finally, the variation in question type allows for tests of transparency. Appendix Table 1 summarizes key elements of the data.

Random Ordering. Ensuring that the exam quality is conditionally random is important so as to avoid the possibility that unobserved heterogeneity might confound estimates of antecedence or contrasts. For example, one could imagine positive autocorrelation in ability across alphabetical order due to ethnic clustering of surnames. While the presence of positive autocorrelation would downward bias estimates of contrasts, negative autocorrelation in ability could generate spurious evidence for the effect of interest.

Accordingly, this data lends itself to a natural control, problem sets, which one can exploit as a test for random ordering. Problem sets, unlike final exams, are (presumably) not alphabetized before grading, and are evaluated by a collection of student instructors—determined by a student's assigned section—as opposed to a single instructor. While ruling out systematic patterns across alphabetical order in problem set scores does not ensure that ability is necessarily random, it does suggest such randomness.

A formal test for randomness in observable quality across alphabetical order takes the following form:

$$PS\ Score_{s,t} = \alpha + \sum \beta_i PS\ Score_{s,t-i} + \eta_s + \varepsilon_{s,t}$$

where $PS\ Score_{s,t}$ indicates the PS score in semester s for student t where students within a semester are ordered alphabetically. Fixed effects control for semester specific variation. The specification is meant to replicate later analysis on final exam scores. The later analysis, however, relies in part on a restricted sample as well as more restrictive definitions of exemplars. Due to the small sample size permitted by a single data point for each student, an analysis of PS scores cannot completely parallel the analysis of exams.[10]

[10]Estimating the same model for the restricted sample can be done for the lag specification, as well as

Table 2

TEST OF RANDOMNESS IN GRADING ORDER

	DEPENDENT VARIABLE - PROBLEM SET QUESTIONS (OLS)								
		FULL SAMPLE							
	LAGS	LOW EXEMPLAR		HIGH EXEMPLAR		LOW EXEMPLAR		HIGH EXEMPLAR	
		Two Exams				Three Exams			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
PS Score - Lag 1	-0.024 (0.030)								
PS Score - Lag 2	-0.001 (0.030)								
PS Score - Lag 3	0.002 (0.030)								
Lag PS Score < 25%		0.491 (0.910)	0.437 (0.910)			-0.227 (1.930)	-0.311 (1.930)		
Lag PS Score < 35%			-0.849 (0.860)				-1.630 (1.280)		
Lag PS Score > 80%				-1.004 (1.000)	-1.052 (1.010)			0.643 (3.660)	0.661 (3.660)
Lag PS Score > 70%					-1.139 (0.910)				1.594 (1.290)
Treatment Size		65	134	53	102	16	55	6	21
N	N = 1127	N = 1131	N = 1131	N = 1131	N = 1131	N = 1127	N = 1127	N = 1127	N = 1127
R^2	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01

Notes: The dependent variable is the total score across all problem sets in the semester for each student. The first column presents results of an OLS regression of problem set scores on lagged scores for the prior three students by alphabetical order. The next four columns provide results of an OLS regression of problem set scores on a dummy variable indicating that the prior two scores are both either low or high exemplars. Low exemplars are scores falling in the left tail of the distribution for the semester, while high exemplars are exams falling in the right tail of the distribution. The final four columns provide results for the estimation of an analagous model but with a dummy variable indicating the presence of a streak of three high or low exemplar scores. Fixed effects control for semester specific variation. Standard errors are robust.

* significant at 10%; ** significant at 5%; *** significant at 1%

The first column of Table 2 reports the result of the above estimation using OLS. An F-test fails to reject the null hypothesis of no autocorrelation in exam quality across the three lagged problem set scores (F = .17, p =.68). Given that the subsequent analysis focuses on the influence of high and low scoring past exams ("exemplars"), as well as the influence of streaks of such exams, the next several columns in the table tests for random ordering after streaks of high scoring problem sets (i.e. exams falling in the upper 80th%-tile, or 70th%-tile of scores for that semester) or low scoring problem sets (i.e. exams falling in the

for the exemplar streaks of two exams. These results provide similar, but far more imprecise, evidence compared to that reported in Table 2. The estimation on the restricted sample cannot be meaningfully completed for exemplar streaks of three exams due to the limited sample.

lower 25th%-tile, or 35th%-tile of scores for that semester). The asymmetry in thresholds used to define "high" and "low" exemplars is due to the asymmetry in the occurrence of streaks of such exemplars.[11]

The table offers no significant evidence for negative correlation in problem set scores across the alphabet. Columns 2 to 5 suggest modest negative, but insignificant, levels of correlation after exemplar streaks of two exams. However, Columns 6 to 9 suggest modest, and insignificant, positive correlation in ability after both high and low scoring streaks of three exams. The estimated coefficients are small relative to the average problem set score of about 40. A later analysis of multiple choice questions both serves as a second test of random order, as well as a test of transparency.

Antecedent Effects – Lags. An essential, but not sufficient, indication that graders are subject to contrasts in the evaluation of final exams is the existence of an antecedent effect (Prediction 1). Given that perceived and intrinsic quality is not observable in this setting, the following panel model tests for the influence of lagged evaluations on the scoring of subsequent exams:

$$Final_{s,t,q} = \alpha + \sum \beta_i (Final)_{s,t-i,q} + \sum \gamma_m Ability^m_{s,t} + \eta_{sq} + \varepsilon_{s,t,q} \quad (1)$$

where $Final_{s,t,q}$ indicates the score in semester s of student t for final exam question q. $Final_{s,t-i,q}$ indicates the score of the ith lagged exam. Fixed effects for each question are included to control for question specific variation in scoring. Meanwhile, $\sum Ability^m_{s,t}$ represents a set of proxies for student ability and is estimated by the student score on the MCQs, PS as well as a student's composite score on the midterms.[12] Standard errors are clustered at the level of each question (s x q).

Column 1 of Table 3 reports results of an OLS estimation for all exams and offers no evidence for antecedence. An F-test of the joint influence of the lagged scores (F = .40, p = .75) is insignificant.

Antecedent Effects – Exemplars. The lack of a parametric link between current and lagged exam scores does not signal the absence of an antecedent effect. It is possible that such effects are elicited only after an extreme exam, or streak of such exams. The estimation of the equation below explores the possibility that contrast effects are triggered by recent exposure to exemplars:

[11] For most questions, there is greater density on the right tail of the scoring distribution than the left tail. As such, in order to roughly equate the size of high and low streaks, the streaks are defined with asymmetric thresholds.

[12] Potentially the midterm is subject to contrast effects as well. Using the total midterm score across 7 to 10 questions in each of two midterms should not bias the estimate significantly. Estimating the equation without the midterm scores as a control does not qualitatively alter the results, but does increase the standard errors.

$$Final_{s,t,q} = \alpha + \beta D^{F,k,n}_{s,t-1,q} + \sum \gamma_m Ability_{s,t,m} + \eta_{sq} + \varepsilon_{s,t,q} \quad (2)$$

where a dummy variable $D^{F,k,n}_{s,t-1,q}$ indicates that the prior n exam scores fall either above or below the kth—defined as the 80th or 25th—percentile for a given question. This specification more flexibly allows for a link between current and past exam scores.[13]

Table 3

ANTECEDENT EFFECTS IN GRADING FOR SINGLE EXAMS AND STREAKS

	DEPENDENT VARIABLE - EXAM SCORE (OLS)								
	FULL SAMPLE			RESTRICTED SAMPLE (ORDER > 100)					
	LAGS	EXEMPLARS		EXEMPLARS					
	SINGLE EXAM			SINGLE EXAM		STREAKS			
						Two Exams		Three Exams	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Exam Score - Lag 1	0.003								
	(0.011)								
Exam Score - Lag 2	0.006								
	(0.009)								
Exam Score - Lag 3	-0.004								
	(0.018)								
Lag Exam		-0.037		0.164		0.308		1.265***	
Score < 25%		(0.120)		(0.180)		(0.280)		(0.316)	
Lag Exam			-0.069		-0.195		-0.285		-0.410
Score > 80%			(0.091)		(0.100)		(0.195)		(0.306)
N	N = 7888	N = 7971	N = 7971	N = 5172	N = 5172	N = 5172	N = 5172	N = 5172	N = 5172
R^2	0.65	0.65	0.65	0.66	0.66	0.66	0.66	0.66	0.66

Notes: The dependent variable is the score for a single question of the final exam for each student. The first column presents results of an OLS regression of exam scores on lagged scores for the prior three students graded. The next two columns provide results of an OLS regression of exam scores on a dummy variable indicating that the prior score is either a low or high exemplar. Low exemplars are scores falling in the left tail (< 25%-tile) of the distribution for the question, while high exemplars are exams falling in the right tail (> 80%-tile) of the distribution. Columns 4 and 5 provide results of the estimation of an analagous model but estimated on a restricted sample which excludes the first 100 exams. The final four columns report results of an estimation of exemplar streaks of two and three exams for the restricted sample. Fixed effects control for question specific variation across all specifications. Standard errors are robust and clustered at the question level.

* significant at 10%; ** significant at 5%; *** significant at 1%

With exemplar exams defined as specified, an antecedent effect would be indicated by a negative point estimate for the lagged high exemplar and a positive point estimate for

[13] At first glance, this estimation strategy may seem to suffer from a bias due to "sampling without replacement." However, while such an estimation is biased for very small samples, for larger samples, the bias is negligible. To see this, note that after a low scoring exam, the next exam can be low scoring as well. Simulations which demonstrate the absence of a measurable bias, unreported here, are available from the author.

the lagged low exemplar. For example, if a past exam is high, then a grader subject to misperception would perceive the subsequent exam to be contrasted downwards. While Columns 2 and 3 do not suggest an effect for either single high or low exemplars, the remaining set of columns estimate the model after restricting the sample by excluding early exams. One possibility why later, but not early, exams may be subject to contrast effects is that grader fatigue sets in after several exams and leads to a higher likelihood of perceptual errors. A second possibility is that the distribution of exemplars is such that the estimation of later exams has more power than that of the full set of exams. As such, subsequent estimations focus only on the restricted sample of exams.

The results of Columns 4 and 5 are more suggestive of an antecedent effect consistent with contrasts for either a single high or low exemplar, but are not significant. It is conceivable, however, that contrast effects are elicited only after a *streak* of exemplars. For instance, a grader who has evaluated a series of poor exams may be more likely to misperceive a subsequent exam due to contrast than the grader who has faced only a single poor exam. The section on model extensions discusses how the model could be adapted to account for additive antecedent effects across periods.

The final four columns of Table 3 explore the effect of exemplar streaks. The columns report the estimation of Equation 2 with a dummy variable indicating that the most recent two or three exams are above or below the given threshold. The monotonic rise in the size of the estimated coefficients across the last six columns suggests that the antecedent effect is additive across streaks of exemplars. Column 8 indicates that an exam following a streak of three poor exams, for example, elicits a score 1.3 points higher than that predicted by ability. The estimate represents an 8% rise relative to the average score of a question (16.4), and is about 18% of the standard deviation for an average question. Coefficients for low exemplar streaks are in the expected direction and are monotonically increasing in magnitude as well, but are smaller and not significant.

The lack of statistical significance for many of the estimates of Table 3 may be partially due to the small incidence of streaks of 2 or 3 exams, as well as to the unrestricted definition of exemplars. Table 4 explores the influence of a streak of three exemplars after varying the levels of threshold restrictiveness. The first three columns estimate antecedent effects triggered by streaks of low exemplars. Column 1 reports that an exam following a streak of scores in the bottom 15% percent of the scoring distribution elicits a 2.0 point rise above that predicted by proxies for ability. This represents an added lenience of about 13% relative to the baseline score. Columns 2 and 3 suggest that the effect of lesser exemplar streaks—that is a series of exams below the 25 or 35 percentile, exclusive of the already accounted for streaks—also produce higher lenience but of more moderate size.

Conversely, the last three columns of Table 4 report effects elicited by streaks of high

exemplars. Column 4 reports that an exam following a streak of scores in the upper 10% of the scoring distribution produces a 1.0 point grade loss relative to that predicted by student ability. This represents a 6% distortion relative to the mean exam. The final two columns point to modest or no effect for lesser exemplar streaks—that is exams above the 80th or 70th percentile exclusive of already accounted for streaks. The treatment sizes at the bottom of the table mark the rarity with which exams are preceded by streaks of the identified nature. Table 4 offers evidence consistent with a contrast effect, but the effect is limited to the small sample of exams that follow streaks of very high and low exemplars.

Table 4

ANTECEDENT EFFECTS IN GRADING FOR EXEMPLAR STREAKS OF THREE EXAMS

	DEPENDENT VARIABLE - EXAM SCORE (OLS)					
	RESTRICTED SAMPLE					
	LOW EXEMPLARS			HIGH EXEMPLARS		
	(1)	(2)	(3)	(4)	(5)	(6)
Lag Exam Score < 15%	2.035*** (0.660)	2.043*** (0.660)	2.056*** (0.650)			
Lag Exam Score < 25%		0.825* (0.450)	0.838* (0.450)			
Lag Exam Score < 35%			0.467 (0.330)			
Lag Exam Score > 90%				-1.045* (0.570)	-1.047* (0.570)	-1.037* (0.570)
Lag Exam Score > 80%					-0.195 (0.390)	-0.189 (0.390)
Lag Exam Score > 70%						0.432 (0.330)
Treatment Size	21	58	188	36	138	355
N	N = 5172	N = 5172	N = 5172	N = 5172	N = 5172	N = 5172
R^2	0.66	0.66	0.66	0.66	0.66	0.66

Notes: The dependent variable is the score for a single question of the final exam for each student. The six columns present results of an OLS regression of exam scores on a dummy variable indicating that the prior three scores are all either low or high exemplars. Low exemplars are scores falling in the left tail of the distribution for the question, while high exemplars are exams falling in the right tail of the distribution. Each estimation relies on a restricted sample which excludes the first 100 exams. Fixed effects control for question specific variation across all specifications. Standard errors are robust and clustered at the question level.

* significant at 10%; ** significant at 5%; *** significant at 1%

Figure 1 graphically compares the direction and magnitude of antecedent effects across exam order for high and low exemplar streaks. The figure plots the residuals estimated from Equation 2 after excluding the exemplar dummy variable for all exams which follow a streak of exemplars. Exams preceded by low scoring exams generally produce positive

residuals, while exams preceded by high scoring exams generally produce negative residuals. Importantly, the spread of the residuals does not diminish, and may actually increase, as more exams are graded. This is evidence against the possibility that learning is fully responsible for generating these effects.

