Is an Improving Grade Sequence Preferable to a Better Grade?

Linda H. Moya

Carnegie Mellon University

The relationship between the trend of a sequence of related outcomes and judged assessments of fairness of those outcomes was studied in the context of an educational framework. Specifically, a short course was modeled as a controlled experiment, in which participants completed several "assignments" and, a "final", for which they received an overall "course grade". After the course grade was received, participants judged the fairness of various aspects of the course – whether or not they felt the course grade was fair, the final test was fair, the process was fair, and whether or not the course grade reflected their actual ability. The main conclusions are that given two grade sequences with the same overall average, participants significantly prefer the sequence in which the trend increases (their grades improve) from beginning to end; furthermore, given a sequence in which the trend increases with a lower overall average grade versus a sequence in which the trend decreases with a higher overall average grade, participants still prefer the sequence in which the trend increases. However, this latter result is not significant when controlling for performance. In short, depending on performance, the process may matter more than the outcome, that is, an improving grade sequence may be preferable to a better grade.

I. Introduction

The judged fairness of a process or sequence of a series of related outcomes has been the subject of extensive study in the social sciences. Within the field of decision research, two such streams of research are intertemporal choice, which deals with valuing today the trend of a series of outcomes extending into the future, and experienced utility, which deals with the assessment of a series of outcomes that have occurred in the past. Within the education literature, the importance of fairness in process is framed in terms of procedural justice, while outcome fairness is framed in terms of distributive justice. Like experienced utility, the focus of the education and related literature has been on the assessment of a series of outcomes that have occurred in the past. The present research brings together these three bodies of literature to evaluate and explain a representative actual experience in terms of the three theories.

Specifically, the present research studies a controlled experiment that models students taking a college-level course. The assessment mechanisms for such courses typically include periodic graded assignments and graded tests, which are then factored into an overall course grade. The course grade conveys to the student how well they performed in the course. In terms of decision research, the assignments and the final comprise a sequence of outcomes, and in retrospect, the student will undoubtedly have a judged perception of fairness of those outcomes.

In the fall of 2003 the author of this paper taught a short course approximately seven weeks long, and utilized typical assessment tools: several assignments and a final test. The students' course grade was comprised of the weighted average of these outcomes. Interestingly, two students who received the very good grade of A- for the course were nonetheless extremely unhappy with the grade, and presumably with the process that produced the grade. Reflecting on their dismay, I theorized that a key reason behind their unhappiness was that their course grade was heavily influenced by their grade on the final test, which, for these two, was lower than the grades they had received on the assignments. In terms of decision research, they had experienced

a decreasing trend in the series of course outcomes, and had the trend been increasing instead, they would have been much happier with the same course grade.

To test my theory, I designed an experiment designed to emulate the salient aspects of a school course, while enabling increased experimental control, relative to an actual course, on variables that could have relevance in the retrospective judgment of fairness. The detailed experiment design is provided in section IV. I also reviewed the three research streams outlined above (intertemporal choice, experienced utility, and the education and related literature dealing with fairness of assessment mechanisms) for the purpose of developing hypotheses and explaining my results in terms of the academic theories. The theory and hypotheses are given in section III. The experiment analysis is given in section V. First, however, is a literature review of the key findings in the three research streams.

II. Literature Review: Intertemporal Choice, Experienced Utility, and Fairness of Assessment Mechanisms

The literature on intertemporal choice and experienced utility is complimentary in significant ways, but also qualitatively different. Intertemporal choice is a decision between choices extending into the future, while experienced utility is a rating of a past experience or set of experiences (Ariely and Loewenstein, 2000).