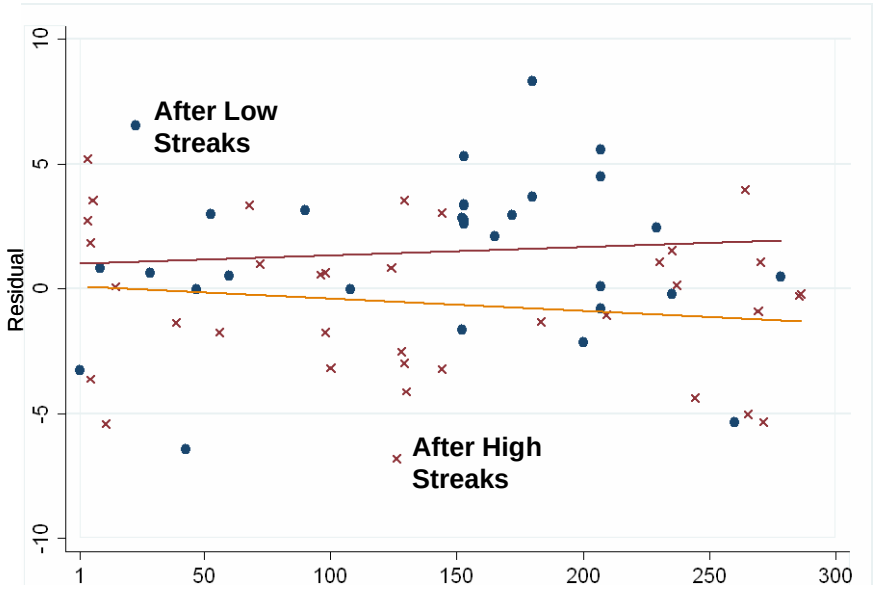

Figure 1, Unexplained Residual After Streaks of Three High (>90%-tile) or Low (<15%-tile) Exams across Rounds

Placebo Check and Test of Decay. A concern in this analysis is that serial correlation across periods may result in spurious rejection of the null hypothesis of no effect. Figures 2 and 3 provide a non-parametric test of the hypothesis that $\beta = 0$ for streaks of either low or high exemplars in Equation 2.[14] A first step in this approach is to estimate placebo regressions generated by constructing placebo streaks of exemplars. For example, Column 1 of Table 4 reports the coefficient estimate of β for a low scoring exam streak stretching across periods (t-1), (t-2) and (t-3). One could estimate the same equation, but with a dummy variable instead indicating a placebo streak of low exemplars stretching across periods (t-2), (t-3), and (t-4). Denote this estimate of the new dummy variable $\widehat{\beta}_p$. This procedure can be repeated nearly 100 times by estimating Equation 2 with streaks stretching back to periods (t-98), (t-99), and (t-100). These coefficient estimates should not be systematically positive since such distant exemplar streaks should not influence the current evaluation. The number of placebo estimates is constrained by the number of

[14] This non-parametric approach is inspired by that followed by Bertrand et al. (2002) in their analysis of double difference estimates. Another implementation of the method can be found in Chetty, Looney, and Kroft (2007).

available lags given the sample is restricted to exclude the first 100 observations.

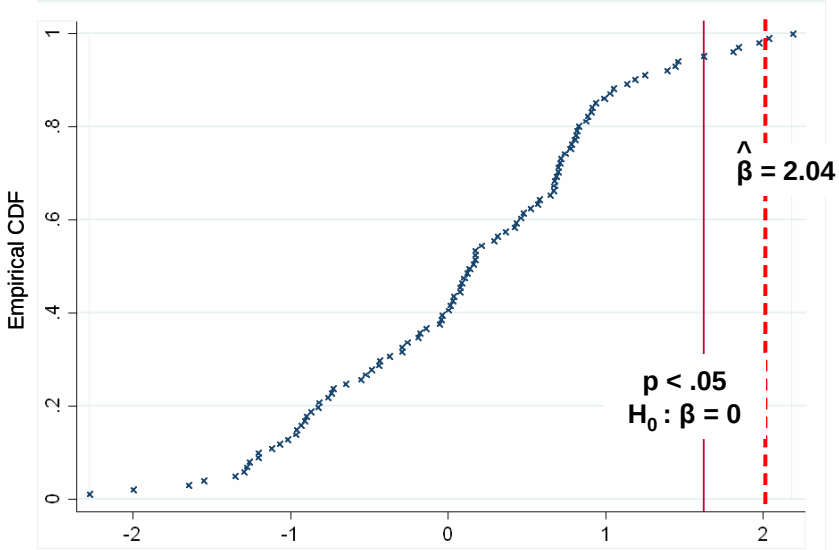

Figure 2, Distribution of Placebo Coefficients for Low Streaks (< 15%-tile)

Figure 2 plots the distribution of placebo estimates $\{\widehat{\beta}_p\}$ in the sample for streaks of low exemplars (defined as a streak of three scores in the bottom 15%-tile). The resulting empirical CDF provides a visual test of the hypothesis that $\beta = 0$. The CDF resembles a normal distribution which implies that the parametric assumptions underlying the t-tests in Table 4 are reasonable. The non-parametric test allows one to reject the hypothesis of no effect ($p = .02$).

Figure 3 displays the analogous distribution of placebo estimates for streaks of high exemplars (defined as a streak of three scores in the upper 10%-tile). The empirical CDF rejects the hypothesis that $\beta = 0$ for such streaks but with less precision ($p = .12$). The large p-value for this non-parametric test is not surprising given that this test is likely to have considerable less power than the corresponding t-test. Nevertheless, for streaks of both low and high exemplars, the results of the placebo analysis are consistent with the findings of Table 4.

Figures 2 and 3 also provide evidence that the antecedent effect decays over time. The presence of decay implies that the observed findings cannot be explained through standard rational explanations. Evidence for decay is discussed further below.

Contrast Effects. The antecedent effects found above are consistent with contrast effects as well as alternative explanations emerging from rational preferences—including the possible existence of a quota for high and low evaluations—and updating. Predictions 3 (Decay), 4 (Experience) and 5 (Transparency) offer tests through which to differentiate

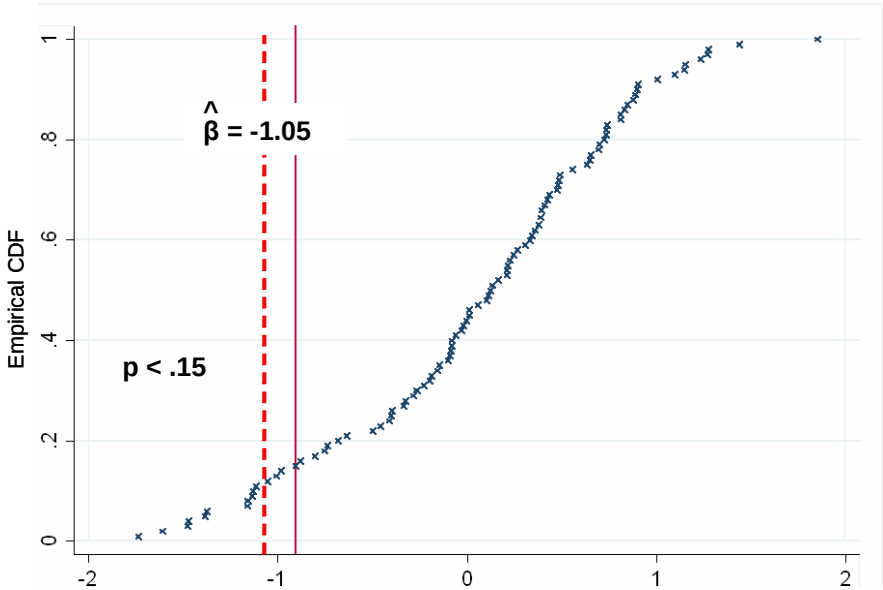

Figure 3, Distribution of Placebo Coefficients for High Streaks (> 90%-tile)

between these explanations. Specifically, the first of these predicts that for the contrast-effects model, the antecedent effect should decay as the intervening distance between the current and prior exam increases. This prediction is born out by Figures 2 and 3. In Figure 2, only one of the 100 lagged streaks of low exemplars is associated with a higher coefficient estimate than the original streak estimated in Table 4. This implies a decay in the influence of the streak as the intervening distance between current exam and the end of the streak grows. Though less stark, the same interpretation applies to Figure 3. It is worth noting that the coefficient estimate for the first lagged streak (i.e. a streak stretching from exams (t-2), (t-3), and (t-4)) for both high and low exemplars suggests an immediate partial decay (low streaks: $\widehat{\beta}_{lag1} = .10$, high streaks: $\widehat{\beta}_{lag1} = -.80$).

Prediction 4 claims that, for a decision-maker subject to contrasts, the antecedent effect will not disappear as the decision-maker accrues experience. The above estimations, in highlighting the strength of the antecedent effect in later, as opposed to earlier, exams, offers one piece of evidence consistent with the prediction. Figure 1 provides additional, visual, support for the prediction.

Finally, Prediction 5 holds that transparent questions, which permit little grading subjectivity, should not elicit contrasts. It is, however, difficult to categorize final exam questions by the degree of evaluative transparency. One clearly demarcated category of questions is the set of multiple choice questions on each final exam. Table 5 replicates the analysis reported in Table 3 but for multiple choice questions where $Ability_{s,t,m}$ now

includes only midterm and problem set scores. As with problem sets, the sample size is limited since each student has only one MCQ score, so the model is estimated for the full set of exams. The table reports estimated coefficients for lagged scores, high and low exemplars, as well as streaks of exemplars. Table 5 provides no consistent evidence for contrast effects in the grading of multiple choice questions. Analagous estimates for the restricted sample of exams (unreported here) are imprecise, as they are based on a small number of exemplar streaks, but similarly do not support the existence of contrasts.

Table 5

CONTRAST EFFECTS FOR TRANSPARENT GRADING

	DEPENDENT VARIABLE - MULTIPLE CHOICE QUESTIONS (OLS)						
	FULL SAMPLE LAGS	FULL SAMPLE EXEMPLARS					
		Single Exam		Two Exams		Three Exams	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
MCQ Score - Lag 1	0.031 (0.032)						
MCQ Score - Lag 2	0.0536* (0.031)						
MCQ Score - Lag 3	0.012 (0.031)						
Lag MCQ Score < 25%		-0.149 (0.350)		0.220 (0.790)		-0.316 (0.600)	
Lag MCQ Score > 80%			0.327 (0.310)		0.212 (0.530)		0.392 (0.950)
N	N = 1127	N = 1138	N = 1138	N = 1131	N = 1131	N = 1127	N = 1127
R^2	0.14	0.48	0.48	0.48	0.48	0.48	0.48

Notes: The dependent variable is the score for the multiple choice questions (MCQ) of the final exam for each student. The first column presents results of an OLS regression of MCQ scores on lagged scores for the prior three students graded. The next two columns provide results of an OLS regression of MCQ scores on a dummy variable indicating that the prior score is either below or above a low or high exemplar. Low exemplars are scores falling in the left tail (< 25%-tile) of the distribution for the question, while high exemplars are exams falling in the right tail (> 80%-tile) of the distribution. The final four columns report results of an estimation of exemplar streaks of two and three exams for the full sample. Fixed effects control for semester specific variation across all specifications. Standard errors are robust.

* significant at 10%; ** significant at 5%; *** significant at 1%

Distortion of Final Grades. If contrast effects do exist, how do such biases affect students? Table 6 calculates the number of distorted grades produced by biases implied by the above estimates for exemplar streaks of varying restrictiveness. The grades are recalculated using the precise algorithm used in the course after correcting for the estimated contrast effects for those exam questions following exemplar streaks.[15] The net effect is

[15] The (highly conscientious) instructor for the course generates a composite score for each student by

that the contrast effects distort grades for 1-2 percent of students (e.g. from a B to a B+). The incidence of these grade distortions fall largely on students in the middle to upper range of the grading distribution.

Table 6

INFLUENCE OF CONTRAST EFFECTS ON STUDENTS

	SCENARIO 1 (> 90% , < 15%) (+2.0, -1.0)	SCENARIO 2 (> 80% , < 25%) (+1.3, -.4)
Questions Influenced:	42	142
Students Influenced:	31	108
Average Score Shift:	1.4%	2.4%
Total Grade Distortion:	12 (6 up, 6 down)	17 (11 up, 6 down)
	12 to 20 students (~1-2% of students)	
Incidence of Distortion:	4 As, 5 Bs, 3 Cs	7 As, 6 Bs, 4 Cs

When interpreting the size and importance of these distortions, it is worthwhile to note that, in the absence of systematic ordering patterns across classes, a student suffers from no ex-ante distortion in expected grades. Particularly, over the course of 30 to 40 undergraduate courses, one would expect that positive and negative distortions, to the extent that they occur, would negate one another. On the other hand, one could imagine that distortions are additive within a course if such biases existed across all the graded components of a course. There may also be schooling environments where systematic ordering does exist across classes and over years (e.g. elementary school). Moreover, to the extent that similar effects occur in other evaluative domains such as hiring or employee evaluation, effects of this size could substantively distort welfare for those affected.

4.2 Judicial Sentencing Decisions

One setting of clear public policy importance where relativity in sequential assessments can be tested is that of judicial decisions. Experimental evidence suggests that legal professionals in lab settings may be sensitive to irrelevant context (Englich and Mussweiller 2001). Judges in courtrooms characterized by short and high frequency hearings of charges with varying severity may be particularly sensitive to the sequential context produced by the order in which cases are heard.

weighing the student's final exam, midterm exam and problem set scores. Additional points are earned through improvement on the second midterm and the final exam. Finally, problem set scores are adjusted for student specific fixed effects.

The present analysis, originally reported in Bhargava and Cann (2007), examines sentencing patterns for minor traffic and non-traffic violations heard in lower level PA courts from 2001 to 2005. In the PA judicial system, lower level judges primarily hear minor non-criminal charges, but on occasion also preside over preliminary hearings and arraignments for severe criminal offenses committed in their jurisdiction. If not dismissed, withdrawn or plea-bargained, criminal offenses are eventually passed on to higher judicial authorities such as the Court of Common Pleas.

Table 7

DISTRIBUTION OF HEARINGS IN PA DISTRICT COURTS, 2001 - 2005

Offense	%	Guidelines	Examples
Criminal Felony	3		
Felony 1		> 10 yrs	Murder, Rape, Aggravated Assault,
Felony 2		> 7 yrs, < 10 yrs	Manslaughter, Statuary Rape, Arson
Felony 3		> 5 yrs, < 7 yrs	Infanticide, Insurance Fraud, Animal Cruelty
Criminal Misdemeanor	4	< 5 yrs	Disorderly Conduct, False Identification, DUI
Summary Traffic	24	< 90 days	Speeding, Careless Driving, Improper Permit
Summary Non-Traffic	11	< 90 days	Littering, Unauthorized Sale of Tickets, Criminal Mischief
Civil and Landlord Tenant Disputes	16	None	X
Non-Summary Traffic and Non-Traffic	19	None	X
Other or Unlabeled	22	Varies	X

Note: Categories describe hearings in district courts for PA counties outside of Philadelphia and Pittsburgh from 2001 to 2005. Calendars of hearings by day by judge provided by the AOPC. Sentencing guidelines taken from AOPC publications and website. Share statistics calculated by authors. For certain records, hearing type could not be inferred.

Magisterial District Courts constitute the initial level of the judicial system in PA (see Appendix Table 2 for a detailed description of the PA hierarchy of courts). Judges presiding over such courts (MDJs) are primarily responsible for adjudicating minor civil and traffic cases in Pennsylvania counties outside of Philadelphia and Pittsburgh.[16] Table 7 describes the various cases heard by a MDJ. The district courts conduct non-jury hearings making MDJs solely responsible for rendering judgments and assigning penalties and fines. Approximately 500 full time judges dispose of roughly 12 to 15 cases a day totaling over 7 million cases during the five year sample period. Of these, summary offenses—traffic and

[16] Philadelphia has distinct traffic courts for traffic related offenses and municipal courts for less serious, non-traffic offenses. While Pittsburgh has MJDs, minor criminal cases and preliminary hearings for more serious offenses are heard in separate municipal courts. Accordingly, this analysis only considers district courts in counties outside of Philadelphia and Pittsburgh.

minor non-traffic violations punishable by up to 90 days imprisonment and/or a fine not exceeding $300—constitute the largest category of cases, and represent over one-third of total charges.[17] The courts additionally hear civil and landlord disputes, non-summary hearings for traffic and minor non-traffic cases (e.g. primarily hearings for defendants unable to make payments on existing fines), and conduct arraignments, bail hearings and other preliminary hearings for more serious criminal matters. Of hearings involving criminal charges, about 40% are arraignments, and 57% are preliminary hearings. The Administrative Office of Pennsylvania Courts (AOPC) compiles and maintains case-level statistics for the courts considered in this analysis.

Research Design. One test of the influence of sequential context on judicial decision-making is whether judges exhibit abnormal lenience after exposure to a criminal infraction. One might expect criminal charges to prompt a contrast effect given that such violations are considerably more severe and rare than other charges on the judicial docket. Criminal infractions are split between felonies and misdemeanors. Felonies carry a recommended prison sentence of at least five years and include charges such as murder, assault, arson, and severe cases of animal cruelty. Misdemeanors carry recommended sentences under five years and include more minor charges such as disorderly conduct and driving under the influence. The most frequent criminal felonies or misdemeanors in our data involve drug use (16%), driving under the influence (16%), and assault (11%). A typical judge hears either a criminal felony or misdemeanor on about 1-2% of days.[18] On such days, the judge will often preside over multiple criminal hearings.

Lenience in summary trials is the ideal dependent variable for the analysis for three reasons. First, summary trials involve transparent declarations of guilt.[19] This is in contrast to non-summary offenses involving private parties (e.g. civil and landlord disputes), where declarations of guilt are not made and lenience cannot clearly be inferred, as well as hearings for criminal offenses which are typically transferred to higher courts for final judgment. Second, judges in summary trials are afforded considerable discretion. While bound by guidelines which assign fines to charges, judges retain the ability to reduce the severity— and consequently the associated fine—of final charges for any traffic violation. Of hearings which go to trial, judges dismiss 22% of charges due to either insufficient evidence

[17] This fraction is higher if one considers summary offenses which are attached to a criminal charge against a particular defendant. District judges either dispose of such cases themselves or send such cases to a higher court along with the accompanying criminal charge. These summary offenses are included in "Other or Unlabeled" in Table 7 and are excluded for the purpose of the analysis.

[18] The actual frequency of days with criminal exposure is higher than this if one were to account for those charges for which judges have discretion to upgrade or change to a misdemeanor or felony charge, or records for which charge type is unavailable.

[19] This is ignoring those cases which (i) do not proceed to trial because of a guilty plea, (ii) are withdrawn by the prosecutor, or (iii) are dismissed by the judge typically due to insufficient evidence, absence of key witnesses or, when applicable, the officer who authored the citation.

or absence of key prosecutorial person(s), render a "not guilty" verdict in an additional 23% of charges, and decide either partial (i.e. guilty but with a reduced charge) or full guilt in the remaining 55% of charges. Finally, summary trials are generally short—often 15 to 30 minutes—which permits several decisions to be rendered during the course of a day. The transparency of lenience, the permitted judicial discretion, and the high frequency of such decisions make summary offenses amenable to analysis.

Random Ordering. As in grading and dating, it is important that the ordering of evaluated targets is quasi-random. Such ordering makes it unlikely that an unobserved variable would confound estimates of a contrast. Interviews with court administrative clerks, the AOPC, and the administrative MDJ office suggests that one notable feature of case scheduling is that courts often devote specific days of the week to traffic hearings or private party disputes. However, within any given day, including days earmarked for particular offense categories, it is usually the case that other charges are also heard.[20] The ordering of charges within a day is dictated by police officer or prosecutorial availability rather than any offense characteristic.[21] While judges have discretion in the scheduling of non-criminal offenses, criminal offenses are scheduled by the District Attorney.[22] These hearings are scheduled with advanced notice ranging from a few days to several weeks. Judges are required to interrupt their dockets in order to accommodate such hearings as needed.

In principle it is possible to test for random ordering of charges empirically. A challenge is that one must both characterize the relative severity across summary offenses as well as the link between infraction severity and lenience. Indirect empirical tests of random order are discussed and presented in Bhargava and Cann (2007). The tests provide no evidence that charges are ordered in a manner which would generate the findings reported below.

Antecedent and Contrast Effects. The basic empirical result is illustrated in Figures 4 and 5. The figures plot residual leniency in summary offenses after controlling for judge x charge-specific variation as a function of case order within a day. Figure 4 compares residual leniency after exposure to a felony charge on days where a single felony is heard to the residual leniency for summary offenses not immediately following the felony. The evidence is consistent with an antecedent effect—greater relative lenience immediately after felony exposure. This effect does not appear to diminish in later cases during a day. Figure 5 implies the antecedent effect disappears on days for which the docket includes

[20] Defendants have the right to request that hearing times and dates be modified, but such requests are dealt with differently by different courts. Given that most summary trials do not involve lawyers, it is unlikely that any coordination in scheduling across cases exists.

[21] We learned details of scheduling procedures from interviews with the AOPC, court clerks and the MDJ administrative office.

[22] Frequently, District Attorneys try to schedule multiple criminal hearings on the same day since such hearings require the presence of a Public Defender as well as the District Attorney.

multiple felony hearings. The possibility that the salience of a felony depends on the frequency of felonies within a day reconciles these two plots. When only a single criminal hearing occurs during a day, it is likely to appear more psychologically salient to the judge. The salience of an extreme target increases the probability that it may trigger a perceptual contrast (Higgins 1996).

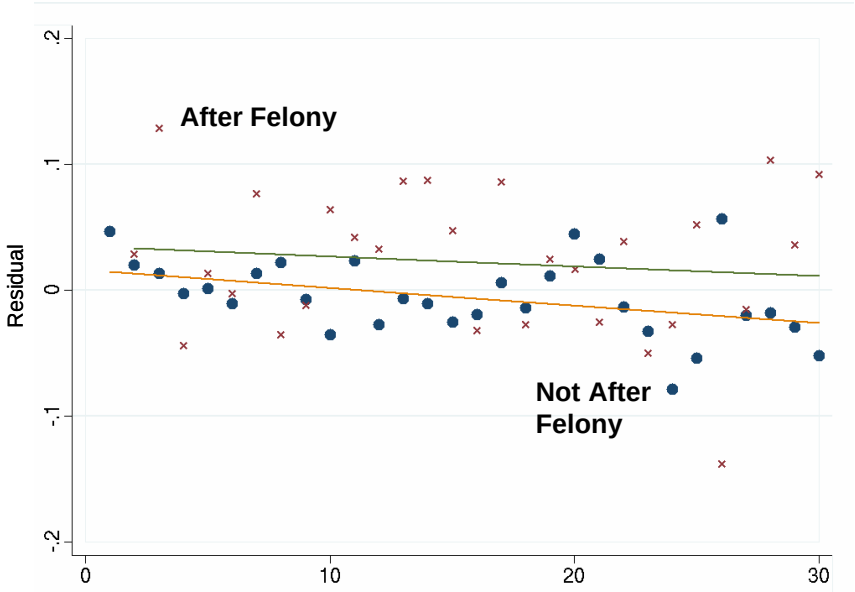

Figure 4, Unexplained Residual Leniency on Days with 1 Felony by Case Order

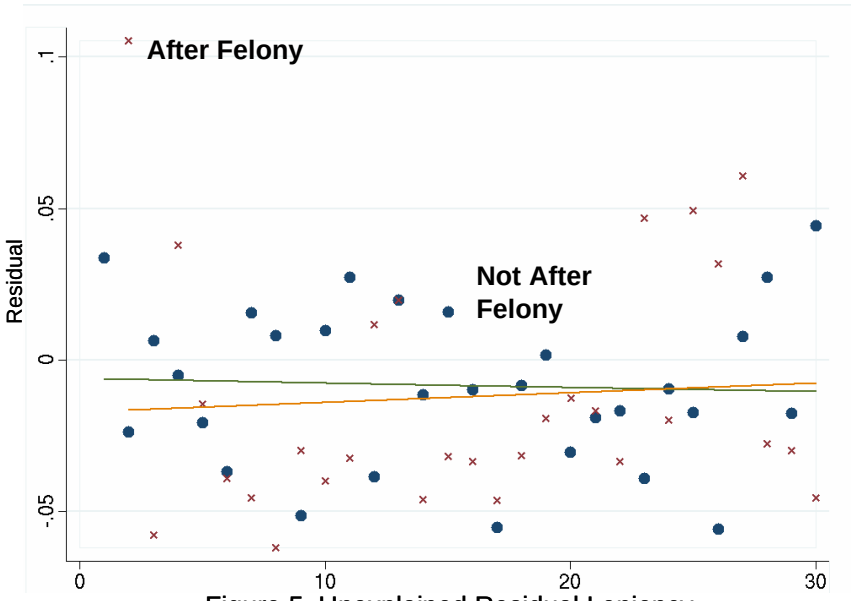

Figure 5, Unexplained Residual Leniency on Days with > 1 Felony by Case Order

The specification below formally tests for the link between judicial lenience and past infractions as well as the rate at which this link decays:

$$Lenience_{j,c,d,t} = \alpha + \sum \gamma_i D^{crime}_{j,d,t-i} + \lambda Single_{j,d,t} + \sum \delta_i (Single_{j,d,t} * D^{crime}_{j,d,t-i}) + \eta_{jxc} + \eta_d + \varepsilon_{j,c,d,t} \quad (3)$$

The dependent variable $Lenience_{j,c,d,t}$ is a binary variable indicating whether a judge j grants either high (i.e. a not-guilty verdict) or any (i.e. a not-guilty or a guilty verdict but with a reduced charge) lenience to a defendant for charge c on day d at period t.[23] While summary trials are generally scheduled every 30 minutes, multiple hearings are often scheduled concurrently.[24] For this reason a period refers to a time period as opposed to an individual hearing. $D^{crime}_{j,d,t-i}$ is an indicator variable which denotes the occurrence of a criminal felony or misdemeanor i periods in the past. To ensure that the criminal hearings constitute a sharp negative exposure, dismissed or withdrawn hearings are excluded. $Single_{j,d,t}$ is an indicator variable denoting a day during which only a single felony or misdemeanor is heard. The interaction effect is included to identify differential antecedent effects for days with a single criminal hearing. Fixed effects are included to control for judge x charge specific variation, as well as variation produced by the day of week and hour of day.

Table 8 estimates the above model for both felonies and misdemeanors. The first three columns report estimates for high lenience while the next three columns report results for any lenience. Columns 3 and 6 provide the most flexible pair of specifications. The baseline coefficients indicate that on days with multiple felonies or misdemeanors, criminal hearings do not result in heightened lenience for subsequent summary judgments. In fact, there is some evidence for small decreases in lenience. This result is consistent with an assimilation effect and is explored further in Bhargava and Cann (2007).

[23] Reduced charges only occur for traffic violations. Therefore, any and high lenience are equivalent for non-traffic violations but not for traffic violations.

[24] This is due, according to interviews, to the fact that many defendants do not show up and some cases take only minutes to dispose. Within these concurrently scheduled hearings order is not known.

Table 8

CONTRAST EFFECTS IN SUMMARY TRIAL JUDICIAL DECISIONS

	DEPENDENT VARIABLE					
	HIGH LENIENCY			ANY LENIENCY		
	Linear Probability Model					
	(1)	(2)	(3)	(4)	(5)	(6)
Felony - Lag 1	0.017	-0.020*	-0.018*	-0.019	-0.034**	-0.019*
	(0.018)	(0.011)	(0.011)	(0.028)	(0.013)	(0.010)
Felony - Lag 2	0.004	-0.010	-0.005	-0.018	-0.023*	-0.011
	(0.019)	(0.010)	(0.009)	(0.027)	(0.013)	(0.011)
Misdemeanor - Lag 1	0.041**	-0.009	-0.015*	0.021	-0.015	-0.014
	(0.016)	(0.009)	(0.008)	(0.034)	(0.015)	(0.012)
Misdemeanor - Lag 2	0.026	0.004	0.003	0.000	-0.008	0.002
	(0.017)	(0.010)	(0.009)	(0.028)	(0.011)	(0.009)
Single Felony	-0.024*	-0.010	-0.009	-0.048***	-0.011	-0.002
	(0.015)	(0.007)	(0.007)	(0.016)	(0.009)	(0.007)
Single Misdemeanor	-0.011	0.014**	0.009*	-0.037**	0.020***	0.019***
	(0.014)	(0.005)	(0.005)	(0.017)	(0.007)	(0.006)
Single Crime x Felony - Lag 1	-0.012	0.046***	0.054***	0.034	0.079***	0.055***
	(0.026)	(0.017)	(0.016)	(0.034)	(0.018)	(0.016)
Single Crime x Felony - Lag 2	-0.019	-0.015	-0.012	0.004	0.002	-0.004
	(0.023)	(0.018)	(0.018)	(0.025)	(0.024)	(0.023)
Single Crime x Misdemeanor - Lag 1	-0.062***	0.007	0.022*	-0.046	0.019	0.027*
	(0.021)	(0.014)	(0.012)	(0.040)	(0.019)	(0.016)
Single Crime x Misdemeanor - Lag 2	-0.019	0.003	0.004	0.005	0.014	0.005
	(0.022)	(0.016)	(0.014)	(0.028)	(0.020)	(0.017)
Judge x Charge Fixed Effects			X			X
Judge Fixed Effects		X	X		X	X
Hour, DOW Fixed Effects	X	X	X	X	X	X
N	N = 1168086	N = 1168086	N = 1168086	N = 1075205	N = 1075205	N = 1075205

Notes: The dependent variable for the first three columns is a binary variable indicating whether the judge delivers a "not guilty" verdict for a particular case that proceeds to trial, while the dependent variable for the final three columns indicates whether the judge delivers a "not guilty" verdict or reduces the initial charge for cases proceeding to trial. Only judgments of summary trials are considered here. Felony (Misdemeanor) Lag 1 and Lag 2 indicate whether one of the prior two time periods features a hearing for a felony (misdemeanor) offense. Single Crime refers to days for which only a single felony or misdemeanor occurs. The analysis controls for judge, and judge x charge fixed effects as well as hour and day of week fixed effects as indicated. Standard errors are robust and are clustered at the judge level.

* significant at 10%; ** significant at 5%; *** significant at 1%

On days with a single criminal felony, a felony produces a statistically significant increase in high lenience of 5.4% and any lenience of 5.5% in subsequent relative to non-subsequent decisions. Summing the coefficient of the interaction with the coefficients of the constituent variables indicates that judges are 2.7% more lenient (3.4% for any lenience) following a felony on single-felony days compared to lenience on other days. An F-test rejects the null that lenience is not serially correlated across felony hearings on salient days (F = 6.5, p

$< .02$ and $F = 8.4$, $p < .01$ respectively). The increase in lenience prompted by a felony compares to a typical high lenience of 29% and any lenience of 45% across the sample and implies a distortion of 8 to 9%. Even though it is unlikely that the observed effects are due to Bayesian learning because of both the in-sample and likely out-of-sample experience of a typical lower court judge, the rapid decay of the effects after a single period is further evidence against such an explanation.[25]

Similarly, on days with a single misdemeanor, a misdemeanor prompts an increase in high lenience of .9% and any lenience of 1.9% in subsequent relative to non-subsequent decisions. Summing again across pertinent coefficient values indicates that judges are 1.6% more lenient (3.2% for any lenience) following a misdemeanor on single-misdemeanor days compared to lenience on other days. An F-test rejects the null that lenience is not serially correlated across misdemeanor hearings on salient days ($F = 4.2$, $p < .05$ and $F = 14.0$, $p < .01$ respectively). As with felonies, the effect appears to decay after one period.

The 8 to 9% distortion in lenience (4 to 6% distortion in the likelihood of a guilty verdict) induced by felony exposure is comparable to the size of other biases found in research on judicial decision-making. It is about 2/3 as large as the 12% difference found in the proclivity of black as compared to white judges in PA criminal courts to sentence defendants to prison (Steffensmeier and Britt 2001). The effect sizes are also in the range of studies that have found sentencing differences due to defendant race (Steffensmeier and Britt 2001).

History of Felony Exposure and Dismissed/Withdrawn Charges. One dimension across which the sensitivity of lenience to felony exposure might plausibly differ across judges is past familiarity to felonies. Table 9 estimates the relationship between high lenience and a past felony for single crime days across history of felony exposure. The first two columns suggest that the antecedent effect is much larger for judges in the lower half of the distribution of exposure when exposure is measured by the percentage of in-sample days with at least one felony. This finding is consistent with the possibility that decision-makers for whom felonies have heightened salience display larger contrast effects. Salience has long been understood to be, at least in part, a function of stimulus novelty (Higgins 1996).

The final column of Table 9 suggests that there is no relationship between high lenience and past felony exposure for felony charges that were either dismissed or withdrawn. Presumably, exposure to dismissed or withdrawn charges is not equivalent to exposure to charges either held for further court proceedings or resulting in a guilty plea. Different dispositions are likely the result of differences in the hearings across several dimensions. One such dimension is the criminal severity of the charge or the guilt of the defendant. A

[25] It is even more unlikely that a judge is subject to a quota constraint in the scenario under consideration. Judges do not render any verdicts with respect to the criminal cases, so it is implausible that a criminal hearing would constrain future judgments of summary offenses.

judge will withdraw or dismiss a charge when the case against the criminal is weak, incomplete or otherwise procedurally flawed. Alternatively, differences across disposition may reflect differences in the length of the proceeding, the nature of the crime committed, or other procedural details. With such caveats in mind, the lack of sensitivity of subsequent lenience to a dismissed or withdrawn felony is consistent with contrast effects which are induced exclusively by exposure to serious criminality. This result is explored in greater depth in Bhargava and Cann (2007).

Table 9

CONTRAST EFFECTS IN JUDICIAL DECISIONS BY FELONY EXPOSURE AND FOR DISMISSED OR WITHDRAWN CHARGES

	DEPENDENT VARIABLE - HIGH LENIENCE		
	FELONY EXPOSURE		DIS / WITH
	1 Felony / Day		
	< 50%	> 50%	All Judges
	(1)	(2)	(3)
Felony - Lag 1	0.098**	0.038**	0.000
	(0.040)	(0.018)	(0.010)
Felony - Lag 2	-0.035	-0.032	-0.022*
	(0.040)	(0.025)	(0.013)
Number of Judges	289	306	595
Judge x Charge Fixed Effects	X	X	X
Hour, DOW Fixed Effects	X	X	X
N	N = 1547	N = 7644	N = 22789

Notes: The dependent variable across all columns is a binary variable indicating whether the judge delivers a "not guilty" verdict for a particular summary offense that proceeds to trial. The first two columns test whether lenience is sensitive to a past felony for judges in the upper and lower half of the sample distribution with respect to felony exposure. The final column estimates the sensitivity of lenience to past felony exposure for felonies which were either dismissed or withdrawn.

* significant at 10%; ** significant at 5%; *** significant at 1%

Discretion and Contrast Effects. Many federal and state court judges in the US have considerable discretion in the sentencing of criminal and civil infractions. This discretion may be increasing. A December 2007 U.S. Supreme Court decision granted federal judges even greater freedom to deviate from federal sentencing guidelines set by the US Sentencing Commission.[26] If the perceptual bias identified in the sentencing of minor offenses extends to the adjudication of more serious criminal cases, contrast effects could result in substantive unfairness for a non-negligible population of defendants. The recent pattern of heightened discretion further highlights the importance of research on the role of irrelevant context in determining judicial outcomes.

[26] Kimbrough v. U.S., Case No. 06-6330 (Dec. 10, 2007).

4.3 Speed Dating Decisions

Research Design. A final sequential decision-making environment where one can detect cognitive errors in perception is that of romantic *speed dating* (this analysis was originally reported in Bhargava and Fisman 2007). Speed dating refers to a structured matchmaking process in which men and women meet multiple partners through a series of short, sequential interactions.[27] Participants privately indicate romantic interest at the end of each interaction, and organizers distribute contact information in cases of mutual interest. The data in this analysis is gathered from 20 experimental speed dating sessions organized over several weekday evenings at Columbia University from 2002 to 2004 (Fisman et. al. 2005, 2006). Sessions range from 18 to 44 participants for a total of 474 subjects who collectively make 7684 decisions.

The experimental sessions depart slightly from the procedural details of a typical commercial speed dating event. At the close of each four minute interaction, organizers require subjects to privately evaluate each partner along a variety of dimensions including physical attractiveness.[28] Two independent research assistants additionally evaluate each subject during the course of each session to generate a measure of "objective" attractiveness on a 1-10 ordered scale. While subjects seat themselves, regressions of subject attractiveness on lagged attractiveness of prior ordered subjects reveal no systematic pattern in ordering.[29] As such, this environment provides an ideal set of decisions from which to test for contrast effects in the perception of attractiveness— subjects are randomly ordered, measures of attractiveness are available, the stakes are non-trivial, the setting is real, and preferences are revealed rather than surveyed.

The figures, as well as most of the more formal results which follow, focus on male subjects. Kenrick et. al. claim that gender asymmetries associated with contrast effects in the perception of physical attractiveness are consistent with evolutionary theories of sexual selection (1994). They argue that because males attend to "facial and bodily" attractiveness to a greater extent than females—a contention long supported by evolutionary theorists—male evaluations of romantic partners are more sensitive to sequential contrasts than female counterparts. One study consistent with this view finds that male subjects exposed to photographs of female models are less attracted to their girlfriends but that female subjects are not similarly affected by exposure to photographs of male models (Kenrick et.

[27] While the first organized speed dating event supposedly occurred in 1998 in Beverly Hills, the format has since gained popularity across the United States. As of 2003, one of nation's largest commercial organizers of speed dating, *8minuteDate*, reported over 60,000 customers.

[28] Questionnaires elicited subject ratings of 6 different partner attributes (also on a 1-10 point ordered scale) including intelligence, sincerity, ambition, shared interests, fun-lovingness, as well as physical attractiveness.

[29] Results of this estimation can be found in Bhargava and Fisman (2007).

al. 1989).

The data in the present analysis suggests that both male and female dating decisions are determined by partner attractiveness. Male subjects appear slightly more sensitive to partner attractiveness than females, but this difference is not statistically significant. However, since the evidence for contrast effects is strong for males and non-existent for females, the analysis focuses on the male subjects. This gender difference is discussed in more depth in Bhargava and Fisman (2007).

The basic empirical result is illustrated in Figures 6 and 7. Figure 6 compares the average affirmative response for male subjects after exposure to a high exemplar (i.e. a partner of attractiveness > 6) to the average affirmative response after exposure to a low exemplar (i.e. a partner of attractiveness < 5).[30] In order to control for selection, the sample is confined to subjects with at least 18 rounds in the sample (representing approximately 2/3 of male subjects) and only the first 18 rounds are reported. Fitted lines are drawn for the two exposure groups as well as for the entire sample. The figure conveys evidence consistent with an antecedent effect—greater lenience after low exemplars, and less lenience after high exemplars. While this antecedent effect appears smaller in later rounds, it does not fully disappear with dating experience.

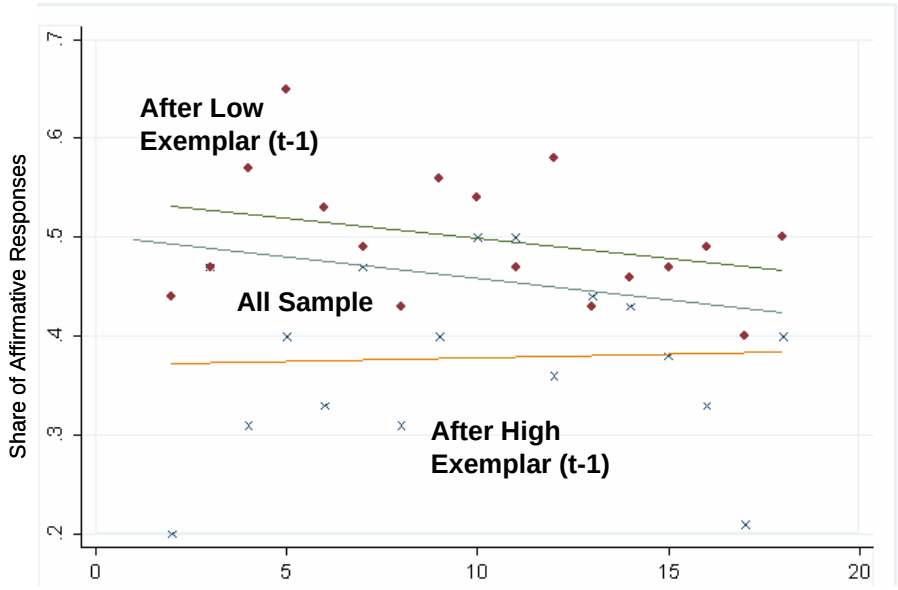

Figure 6, Average Male Affirmative Response After Lag 1 Exemplars across Rounds (subjects w/ rounds > 18)

[30] The assymetric cutoff points reflect assymetries in the distribution of attractiveness ratings.

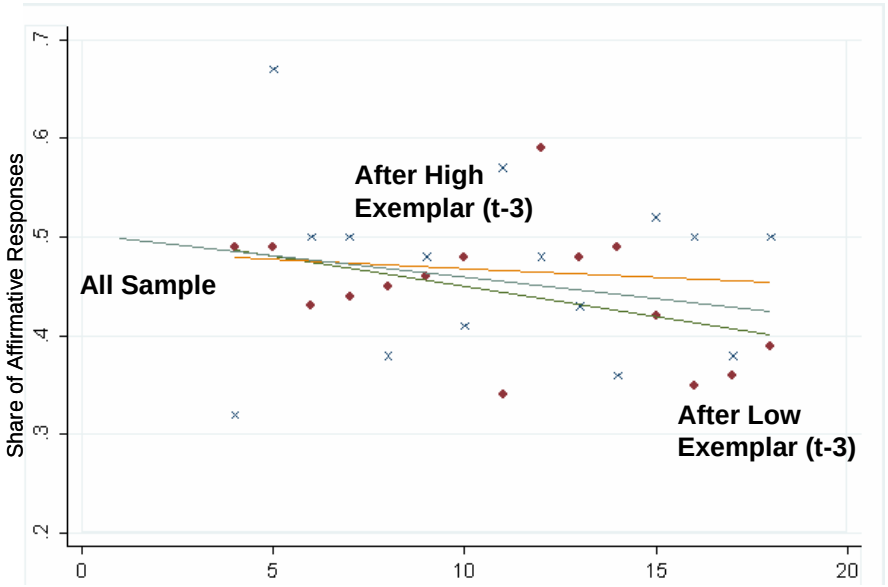

Figure 7, Average Male Affirmative Response After Lag 3 Exemplars across Rounds (subjects w/ rounds > 18)

Figure 7 provides insight into the rate at which the antecedent effect decays. The plot depicts dater responses after exposure to a high or low exemplar three periods in the past. The figure shows no evidence for an enduring antecedent effect. Plots after longer lags similarly show no evidence for such an effect, while the plot for the second lag shows a mild effect (unreported here). The one to two period decay implied by these figures is consistent with a model of contrasts but not the possibility that the effects are driven by subject quotas for matches or other explanations emerging from the standard model. The quick decay of the effect, as well as the persistence of the effect across rounds, support the existence of contrasts.

Antecedent Effect. The specification below formally tests for the antecedent effect between current decisions and past partner attractiveness:

$$Dec_{w,s,p} = \alpha + \gamma Att_{w,s,p} + \sum \beta_i Att_{w,s,p-i} + \eta_s + \varepsilon_{w,s,p} \tag{4}$$

where $Dec_{w,s,p}$ denotes the "yes/no" decision which subject s makes in regard to partner p in session w, and $Att_{w,s,p}$ indicates average partner attractiveness. The regression is estimated with a linear probability model and controls for current partner attractiveness and subject fixed effects.[31] Table 10 reports the results of the estimation.

[31] An alternative, and perhaps more direct, test is to explicitly examine the impact of past decisions on current decisions ($\partial g_t^c / \partial g_{t-1}^c$). However, because of the complications produced by estimations of binary dynamic panel data models with fixed effects (originally identified by Nickell (1981), see Chamberlain (1985) for a discussion), $\partial g_t^c / \partial q_t$ is calculated.

The first column of the table provides a baseline estimate of the importance of current and lagged partner attractiveness on subject decisions to date. The results corroborate Prediction 1 and indicate that a one unit rise in current partner attractiveness, all else equal, produces an approximately 11% rise in the likelihood of an affirmative decision.[32] Consistent with Prediction 2, the point estimate for the first lag, $Att_{w,s,p-1}$, suggests that a one unit rise in the attractiveness of the last partner lowers the probability of a subsequent affirmative response by about 2%.

Column 2 reports the analogous estimation but only for male subjects. The estimates indicate that a rise in the marginal attractiveness of the last partner produces a 3% drop in the likelihood of a subsequent affirmative decision. For example, under this estimate, a male subject facing a highly attractive partner with a rating of 9, would be roughly 13% less likely to accept a subsequent partner with a more modest rating of 5 than in the counterfactual where the partner had instead been preceded by an equally modest partner. The influence of the last partner is notable in light of the overall average acceptance rate for males of 48% as well as in relation to the 11% marginal influence of current partner attractiveness. Estimates for more distant partner attractiveness are not significant which suggests that the antecedent effect dies out quickly.

Only male subjects appear to exhibit strong antecedent effects. An F-test for the lagged model of Column 1 and 2, but estimated solely for female subjects (unreported here) fails to reject the null hypothesis of no autocorrelation across lagged attractiveness ($H_0 : \beta_1 = \beta_2 = \beta_3 = 0$, F = 1.00, p = .39). The remainder of the table reports estimates primarily for male subjects.

The remaining columns report the effects of prior exposure to either single or streaks of exemplar partners on subject decisions.[33] The estimation takes the following form:

$$Dec_{w,s,p} = \alpha + \gamma Att_{w,s,p} + \beta D^{k,i,n}_{w,s,p-1} + \eta_s + \varepsilon_{w,s,p} \tag{5}$$

where $D^{k,i,n}_{w,s,p-1}$ indicates whether the last n partners have an attractiveness rating either greater or less than k.

[32] Additional estimates—unreported here—indicate that the influence of current partner attractiveness on decisions is monotonic and approximately linear for increasing levels of attractiveness.

[33] The table reports results of estimates for streaks of 3 partners. Estimations for streaks of 2 partners is omitted here, but included in the original paper (Bhargava and Fisman 2007).

Table 10

ANTECEDENT EFFECTS IN MALE DATING DECISIONS AND ASSESSMENTS

	DEPENDENT VARIABLE - DECISION TO DATE (OLS)						ATTRACTIVENESS RATING (OLS)	
	SINGLE PARTNER LAGS		SINGLE PARTNER EXEMPLAR		STREAKS OF THREE EXEMPLAR		SINGLE PARTNER LAGS	
	All	Males	Males		Males		All	Males
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Partner Attractiveness	0.109***	0.113***	0.114***	0.116***	0.115***	0.116***	0.693***	0.678***
	(0.008)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.036)	(0.051)
Partner Attractiveness - Lag 1	-0.020**	-0.032***					-0.027	-0.080**
	(0.008)	(0.011)					(0.034)	(0.039)
Partner Attractiveness - Lag 2	-0.008	0.004					-0.064	-0.032
	(0.008)	(0.011)					(0.035)	(0.048)
Partner Attractiveness - Lag 3	-0.004	-0.005					-0.017	-0.020
	(0.008)	(0.011)					(0.036)	(0.048)
Partner Attractiveness - Lag 1 > 6			-0.062*		-0.112*			
			(0.037)		(0.057)			
Partner Attractiveness - Lag 1 < 5				0.081***		0.084		
				(0.029)		(0.063)		
Treatment Size			619	1049	104	154		
N	N = 6234	N = 3108	N = 3592	N = 3592	N = 3108	N = 3108	N = 6056	N = 3018
R^2	0.34	0.35	0.33	0.34	0.35	0.35	0.49	0.48

Notes: The dependent variable for the first six columns is a binary variable indicating the "yes" / "no" decision to date for each subject across each interaction. The first two columns present results of a linear probability model of the dating decision on current partner attractiveness and lagged attractiveness of prior partners for all subjects as well as just male subjects. Columns 3 and 4 estimate a linear probability model of male subject dating decisions on current partner attractiveness and a dummy variable indicating that the prior partner is a low or high exemplar. Low exemplars are partners with attractiveness ratings < 5 while high exemplars are partners with attractiveness ratings > 6. The next two columns provide results of an analagous estimation but with a dummy variable indicating a streak of three exemplar partners. The final two columns repeat the estimation of Columns 1 and 2, but substitute the subject rating of partner attractiveness as the dependent variable. Fixed effects control for subject specific decisions across all specifications. Standard errors are robust and clustered at the partner level.

* significant at 10%; ** significant at 5%; *** significant at 1%

Columns 3 and 4 provide evidence for an antecedent effect after a single exemplar partner. Column 3 suggests that a highly attractive partner in the prior period with a rating greater than 6— representing about 1/3 of the total sample of female partners— elicits a 6% drop in the likelihood of a subsequent acceptance beyond that predicted by current partner attractiveness. A less attractive prior partner, with a rating less than 5, increases the likelihood of subsequent acceptance by 8%. The asymmetry in the threshold definitions reflect the asymmetric distribution of attractiveness across the sample.

The next two columns report analogous results for streaks of three highly attractive or unattractive partners. A streak of attractive partners elicits (an imprecisely measured) fall of 11% in the subsequent likelihood to date. This is equivalent to the influence of a unit rise in current partner attractiveness and is almost twice as large as the influence of a single lagged attractive partner. Column 6 does not offer parallel evidence for the additive influence across streaks of unattractive partners. This estimate is equally as imprecise as in the high exemplar case and is due, at least partially, to the relative scarcity of exemplar streaks.[34]

Finally, in order to confirm that the observed effects operate through distorted perceptions or judgments of attractiveness, as opposed to an alternative mechanism, one can examine the influence of recent partner attractiveness on subject ratings of current partner attractiveness. Such a test doubly serves to reject the presence of quotas or learning since only a model of perceptual contrasts is consistent with $\partial \tilde{q}_t^c / \partial q_{t-1} < 0$.[35]

Accordingly, the final two columns of Table 10 estimate Equation 3 after substituting subject decisions with subject ratings of attractiveness as the dependent variable. For male subjects, the estimation indicates an antecedent effect with respect to subject ratings. As compared to the original model in Column 2, the size of the effect relative to the influence of current partner attractiveness is much smaller (approximately 11% compared to 27%). This may be partially attributable to nonlinearity in the contrast effect. On the whole, Table 10 suggests that prior exposure to partners distorts impressions of physical attractiveness as well as decisions to date.

Contrast Effects. While the estimates above provide evidence consistent with contrast effects, Prediction 3 (Decay), Prediction 4 (Experience), and Prediction 5 (Transparency) offer direct tests through which to reject rational alternative explanations. Pre-

[34] The threshold for the attractive and less attractive streaks was selected to produce a roughly equivalent volume of streaks in each category. For instance, there are no streaks of three partners with an attractiveness rating less than 4 in the entire sample.

[35] One possible mechanism through which learning could cause antecedent effects in ratings, as well as decisions, is through semantic contrast effects. That is, subjects might arguably redefine the mapping between attractiveness and the 1-10 scale during each round. However, the stable relationship between "objective" and subjective partner ratings across rounds (not reported here) seems to argue against semantic contrasts.

diction 3 holds that the influence of a partner on a future decision should decay as the intervening distance between the decisions increases. The small and insignificant estimates for the second and third lags in Columns 1 and 2 provide initial evidence for this decay. The decay is also implied by Figures 6 and 7.

The figures also offer preliminary evidence that experience does not fully account for the antecedent effect, as do the final two columns of Table 10. The following regression more directly tests whether the observed antecedent effect diminishes with dater experience:

$$Dec_{w,s,p} = \alpha + \gamma Att_{w,s,p} + \beta Att_{w,s,p-1} + \theta Exp_{w,s,p} + \lambda(Att * Exp)_{w,s,p-1} + \eta_s + \varepsilon_{w,s,p} \quad (6)$$

where $Exp_{w,s,p}$ refers to the experience accumulated by subject s prior to meeting partner p in session w. Other variables are as previously defined. The prediction of learning would imply a positive estimate for λ— that is, as a subject progresses through a session and ascertains the underlying distribution of dater attractiveness, the influence of the most recent partner on the current partner decision should diminish (i.e. become less negative). Table 11 reports the results of the estimation. The coefficient estimates of the interaction term in Columns 1 and 2 are positive. While they are small enough such that the effect does not fully diminish by session end, the Column 2 estimate implies that as much of 4/5ths of the effect disappears by the last round.

Columns 3 through 6 report the estimation of a similar model, but with a dummy variable indicating the presence of either a single or streak of high or low exemplar partners in prior periods. While not significant, the interaction estimates are directionally consistent with learning (i.e. positive for high lagged exemplars, and negative for low lagged exemplars). For streaks, they imply that 20 rounds of experience diminishes up to 2/3rds of the effect. The weakening of the main antecedent effect with experience, coupled with strong evidence for the decay in the effect for more distant lags, suggests the presence of a contrast effect, but one that may diminish by the end of a particular session. The possibility that psychological biases in perception exist, but that their influence diminishes over time, is plausible and is explored in greater depth in Bhargava and Fisman (2007).

The final two columns of Table 11 examine whether the transparency of an evaluation influences the size of the contrast effect (Prediction 5). Absent the explicit distinction between ambiguous and non-ambiguous targets in the grading data, one test for transparency in this setting is whether partners of moderate attractiveness elicit stronger contrasts than those on either extremes of the distribution. The presumption is that highly attractive or unattractive partners are more transparent targets than their counterparts.

Table 11

CONTRAST EFFECTS IN MALE DATING DECISIONS BY EXPERIENCE AND TRANSPARENCY

	DEPENDENT VARIABLE - DECISION TO DATE (OLS)							
	EXPERIENCE TEST SINGLE PARTNER Lags		EXPERIENCE TEST SINGLE PARTNER Exemplar		STREAKS OF THREE Exemplar		TRANSPARENCY SINGLE PARTNER Exemplar	
	All	Males	Males		Males		Males	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Partner Attractiveness	0.107***	0.114***	0.114***	0.115***	0.115***	0.116***	0.114***	0.113***
	(0.008)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.014)	(0.014)
Partner Attractiveness - Lag 1	-0.025**	-0.051***						
	(0.011)	(0.014)						
Partner Attractiveness - Lag 1 > 6			-0.102**		-0.165		-0.083*	
			(0.050)		(0.104)		(0.044)	
Partner Attractiveness - Lag 1 < 5				0.125***		0.150		0.112***
				(0.041)		(0.100)		(0.034)
Partner Attractiveness - Lag 1 x Experience	0.001	0.002**	0.004	-0.004	0.006	-0.005		
	(0.001)	(0.001)	(0.004)	(0.003)	(0.009)	(0.006)		
Exemplar Lag 1- (High/Low) x Exemplar (High)							0.037	-0.170***
							(0.071)	(0.055)
Exemplar Lag 1- (High/Low) x Exemplar (Low)							0.077	-0.079*
							(0.074)	(0.046)
Treatment Size			619	1049	104	154	619	1049
N	N = 7192	N = 3592	N = 3592	N = 3592	N = 3108	N = 3108	N = 3592	N = 3592
R^2	0.33	0.34	0.33	0.34	0.35	0.35	0.33	0.34

Notes: The dependent variable is a binary variable indicating the "yes" / "no" decision to date for each subject across each interaction. The first two columns present results of a linear probability model of the dating decision on current partner attractiveness and lagged attractiveness of the prior partner for all subjects as well as just male subjects. Also included is an interaction of lagged attractiveness and subject experience as measured in number of dates. Columns 3 and 4 estimate a linear probability model of male subject dating decisions on current partner attractiveness, a dummy variable indicating that the prior partner is a low or high exemplar, as well as an interaction of the dummy variable with the experience measure. Low exemplars are partners with attractiveness ratings < 5 while high exemplars are partners with attractiveness ratings > 6. The next two columns provide results of an analagous estimate but with a dummy variable indicating a streak of three exemplar partners. The final two columns provide a test of transparency by estimating the differential contrast effects for current partners of high and low attractiveness as compared to partners of median attractiveness. Fixed effects control for subject specific decisions across all specifications. Standard errors are robust and clustered at the partner level.

* significant at 10%; ** significant at 5%; *** significant at 1%

This test of transparent targets is carried out by estimating the original model of antecedence but with the inclusion of an additional interaction of the exemplar indicator with a dummy variable indicating that the current partner falls in either tail of the distribution (i.e. > 6 or < 5). The estimates of the interactions confirm that, for both low and high lagged exemplars, the contrast effect is muted— or non-existent— if the current partner is either extremely attractive or unattractive. An F-test fails to reject the null of no autocorrelation in lagged partner attractiveness for either the low or high end of the distribution for high exemplars ($F = .66, p = .42; F = 1.05, p = .31$) or for low exemplars ($F = .01, p = .93; F = .46, p = .50$). This is consistent with Prediction 5, as well as psychological evidence which suggests that psychological biases are heightened for ambiguous stimuli.

5 Discussion

Alternative Explanations. A number of alternative explanations exist for the findings in this paper including learning, and the presence of quotas for high or low evaluations. The strongest evidence against both of these explanations, as given by the model, is the rapid decay of the observed effects across all three contexts. In all three settings, the effects decay within a single period. Additional, but less strong, evidence rules out the possibility that the effects fully disappear as decision-makers accumulate evaluative experience.

An unexplored explanation is limited memory. For example, in the extreme instance where a decision-maker only recalls the last target, Bayesian learning would prompt a low evaluation in a period subsequent to a naturally occurring high evaluation. However, the presence of limited memory alone is unlikely to produce these results for a truly rational Bayesian. Simple limitations in memory are not likely to preclude a genuinely rational decision-maker from near optimal updating so long as she is able to commit a small number of sufficient statistics—such as the empirical mean and sample size—to memory.

In order for limitations in memory to induce behavior consistent with a contrast effect, the decision-maker must be subject to additional biases such as selective memory, selective recall, or the presence of some evaluation heuristic. It is possible that such a combination of factors—or other biases such as the gambler's fallacy or base-rate neglect—could account for the observed behavior. In grading, the presence of the effect only after streaks of at least three exams echoes some of the properties of the gambler's fallacy (Rabin and Schrag 2002). Finally, it is possible that the effects may be due to contrast effects prompted not by a cognitive or perceptual error, but instead by a mood induction. Particularly in the case of judicial sentencing, exposure to a criminal hearing may induce a mood which results in sympathy for future defendants.

Comparisons Across Domains. The settings considered in this paper differ along a number of psychologically meaningful dimensions—the semantic content, duration, frequency and welfare importance of each decision, as well as the interval between decisions. Such differences complicate attempts to compare estimated effect sizes across domains or to make claims about broader applicability. However, the breadth of the experimental literature, as well as the field examples cited in this paper suggest that these effects may be widespread.

Table 12 compares the analytical results across the domains. The first panel of the table compares the normalized effect sizes across high and low exemplars as well as across single and exemplar streaks. The table reports the effect sizes relative to the baseline decision (i.e. average score in points for grading, average male affirmative decision for dating, and average rate of high lenience for judging), as well as relative to the standard deviation of the baseline decision.

Contrast effects appear in the expected direction for both low and high exemplars. High exemplars produce negative error in subsequent perception while low exemplars produce a positive error. A common pattern across high exemplars in the dating and grading domains is stronger effects for streaks as opposed to a single target. In grading, similar differential effects exists for streaks of low exemplars. It is difficult to identify additional patterns across the domains. Magnitudes are smallest for grading, and in grading, more extreme exemplars are needed to trigger contrasts than judging or dating. Both of these facts seem congruous with the notion that a human partner is a more salient target than an exam question and that the length of an interaction between an evaluator and target may be linked to the size of the effect.

The second panel of Table 12 compares the decay of the contrast effects across the domains as well as the differential size of the effects for transparent targets across grading and dating. The table indicates that, across the three settings, effects decay immediately. As discussed, the presence of decay helps rule out rational explanations for the findings. Moreover, identifying systematic patterns in the rate of decay could allude to prescriptions on how to avoid contrasts in important settings.

Finally, across grading and dating, there is no evidence for contrast effects for highly transparent targets. Such targets do not allow for much evaluative discretion and do not elicit any strong forms of psychological bias. The absence of an effect for these targets both confirms a feature of contrast effects suggested by the psychology literature and provides a robustness check of the research design. The question of transparency for judicial decisions is slightly more complicated and is discussed in greater depth in Bhargava and Cann (2007).

Table 12

COMPARISON OF CONTRAST EFFECTS ACROSS DOMAINS

	CONTRAST EFFECTS									
	GRADING				DATING				JUDGING	
	HIGH (> 90%)		LOW (< 15%)		HIGH (> 6 ; > 84%)		LOW (< 5 ; < 27%)		LOW (Criminal Offense)	
	Baseline	SD	Baseline	SD	Baseline	SD	Baseline	SD	Baseline	SD
Single Exemplar	-0.01	-0.03	0.01	0.02	-0.13	-0.12	0.17	0.16	0.09	0.06
Exemplar Streak	-0.06	-0.15	0.12	0.29	-0.23	-0.22	0.18	0.17	X	X

	TRANSPARENCY AND DECAY				
	GRADING		DATING		JUDGING
	HIGH Exemplar Streaks	LOW Exemplar Streaks	HIGH Single Exemplar	LOW Single Exemplar	LOW Single Exemplar
High Transparency	No Effect	No Effect	No Effect	No Effect	X
Full Decay	1 Lag	1 Lag	1 Lag	1 Lag	1 Lag

Notes: The first panel refers to estimated effects for regression coefficients reported in Table 4, Table 8 and Table 10, as well as an unreported regression coefficient for the case of a single exam exemplar. Estimates are presented as a ratio of the baseline evaluation (i.e. the average score of 16.4 for grading, or the average male acceptance rate of .48 for dating, or average high lenience of .29), as well as a ratio of the standard deviation of the baseline evaluation (i.e. 7.0 for grading, .50 for dating, and .45 for judging). The second panel of the table summarizes findings for transparency and decay.

How broadly applicable are these results to other decision-making contexts? This paper has discussed how contrasts might influence evaluations in settings where targets are arranged in quasi-random order. Randomly ordered domains include those explicitly examined (grading, judicial decisions, and dating) as well as other settings such as hiring, investment appraisal, and medical diagnoses. In theory, however, sophisticated agents might exploit the existence of contrast effects in decision-making settings which feature non-random ordering. As an example, a number of settings exist where an understanding of contrast effects would presumably benefit sellers who are able to shape the order in which buyers perceive goods.

Real Estate Survey. One example of a setting with non-random and non-trivial evaluation is the real estate market. Despite a recent stagnation in home sales, the National Association of Realtors (NAR) projects that buyers will purchase nearly 5.8 million existing homes in 2007 with a median value of $220,000.[36] Of these, approximately 95% will be handled by accredited real estate agents. As of 2007, a total of 1.3 million realtors are listed with the NAR.

To assess whether realtors are aware and actively take advantage of psychological biases in perception, an original survey of real estate agents from several Minnesota offices of a large national real estate company was conducted in September 2007. The online survey was distributed to a total of 800 to 1200 realtors from 8 to 12 offices and elicited a response rate of 8 to 13%. Precisely 100 realtors responded to the survey.[37] The survey respondents had a median level of experience from 6 to 10 years and represented a uniform mix of sales productivity.

Of those surveyed, 71% agree that exposing a buyer to a home "which is overpriced, is in a bad neighborhood, or is otherwise unattractive" increases the likelihood that a buyer would favorably perceive the subsequent home. Of realtors aware of this tendency towards contrasts, 47% admit to exploiting this knowledge to increase the chances of a buyer liking a particular home.[38] Realtors with higher sales productivity admit to more frequent exploitation of contrast effects to encourage home buying but these differences are not statistically significant (i.e. 41% of total respondents in the upper two quintiles of sales productivity as compared to 30% of realtors in the lower three quintiles made this claim, $t = 1.00$, $p = .32$).

[36] This projection is from the NAR as of October 2007. The projection does not include new home sales (which number approximately 800,000 for 2007).

[37] The actual number of realtors who received the survey is unknown. A link to the online survey was sent to office managers across twelve offices. It is unclear how many office managers actually distributed the email to their staffs. The survey was described as containing a series of questions on the "psychology of decision making." As an incentive, $50 gift certificates were randomly distributed to participants.

[38] Reassuringly, a number of realtors volunteered that they would not do anything in opposition to their client's interest. Proximity was also cited by many realtors as the overwhelming determinant of the order of home showings.

The findings of this survey are consistent with home buyers who exhibit contrast effects as well as some home sellers who strategically exploit knowledge of these tendencies. Importantly, many agents suggest that the recent proliferation in the accessibility of online housing information has reduced the realtor's role in determining the choice and sequence of homes considered for purchase. To this extent, the role of contrasts in home buying may be on a decline. Beyond the usual caveats associated with self-reported evidence, further investigation is required to rule out alternative rational explanations that might account for the survey responses.

6 Conclusion

Considerable experimental evidence suggests that human perception is fundamentally comparative. Indeed, relativity may be a foundational principle of all social judgments (Kahneman and Miller 1986). This may have relevance for a wide range of important sequential decision-making settings such as employee hiring, medical diagnoses, judicial decisions, product and price assessments, as well as student evaluation. Empirical research has not yet documented this relativity in high-frequency decisions in the field. In this paper, I present evidence for such comparative assessment, or contrast effects, across three domains: exam evaluation, judicial sentencing, and dating.

The paper communicates two principle findings. First, there is modest evidence for a negative relationship between scores of evaluated exams, and stronger evidence for such effects in judicial sentencing and dating. Relative to baseline assessments, a streak of low scoring exams produces a 12% score increase while a streak of high scoring exams prompts a 6% score reduction. On days when felonies have high salience, judges in lower PA courts are about 9% more lenient in adjudicating summary trials after exposure to a felony than otherwise. Finally, male speed daters are 17% more willing to date after exposure to an unattractive partner, and 13% less willing to date following an attractive partner. A second finding is that the effects decay fully after a single period across all domains. Consistent with the theoretical framework, the presence of decay helps to rule out possible alternative explanations emerging from rational preferences and updating.

Research on contrast effects is ultimately valuable to the extent to which it informs our awareness of and improves our ability to combat biases associated with social decisions in real-life settings. In the data on exam scores, grades are modestly distorted for 3 to 10% of students, and biases in final exam grading alter the final semester grades of an estimated 1 to 2% of students. In judging, the studied domain for which the policy implications are most pronounced, the effect sizes are similar in magnitude to biases which other researchers have attributed to judge and defendant race. The fact that judges with past experience to felony

exposure exhibit smaller biases suggests that the selective allocation of such cases across judges may be a possible policy intervention. The model implies that simple education with respect to the bias may be sufficient to minimize its impact. Further study is needed to determine the efficacy of other interventions such as mandatory end-of-day sentence reviews, or cooling off periods following worrisome exposure.

Additional research is needed to determine the range of sequential decision-making settings in which individuals are subject to perceptual errors. Conceivably, in settings such as medical diagnosis even single instances of an evaluative error might prove very costly. To the extent that contrast effects do exist, it is important to uncover systematic patterns in the size, duration and contextual triggers of this behavior. It is probable that such effects are mediated by individual differences, as well as situational context. Knowledge of such factors will help formulate effective policy responses.

7 References

Ariely, Dan and George Loewenstein, and Drazen Prelac, ""Coherent Arbitrariness": Demand Curves Without Stable Preferences," *Quarterly Journal of Economics*, Vol. 188, No. 1, pp. 73-106, 2003.

Bhargava, Saurabh, and Andrea Cann, "Law and Order: Contrast Effects in Judicial Decisions," *UC Berkeley Working Paper,* 2007.

Bhargava, Saurabh, and Ray Fisman, "Romantic Relativity: Contrast Effects in Speed Dating," *UC Berkeley Working Paper,* 2007.

Camerer, Colin and Dan Lovallo, "Overconfidence and Excess Entry: An Experimental Approach," *American Economic Review,* Vol. 89, No. 1, pp. 306-318, 1999.

Cash, Thomas, and Diane Cash and Jonathan Butters, "Mirror, Mirror, on the Wall...?" *Personality and Social Psychology Bulletin*, Vol. 9, No. 3, pp. 351-358, 1983.

Damisch, Lysann and Thomas Mussweiler, and Henning Plessner, "Olympic Medals as Fruits of Comparison? Assimilation and Contrast in Sequential Performance Judgments," *Journal of Experimental Psychology: Applied*, Vol. 12, No. 3, pp. 166-178, 2006.

Englich, Birte, and Thomas Mussweiler, "Sentencing under Uncertainty: Anchoring Effects in the Court Room," *Journal of Applied Social Psychology*, Vol. 31, pp. 1535-1551, 2001.

Fisman, Ray and Sheena Iyengar, and Emir Kamenica, and Itamar Simonson, "Racial Preferences in Dating," *Review of Economic Studies*, forthcoming.

Fisman, Ray and Sheena Iyengar, and Emir Kamenica, and Itamar Simonson, "Gender Differences in Mate Selection: Evidence from a Speed Dating Experiment," *Quarterly Journal of Economics*, Vol. 121, No. 2, pp. 673-697, 2006.

Friedrich, Wilkening, and Viktor Sarris and Otto Heller, "Contrast Effects in the Child's Judgment of Lifted Weight," *Psychonomic Science*, Vol. 28, No. 4, pp. 207-208, 1972.

Fryer, Roland, and Matthew Jackson, "A Categorical Model of Cognition and Biased Decision-Making," Forthcoming in *Contributions in Theoretical Economics,* B.E. Press, 2007.

Herr, Paul, "Consequences of Priming: Judgment and Behavior," *Journal of Personality and Social Psychology*, Vol. 51, pp. 1106–1115, 1986.

Higgins, Edward, "Knowledge Activation: Accessibility, Applicability, and Salience," in Edward Higgins and Arie Kruglanski (Editors), *Social Psychology: Handbook of Basic Principles*, New York: Guilford, 1996.

Hood, J.D., "Studies in Auditory Fatigue and Adaptation," *Acta OtoLaryngologica (Supplement)*, Vol. 92, pp. 1-57, 1950.

Kahneman, Daniel and Dale Miller, "Norm Theory: Comparing Reality to Its Alternatives," *Psychological Review*, Vol. 93, No. 2, pp. 136-153, 1986.

Kamenica, Emir, "Contextual Inference in Markets: On the Informational Content of Product Lines," *Working Paper, University of Chicago*, 2007.

Keil, Andreas and Thomas Mussweiler, and Kai Epstude, "Alpha-band Activity Reflects Reduction of Mental Effort in a Comparison Task: A Source Space Analysis," *Brain Research,*Vol. 1121, pp. 117-127, 2006.

Kenrick, Douglas, and Steven Gutierres, "Contrast Effects and Judgments of Physical Attractiveness: When Beauty Becomes a Social Problem," *Journal of Personality and Social Psychology*, Vol. 38, pp. 131-140, 1980.

Kenrick, Douglas, and Steven Gutierres and Laurie Goldberg, "Influence of Erotica on Ratings of Strangers and Mates," *Journal of Experimental Social Psychology*, Vol. 25, pp. 159-167, 1989.

Kenrick, Douglas, and Daniel Montello, and Sara Gutierres, and Melanie Trost, "Effects of Physical Attractiveness on Affect and Perceptual Judgments: When Social Comparison Overrides Social Reinforcement," *Personality and Social Psychology Bulletin*, Vol. 19, No. 2, pp. 195-199, 1993.

Kenrick, Douglas and Steven Neuberg, Kristin Zierk, and Jacquelyn Krones, "Evolution and Social Cognition: Contrast Effects as a Function of Sex, Dominance, and Physical Attractiveness," *Personality and Social Psychology Bulletin*, Vol. 20, No. 2, pp. 210-217, 1994.

Koszegi, Botond and Matt Rabin, "A Model of Reference-Dependent Preferences," *Quarterly Journal of Economics*, Vol. 121, No. 4, pp. 1133-1165, 2006.

Laibson, David, "Golden Eggs and Hyperbolic Discounting," *Quarterly Journal of Economics*, Vol. 112, No. 2, pp. 443-478, 1997.

Melamed, Leslie, and Martin Moss, "The Effect of Context on Ratings of Attractiveness of Photographs," *Journal of Psychology*, Vol. 90, pp. 129-136, 1975.

Mussweiler, Thomas, "Comparison Processes in Social Judgment: Mechanisms and Consequences," *Psychological Review*, Vol. 110, pp. 472-489, 2003.

Mussweiler, Thomas, "The Durability of Anchoring Effects," *European Journal of*

Social Psychology, Vol. 31, pp. 431-442, 2001.

Quattrone, George, and Edward Jones, "The Perception of Variability within In-Groups and Out-groups: Implications for the Law of Small Numbers," *Journal of Personality and Social Psychology*, Vol. 38, pp. 141-152, 1980.

Rabin, Matthew, and Joel Shrag, "First Impressions Matter: A Model of Confirmatory Bias," *Quarterly Journal of Economics*, Vol. 114, No. 1, pp. 37-82, 1999.

Rabin, Matthew, and Joel Shrag, "Inference by Believers in the Law of Small Numbers," *Quarterly Journal of Economics*, Vol. 117, No. 3, pp. 775-816, 2002.

Hovland, Carl, and Muzafer Sherif and Daniel Taub, "Assimilation and Contrast Effects of Anchoring Stimuli on Judgments," *Journal of Experimental Psychology*, Vol. 55, No. 2, 1958.

Simonsohn, Uri, and George Loewenstein, "Mistake #37: The Effect of Previously Encountered Prices on Current Housing Demand," *The Economic Journal*, Vol. 116, pp. 175-199, 2006.

Simonsohn, Uri, "New Yorkers Commute More Everywhere: Contrast Effects in the Field," *The Review of Economic and Statistics*, Vol. 88, No. 1, pp. 1-9, 2006.

Simonson, Itamar and Amos Tversky, "Choice in Context: Tradeoff Contrast and Extremeness Aversion," *Journal of Marketing Research*, Vol. 29, pp. 281-295, 1992.

Steffensmeier, Darrell and Chester Britt, "Judges' Race and Judicial Decision Making: Do Black Judges Sentence Differently?" Social Science Quarterly, Vol. 82, No. 4, pp. 749–764, 2001.

Strack, Fritz, and Thomas Mussweiler, "Explaining the Enigmatic Anchoring Effect: Mechanisms of Selective Accessibility," *Journal of Personality and Social Psychology*, Vol. 73, pp. 437-446, 1997.

Tversky, Amos, and Daniel Kahneman, "Belief in the Law of Small Numbers," *Psychological Bulletin,* Vol. 76, No. 2., pp. 105-110, 1971.

Tversky, Amos, and Itamar Simonson, "Context-Dependent Preferences," *Management Science,* Vol. 39, No. 10., pp. 1179-1189, 1993.

8 Appendix - Model Extension for Quota Constraints

A shortcoming of the utility function specified in the model above is that it the does not reflect the possibility that decision-makers may be subject to constraints for high or low evaluations. The presence of such constraints or "quotas" are likely in many sequential decision settings including those examined in this analysis. For instance, a grader may prefer to limit the number of assigned As or Fs independent of an interest in accuracy, while a judge may be constrained in the number of defendants she can sentence to prison because of limited prison capacity or other institutional pressures. Conceivably a speed dater can only accommodate a limited number of successful matches due to emotional, financial, and temporal constraints.

The Environment. A preference for accuracy subject to quotas can be described by the constrained utility below:

$$U(g_t, \psi_t) = -\sum_{t=0}^{T}(g_t - \psi_t)^2 \;\; subject\ to \;\; : H_t \leq \overline{H} \;\; for\ all\ t \tag{7}$$

where H_t represents the number of accumulated high grades by period t where $H_t = \sum_{t=0} I(g_t > \overline{g})$ and $\overline{g}$ is the threshold above which grades are considered high. One might expect $\overline{g}$ to be an exogenous function of the finite total size of the distribution T. As before, ψ_t is the idiosyncratic component of the exam score realized in period t. This is equivalent to a white noise stochastic shock which is distributed normally. The desire for grading accuracy is indicated through the minimization of $-\sum_{t=0}^{T}(g_t - \psi_t)^2$. The utility is assumed to be time-separable in that the marginal utility of a grade assigned in one period depends only on that period's grade.

The grader's problem is to maximize the expected value of future utility:

$$\max_{\{g_t\}_{t=0}^{T}} \;\; E\left[U(g_t, \psi_t)\right] \;\; subject\ to: \;\; H_t \leq \overline{H} \;\; for\ all\ t \tag{8}$$

where the discount rate of 1 reflects the grader's indifference in utility gained across periods. E is the expectation with respect to the probability distribution of the random variable ψ_t. In this formulation, g_t is the control variable which must be chosen in each period by the decision-maker, and H_t is a state variable which describes the system at any given point in time.

Assume that the timing of events and information is as follows: (1) At the beginning of period t, the idiosyncratic score ψ_t is realized. (2) The decision-maker observes this score (or more precisely $\widehat{\psi}_t$) as well the number of previously assigned high grades H_t. (3) The grader assigns a grade g_t and updates H_{t+1} accordingly. (4) The next period occurs.

Characterizing A Solution. In principle, one could treat this problem as a constrained optimization. A grader could at time t outline a set of contingency plans, $\{g(\psi_t, H_t)\}_t^T$, for each future period which would assign grades based on the possible realizations of the idiosyncratic score ψ_t and the accumulated number of high grades H_t. A much more tractable strategy, however, is to use a finite-period dynamic programming approach which takes advantage of the recursive nature of the problem.

One can define a policy rule to govern the choice of the control variable g_t in each period as a function of that period's exogenous exam score draw and the state variable. Assuming a decision-maker subject to such a policy rule, and no psychological error in perception, a heuristic argument can be used to show that evaluations are (weakly) negatively influenced by past evaluations, and that for a given period, the influence of a past high grade is invariant to the order in which it occurs:

Proposition 3 (Constrained Antecedence - Evaluation) *(i) An increase in the evaluation of an exam leads to a weakly lower evaluation of the immediate subsequent exam:* $\partial g_t/\partial g_{t-1}|_{g_{t-i,i>1}} \leq 0$ *and (ii) the influence of the evaluation of an exam on the evaluation of a subsequent exam, holding other past evaluations constant, is not a function of the distance between the exams:* $\partial g_t/\partial g_{t-k}|_{g_{t-i,i>k}} = \partial g_t/\partial g_{t-l}|_{g_{t-i,i>l}}$ *so long as* $\Delta I(g_{t-k} > \overline{g})|_{g_{t-i,i>k}} = \Delta I(g_{t-l} > \overline{g})|_{g_{t-i,i>l}}$ *for all* k, l.

Proposition 3 outlines behavior roughly similar to that produced by the standard model with preferences over just accuracy. Part (i) of the proposition claims that an exam score (weakly) negatively influences the immediate subsequent exam. The influence of an exam on a subsequent exam occurs through two channels. The first is a negative relationship between past and present exams induced as a grader learns the systematic score s. This relationship dissipates as more exams are evaluated and the grader discovers the shape of the distribution. A second channel of influence acts through the constraint of the grader. Suppose an exam in a particular period earns a high grade $g > \overline{g}$. This assignment will either bind or fail to bind the quota constraint. In the subsequent period, if the constraint is binding, the grader will assign lower grades on average since she can no longer award a high grade.

Consider then the case that the constraint is not binding. The grader will assign a high grade only if it is sufficiently above the threshold so as to warrant the loss of freedom to assign another high grade in the future less the current period penalty due to accuracy loss. The functional threshold above which a score must reach in order for the grader to assign the high grade in a given period is a positive function of the number of previously assigned high grades. If the last period grade causes a rise in the number of assigned high grades, the subsequent functional threshold rises as well, and the grader should again be

less likely to award a high grade.

Part (ii) of the above proposition argues that the influence of a past exam does not decay as the intervening distance between past and present exams increases. The intuition for the non-decay of a past exam's influence—whether prompted through learning, or a quota constraint— is that a grader does not differentially treat proximal past exams compared to more distant past exams. For example, a Bayesian will weight all past observations equally when forming an inference. The ordering of past grades, and specifically past high grades, is unimportant then so long as the distribution of such grades is unchanged. Moreover, a grader who faces a quota constraint in a given period is indifferent as to the timing of how such a constraint was formed.

Proof of Proposition 3 (Heuristic). First note that we can use the Bellman equation to characterize the solution of this problem through a value function $V_t(.)$ for a generic period t. A value function expresses the maximal future utility for a decision-maker who has already assigned H_t high grades and has just observed the exam score ψ_t:

$$V_t(H_t, \psi_t) = \max_{\{g_t\}} \{U(g_t(H_t, \psi_t)) + E_t V_t(H_{t+1}, \psi_{t+1})\}$$
$$subject\ to: \ H_t \leq \overline{H}\ for\ all\ t \quad (9)$$

Given the value function from prior periods, a policy rule can be written such that the value function in any generic period is a function z of the period's state variable H_t and the idiosyncratic draw ψ_t: $V_t = z(H_t, \psi_t)$.

The decision-maker's policy can then be described as follows. First, the grader observes H_t and the realization of $\widehat{\psi}_t$. Next, the grader assigns a grade $g_t(H_t, \psi_t)$ based on H_t and ψ_t. There are four possible scenarios which dictate how a grade is assigned: (1) If the quota constraint is not binding, $H_t < \overline{H}$, and the estimated idiosyncratic score is $\widehat{\psi}_t \leq \overline{g}$, then the problem reduces to the unconstrained case where $g^* = \widehat{\psi}_t$. (2) If the quota constraint is not binding, $H_t < \overline{H}$, but the estimated idiosyncratic score is $\widehat{\psi}_t > \overline{g}$, then the decision-maker must maximize across the unconstrained case which yields expected period utility: $V_t(H_t + 1, \psi_{t+1})$ and the constrained case which yields expected period utility: $-(\overline{g} - \psi_t)^2 + V_t(H_t, \psi_t)$. (3) If the quota constraint is binding, $H_t = \overline{H}$, and $\widehat{\psi}_t > \overline{g}$, then the grader maximizes $U(g_t)$ subject to $g \leq \overline{g}$, and sets $g^* = \overline{g}$. (4) If the quota constraint is binding, $H_t = \overline{H}$, and $\widehat{\psi}_t \leq \overline{g}$, then the problem again reduces to the unconstrained case, $g^* = \widehat{\psi}_t$

To prove part (i), note that the influence of a past high grade on a current grade can be decomposed into a learning component, through changes in the inference of $\widehat{\psi}_t$, and a

change in the state variable and hence the value function $V_t(H_t, \psi_t)$. Consider three cases. Suppose first that a past exam does not increase the state variable even allowing for a small perturbation in the past grade.. In scenarios (1) and (4) above, $\partial g_t/\partial g_{t-1}|_{g_{t-i,i>1}} \leq 0$ follows from an argument of learning summarized in Proposition 1 and Corollary 1. The non-positive correlation between evaluations in scenario (3) reflects similar logic and can be shown by taking the appropriate derivative. In scenario (2), assuming $\widehat{\psi}_t > \overline{g}$ holds, the decision rule for the assignment of g_t is not a function of ψ_{t-1} and thus $\partial g_t/\partial g_{t-1}|_{g_{t-i,i>1}} = 0$. Analagous reasoning can be applied to show that $\partial g_t/\partial g_{t-1}|_{g_{t-i,i>1}} \leq 0$ holds for a second case where the past exam does alter the state variable but is far above the threshold $\overline{g}$ such that a small perturbation in g_{t-1} does not cause a movement across the threshold.

Suppose now that the past exam score is just at the threshold $\overline{g}$ so that a small perturbation of g_{t-1} prompts the assignment of a high evaluation which increases the state variable H_t. In a first condition, if the constraint is binding in the current period then one can show that $\partial g_t/\partial g_{t-1}|_{g_{t-i,i>1}} \leq 0$ since if $\widehat{\psi}_t > \overline{g}$ the grader must set $g^* = \overline{g}$, and if $\widehat{\psi}_t \leq \overline{g}$, the relationship follows from learning. The remaining condition to consider is if the past exam is assigned a high score at the threshold, but the constraint in the current period is not binding. If $\widehat{\psi}_t \leq \overline{g}$, then the grader sets $g^* = \widehat{\psi}_t$ and $\partial g_t/\partial g_{t-1}|_{g_{t-i,i>1}} \leq 0$ follows through learning. If $\widehat{\psi}_t > \overline{g}$, the comparative static reduces to scenario (2) above for a given H_t as compared to H_{t+1}. One can demonstrate the negative relationship across periods by showing that $\partial g_t/\Delta H_t \leq 0$ which is equivalent to showing $V_t(H_t, \psi_t) - V_t(H_t + 1, \psi_t) \leq V_t(H_t + 1, \psi_t) - V_t(H_t + 2, \psi_t)$. The latter inequality is true if $V_t(.)$ is non-increasing, and non-convex (or more technically has a property of "increasing differences"). We can easily show that $V_t(H_t, \psi_t) \geq V_t(H_t + 1, \psi_t)$ and thus that $V_t(.)$ is non-increasing since any sequence of solutions $\{g\}_t^T$ available under $V_t(H_t + 1, \psi_t)$ must also be available under $V(H_t, \psi_t)$. One can additionally argue that $V_t(.)$ must have the property of "increasing differences" through a proof by backwards induction on the two dimensions of quota size and the number of periods. This property corresponds to the intuition that the functional cutoff threshold that graders use to assign high grades in scenario (2) is increasing and monotonic in the number of high grades assigned H_t [(i)].

(ii) The order invariance of past high grades can be inferred from the decision rule stated above. Again note that the influence of a past high grade on a current grade can be decomposed into a learning component, through changes in the inference of $\widehat{\psi}_t$, and a change in the constraint and hence the value function $V_t(H_t, \psi_t)$. The order-invariance of the first channel follows from the proofs for Proposition 1 and Corollary 1. A Bayesian weighs all past observations equally. The order-invariance of the second channel follows from the decision-rule outlined above. While decisions to assign g_t may be sensitive to the number of high grades assigned H_t, decisions are insensitive to the order in which such

grades occur. Specifically, in scenario (2) above, the critical comparison involves $V_t(H_t, \psi_t)$ and $V_t(H_t + 1, \psi_t)$. The value function reflects the order-invariance explicitly as it is a function of only ψ_t and the state variable H_t.[(ii)].

9 Appendix - Derivation of the Model

Estimation of $\hat{s}$ in the Standard Model. A decision-maker sequentially observes $\tilde{q}_1, ..., \tilde{q}_t$, where $\tilde{q}_t = \psi_t + s + \varepsilon_t$. Given $\psi \sim N(0, 1/\lambda_\psi)$, $\varepsilon \sim N(0, 1/\lambda_\varepsilon)$, $1/\lambda = 1/\lambda_\psi + 1/\lambda_\varepsilon$, this series of signals forms a random sample from a normal distribution with unknown mean s $(-\infty < s < \infty)$ and known variance $1/\lambda > 0$. The decision-maker uses Baye's rule to update his estimate of $\hat{s}_t$ from each successive observation of $\tilde{q}_t$. Suppose that the prior distribution of the parameter s, $\xi(s)$, is a normal distribution with mean s_0 and variance $1/\gamma$, then the estimated mean of the posterior distribution of s after t exams, $\hat{s}$, is given by:

$$\hat{s}_t = \frac{\gamma}{\gamma + \lambda t} s_0 + \frac{\lambda t}{\gamma + \lambda t} \overline{\tilde{q}_t}$$

Further, the variance ϕ of such a distribution can be expressed as:

$$\phi = var(\xi(s \mid \tilde{\mathbf{q}})) = \frac{1}{\gamma + \lambda t} \tag{10}$$

Proof. The proof for the expression of the estimated mean and variance of a posterior distribution with normal priors is standard and follows Degroot (2002).

Proof of Proposition 0. The proof for (i) follows directly from taking the derivative [(i)]. To prove (ii) note first that $\frac{\partial g_t}{\partial q_t} = \frac{\partial g_t}{\partial \tilde{q}_t} * \frac{\partial \tilde{q}_t}{\partial q_t}$, and $\frac{\partial \tilde{q}_t}{\partial q_t} = 1$. Next note that $g_t = \hat{\psi}_t = \tilde{q}_t - [\frac{\lambda}{\lambda+\gamma t} s_0 + \frac{\gamma t}{\lambda+\gamma t} \overline{\tilde{q}_t}]$, thus, $\frac{\partial g_t}{\partial q_t} = 1 - (\frac{\gamma}{\lambda+\gamma t})$. Since $0 < (\frac{\gamma}{\lambda+\gamma t}) < 1$, the $\frac{\partial g_t}{\partial q_t} > 0$ [(ii)]. The proof for (iii) follows from (i) and (ii) [(iii)].

Proof of Proposition 1. The proofs for (i), and (ii) follow directly from taking the partial derivative of interest and then taking the limit as $t \to \infty$. Note first that $\frac{\partial \tilde{q}_t}{\partial q_t} = 1$ so one can equivalently rewrite the proposition using partial derivatives with respect to $\tilde{q}_t$ as opposed to q_t. Now note that $g_t = \hat{\psi}_t = \tilde{q}_t - [\frac{\lambda}{\lambda+\gamma t} s_0 + \frac{\gamma t}{\lambda+\gamma t} \overline{\tilde{q}_t}]$ so $\left|\frac{\partial g_{t-k}}{\partial \tilde{q}_{t-k-1}}\right| = \frac{\lambda}{\gamma+(t-k)\lambda} < \left|\frac{\partial g_{t-l}}{\partial \tilde{q}_{t-l-1}}\right| = \frac{\lambda}{\gamma+(t-l)\lambda}$ for $k < l < t$ and $\lim_{t\to\infty} \frac{\partial g_t}{\partial \tilde{q}_{t-1}} = \lim_{t\to\infty} \frac{\partial \hat{\psi}_t}{\partial \tilde{q}_{t-1}} = \lim_{t\to\infty} -\frac{\lambda}{\gamma+t\lambda} = 0$ [(i)], and $\frac{\partial g_t}{\partial \tilde{q}_{t-k}} = -\frac{\lambda}{\gamma+t\lambda} = \frac{\partial g_t}{\partial \tilde{q}_{t-l}}$ [(ii)].

Proof of Corollary 1. The proofs for (i) and (ii) critically rely on the assumption that all prior evaluations are held constant except for the evaluation of interest. Accordingly, a perturbation in the evaluation at period $(t-j)$ must be the result of a perturbation in perceived quality $\tilde{q}_{t-j}$. Therefore while $\frac{\partial g_{t-k}}{\partial g_{t-k-1}} = f(\tilde{q}_{t-k-1}, \tilde{q}_{t-k-2}, ..., \tilde{q}_1)$, $\frac{\partial g_{t-k}}{\partial g_{t-k-1}}|_{g_{t-i-1,i>k}} = f(\tilde{q}_{t-k-1}).= \frac{-\lambda}{\gamma+(t-k)\lambda}$. Similarly, $\frac{\partial g_{t-l}}{\partial g_{t-l-1}}|_{g_{t-i-1,i>l}} = \frac{-\lambda}{\gamma+(t-l)\lambda}$. Since $\frac{-\lambda}{\gamma+(t-l)\lambda} < \frac{-\lambda}{\gamma+(t-k)\lambda} < 0$ it follows that $\frac{\partial g_{t-l}}{\partial g_{t-l-1}}|_{g_{t-i-1,i>l}} < \frac{\partial g_{t-k}}{\partial g_{t-k-1}}|_{g_{t-i-1,i>k}} < 0$ for $k < l < t$. Finally, the $\lim_{t\to\infty} \frac{\partial g_t}{\partial g_{t-1}}|_{g_{t-i,i>1}} = \lim_{t\to\infty} -\frac{\lambda}{\gamma+t\lambda} = 0$ [(i)]. By a similar reliance on the constancy condition, $\frac{\partial g_t}{\partial g_{t-k}}|_{g_{t-i,i>k}} = \frac{\partial g_t}{\partial \tilde{q}_{t-k}} = -\frac{\lambda}{\gamma+t\lambda} = \frac{\partial g_t}{\partial \tilde{q}_{t-l}} = \frac{\partial g_t}{\partial g_{t-l}}|_{g_{t-i,i>l}}$ [(ii)].

Proof of Proposition 2. The proof for part (i) follows directly from definition and then taking the derivative: $\partial \tilde{q}_t^c / \partial q_{t-1}^c = -(1-\alpha)\delta_1 < 0$ since $\delta_1 > 0$ [(i)]. Now assume $\alpha < 1$. Note that $g_t^c = \hat{\psi}_t^c = \tilde{q}_t^c - \hat{s}_t^c$ allows one to decompose the partial derivative $\frac{\partial g_t^c}{\partial q_{t-1}}$ into a perceptual and learning component: $\frac{\partial g_t^c}{\partial q_{t-1}} = \frac{\partial \hat{\psi}_t^c}{\partial q_{t-1}} = \frac{\partial \tilde{q}_t^c}{\partial q_{t-1}} - \frac{\partial \hat{s}_t^c}{\partial q_{t-1}}$. Taking the partial derivative of each expression yields $\frac{\partial \hat{s}_t^c}{\partial q_{t-1}} = \frac{\partial \hat{s}_t^c}{\partial \tilde{q}_{t-1}^c} * \left(\frac{\partial \tilde{q}_{t-1}^c}{\partial \tilde{q}_{t-1}} * \frac{\partial \tilde{q}_{t-1}}{\partial q_{t-1}} \right)$. Since $\frac{\partial \tilde{q}_{t-1}^c}{\partial \tilde{q}_{t-1}} > 0$, $\frac{\partial \tilde{q}_{t-1}}{\partial q_{t-1}} > 0$, and $\frac{\partial \hat{s}_t^c}{\partial \tilde{q}_{t-1}^c} = \frac{\lambda}{\gamma+\lambda t}\left(1 + \frac{\partial \tilde{q}_t^c}{\partial \tilde{q}_{t-1}^c}\right) > 0$ by argument and definition, then $\frac{\partial \hat{s}_t^c}{\partial q_{t-1}} > 0$. This confirms intuition—as any score increases, the estimated systematic score across all students should also increase. Additionally, given $\frac{\partial \tilde{q}_t^c}{\partial q_{t-1}} = \left(\frac{\partial \tilde{q}_t^c}{\partial \tilde{q}_{t-1}^c} * \frac{\partial \tilde{q}_{t-1}^c}{\partial \tilde{q}_{t-1}} * \frac{\partial \tilde{q}_{t-1}}{\partial q_{t-1}} \right) < 0$ by argument and definition, it follows that $\frac{\partial g_t^c}{\partial q_{t-1}} < 0$ [(ii)].

Assume, without loss of generality, that $\alpha = 0$. Applying the reasoning above one can decompose each partial derivative $\frac{\partial g_t^c}{\partial q_{t-k}}$ into a perceptual component and a learning component: $\frac{\partial g_t^c}{\partial q_{t-k}} = \frac{\partial \tilde{q}_t^c}{\partial q_{t-k}} - \frac{\partial \hat{s}_t^c}{\partial q_{t-k}}$. Solving for the first expression (perception) gives: $\frac{\partial \tilde{q}_t^c}{\partial q_{t-k}} = \frac{\partial \tilde{q}_t^c}{\partial \tilde{q}_{t-k}^c} * \left(\frac{\partial \tilde{q}_{t-k}^c}{\partial \tilde{q}_{t-k}} * \frac{\partial \tilde{q}_{t-k}}{\partial q_{t-k}} \right)$ which, by definition can be simplified to: $\frac{\partial \tilde{q}_t^c}{\partial q_{t-k}} = \frac{\partial \tilde{q}_t^c}{\partial \tilde{q}_{t-k}^c} = -[\delta_1 \frac{\partial \tilde{q}_{t-1}^c}{\partial \tilde{q}_{t-k}^c} + \delta_2 \frac{\partial \tilde{q}_{t-2}^c}{\partial \tilde{q}_{t-k}^c} + ... + \delta_k \frac{\partial \tilde{q}_{t-k}^c}{\partial \tilde{q}_{t-k}^c}]$ for $k > 0$. Given the specified lag structure, $\delta_k = \delta_1^k$, for $k = 1$, $\frac{\partial \tilde{q}_t^c}{\partial q_{t-k}} = -\delta_1$, and for $k > 1$, since the coefficients for the partial derivatives cancel, $\frac{\partial \tilde{q}_t^c}{\partial q_{t-k}} = 0$.

The second expression (learning) can be expressed as: $\frac{\partial \hat{s}_t^c}{\partial q_{t-k}} = \frac{\partial \hat{s}_t^c}{\partial \tilde{q}_{t-k}^c} * \left(\frac{\partial \tilde{q}_{t-k}^c}{\partial \tilde{q}_{t-k}} * \frac{\partial \tilde{q}_{t-k}}{\partial q_{t-k}} \right) = \frac{\partial \hat{s}_t^c}{\partial \tilde{q}_{t-k}^c}$ where $\frac{\partial \hat{s}_t^c}{\partial \tilde{q}_{t-k}^c} = \frac{\lambda}{\gamma+\lambda t}\left(\frac{\partial \tilde{q}_t^c}{\partial \tilde{q}_{t-k}^c} + ... + \frac{\partial \tilde{q}_{t-k}^c}{\partial \tilde{q}_{t-k}^c} \right)$. Note that for $k = 1$, $\frac{\partial \tilde{q}_t^c}{\partial \tilde{q}_{t-k}^c} = -\delta_1$ and therefore, $\frac{\partial \hat{s}_t^c}{\partial \tilde{q}_{t-k}^c} = \frac{\lambda}{\gamma+\lambda t}(-\delta_1 + 1)$. For $k > 1$, $\frac{\partial \tilde{q}_t^c}{\partial \tilde{q}_{t-k}^c} + \frac{\partial \tilde{q}_{t-1}^c}{\partial \tilde{q}_{t-k}^c} + ... + \frac{\partial \tilde{q}_{t-k+2}^c}{\partial \tilde{q}_{t-k}^c} = 0$,and $\frac{\partial \tilde{q}_{t-k+1}^c}{\partial \tilde{q}_{t-k}^c} = -\delta_1$, so again $\frac{\partial \hat{s}_t^c}{\partial \tilde{q}_{t-k}^c} = \frac{\lambda}{\gamma+\lambda t}(-\delta_1 + 1)$. Therefore, for any $k > 0$, the learning component $\frac{\partial \hat{s}_t^c}{\partial \tilde{q}_{t-k}^c}$ is $0 < \frac{\lambda}{\gamma+\lambda t}(-\delta_1 + 1) < 1$. The intuition underlying this result is that the degree of updating at period t due to some past score is invariant to how far in the past the score was assessed.

Recall that $\frac{\partial g_t^c}{\partial q_{t-k}} = \frac{\partial \tilde{q}_t^c}{\partial q_{t-k}} - \frac{\partial \hat{s}_t^c}{\partial q_{t-k}}$. When $k = 1$, $\frac{\partial g_t^c}{\partial q_{t-k}} = -(1 - \frac{\lambda}{\gamma+\lambda t})\delta_1 - \frac{\lambda}{\gamma+\lambda t}$ since $\frac{\partial \tilde{q}_t^c}{\partial q_{t-k}} = -\delta_1$ and $\frac{\partial \hat{s}_t^c}{\partial \tilde{q}_{t-k}^c} = \frac{\lambda}{\gamma+\lambda t}(-\delta_1 + 1)$. When $k > 1$, then $\frac{\partial g_t^c}{\partial q_{t-k}} = -\frac{\lambda}{\gamma+\lambda t}(-\delta_1 + 1)$. Therefore, when $k < l$, $\left| \frac{\partial g_t^c}{\partial q_{t-k}} \right| \geq \left| \frac{\partial g_t^c}{\partial q_{t-l}} \right|$[(iii)]. Part (iv) of the proposition follows from parts (i) and (ii) above and then taking the limit as $\alpha \to 1$[(iv)].

Proof of Corollary 2. The proofs for (i), (ii) and (iii) critically rely on the assumption that all prior evaluations are held constant except for the evaluation in the period of interest.

Accordingly, the perturbation in the evaluation at period $(t-1)$ must be the result of a perturbation in the perceived quality of the exam during the same period, $\tilde{q}_{t-1}$. Therefore, $\frac{\partial g_t^c}{\partial g_{t-1}}|_{g_{t-i,i>1}} = \frac{\partial g_t^c}{\partial \tilde{q}_{t-1}} = \frac{\partial g_t^c}{\partial q_{t-1}}$. By part (ii) of proposition 2, $\frac{\partial g_t^c}{\partial g_{t-1}}|_{g_{t-i,i>1}} = \frac{\partial g_t^c}{\partial q_{t-1}} < 0$ [(i)]. Parts (ii) and (iii) follow from the same reasoning, as well as the proof of Proposition 2.[(ii), (iii)].

10 Appendix - Tables

Appendix Table 1

SUMMARY STATISTICS FOR GRADING EVALUATIONS

	ALL YEARS					2002		2003		2004		2005	
	N	MEAN	SD	MIN	MAX	N	MEAN	N	MEAN	N	MEAN	N	MEAN
STUDENTS	1139					289		286		276		288	
PS TOTAL	1139	39.6	7.1	5	50	289	40.2	286	39.9	276	39.5	288	38.7
MIDTERM TOTAL	1134	144.9	23.4	54.5	191	289	147.1	281	143.2	276	145.0	288	144.2
FINAL													
Multiple Choice	1139	22.9	6.2	2	40	289	26.3	286	21.1	276	20.5	288	23.6
Total	1139	137.4	29.6	17	194	289	149.7	286	138.0	276	135.1	288	126.9
BY ORDER													
0 to 100													
Finals MC	400	23.3	6.2	8	40	100	26.5	100	21.6	100	20.2	100	24.7
PS Total	400	39.1	7.8	0	50	100	39.6	100	39.3	100	39.0	100	38.4
Midterm Total	400	145.1	24.3	55	191	100	145.1	99	144.3	100	145.1	100	145.7
101 to 200													
Finals MC	400	22.3	6.3	2	38	100	25.5	100	20.8	100	20.4	100	22.3
PS Total	400	39.8	7.1	5	50	100	40.5	100	40.9	100	39.5	100	38.3
Midterm Total	400	143.4	24.2	61	190	100	146.6	99	142.9	100	144.5	100	139.6
201 +													
Finals MC	339	23.2	5.9	6	38	89	27.1	86	20.9	76	20.9	88	23.6
PS Total	339	39.8	6.5	14	50	89	40.5	86	39.3	76	40.0	88	39.5
Midterm Total	330	146.4	21.2	74	188	89	149.8	83	142.2	76	145.5	88	147.8
TOTAL GRADE	1139	72.3	12.0	25	95	288	70.6	276	71.5	286	71.8	289	75.1

Appendix Table 2

OVERVIEW OF PENNSYLVANIA JUDICIAL COURTS

COURT LEVEL	CASE TYPES	NUMBER OF JUDGES
Magisterial District Courts	Traffic, minor criminal cases, civil cases under $8,000, and bail hearings, arraignments and preliminary hearings for more serious criminal offenses	> 500
Court of Common Pleas	Civil cases, serious criminal offenses, appeals from lower courts, matters involving children and family	93 judges
Superior Court/ Commonwealth Court	Appeals from lower courts, original civil cases brought against the State	35 judges
Supreme Court	Appeals from lower courts	7 judges

Notes: Counties outside of Philadelphia and Pittsburgh have additional courts at the level of the MDCs (not listed above). Data gathered from sources on the PA state court website, as well as the AOPC website.

Appendix Table 3

OVERVIEW OF MAJOR CHARGE CATEGORIES IN PA DISTRICT COURTS

CATEGORY	%	DESCRIPTION	ROLE IN ANALYSIS
Summary Offense	37	Traffic or minor non-traffic offenses with less than $300 fine or 90 days imprisonment	Dependent Variable
Preliminary Hearing	14	Hearing to determine whether sufficient evidence exists to charge defendant with a criminal offense.	Priming Event if Criminal
Arraignment	10	Hearing where defendant is informed of criminal charges against him, and defendant enters a plea of guilty or not guilty	Priming Event if Criminal
Bail Hearings	0.5	Hearing to determine whether criminal defendant can be released on bail prior to trial and the amount of such bail	Priming Event if Criminal
Civil Disposition & Landord-Tenant Dispute	11	Non-criminal case in which private party sues another for remedy	Not used

Note: Categories describe hearings in district courts for PA counties outside of Philadelphia and Pittsburgh. Definitions collected from various publications from PA state court websites, as well as the AOPC websites. Statistics on case percentages calculated from 2001 - 2005 sample.