Varey and Kahneman (1992) distinguish between three concepts of utility: *decision utility* is the value that the decision maker assigns to a consequence in a decision context; *experienced utility* is the actual hedonic value of a consequence after it has occurred; *predicted utility* represents the decision maker's beliefs about the experienced utility of a possible consequence. The authors specifically look at whether people correctly incorporate their hedonic beliefs into their decisions, that is, given a decision now, do people make decisions in such a way as to incorporate the dimensions of the experience that actually turn out to be important to them in the future, when that decision is realized? The results of their experiments demonstrated no suggestion that participants attempted to perform a utility-integration calculation in evaluating the experience. Rather participants used the heuristic of focusing on one or a few selected moments as representative for the whole.

Huber et al. (1997) consider how tradeoffs might change from the time of choice to the time of experience, then from the time of experience to later consumption of memories of that experience. The authors argue that salient alternatives at the time of choice often become much less so later. On the day of choice, you are in a particular reference state. Later, on the day of retrospective evaluation, you are likely in another. Predicting one's own adaptation to new reference states can be difficult. The authors argue that at the time of choice, adaptation effects are difficult to merge into one's decision function because, by definition, the adaptation has not occurred.

These two papers demonstrate that intertemporal choice and experienced utility are qualitatively related, yet different in key respects, in terms of Varey and Kahneman (1992), decision utility is not equal to experienced utility. Therefore, I will first review the key findings in intertemporal choice and experienced utility separately, and then summarize them together.

A. Intertemporal choice

Research on intertemporal choice typically uses hypothetical choice scenarios to elicit participants' preferences for sequences of outcomes that extend into the future. The primary

focus has been choice in the monetary or health domains, although other domains have occasionally been studied. The findings are that often, people prefer increasing sequences of outcomes, especially with respect to monetary income, although in the health domain the opposite is often true. Furthermore some research results are contrary to this finding. A key point, however, is that increasing sequences imply negative time preference utility discounting, which stands in stark contrast to traditional economic discount utility theory (DUT) which in turn assumes positive time preference utility discounting in all life domains.

First, consider the evidence that a majority of people prefer sequences of outcomes that rise in desirability through time. Loewenstein and Sicherman (1991) found that all else being equal people prefer increasing over declining or flat monetary profiles of greater present value. Furthermore, they found that people's preferences for rising sequences are stronger for one's salary than for rental income. Hsee and Abelson (1991) evaluated peoples satisfaction with sequences of outcomes with respect to gambling debts/income, class rank and stock. They found that satisfaction with an outcome is positively related not just to the position (i.e. actual level) of the outcome, but also to the displacement (i.e. directional difference between the current level and the reference level). Read and Powell (2002) conducted choice experiments in three monetary domains (one-year salary earnings, one-year lottery winnings, and lifetime earnings) to understand why, when it comes to money, subjects choose a sequence with an increasing trend at the expense of maximizing present value. They found the reasons to be correlated in a sensible way with the choices made. For example, subjects thought about money as if they would spend it at the rate it was received, hence the reason given for choosing an increasing sequence was *appropriateness*: they chose a sequence that corresponded to the shape of their anticipated consumption needs. Another reason commonly given was the *expectedness* of the sequence. In a finding similar to Loewenstein and Sicherman (1991), there was a stronger effect for salary versus lottery winnings in preferring a rising sequence for the year period.

Although there is ample evidence a majority of people prefer sequences of monetary outcomes that rise in desirability through time, Guyse et al. (2002) found otherwise. They considered business school students' preferences for temporal sequences of environmental outcomes (choose a policy related to air quality and near-shore ocean water quality), and compared those with preferences for monetary (choose payment option to be paid by a partner in a business venture) outcomes. Given choices of rising, constant or increasing sequences, they found that participants preferred constant or increasing sequences of air quality and near-ocean water quality. And, in contrast to Loewenstein and Sicherman (1991) and Read and Powell (2002), they found that when it comes to income from the business venture, these business students prefer decreasing sequences of income.

There is also evidence of domain effects, that is, while generally preferring increasing sequences in the monetary domain, people generally prefer decreasing sequences in the health domain. Chapman and Elstein (1995) found that discount rates for health are positive while discount rates for money are negative. Chapman (1996) found that expectations and length of time mediate the effect of preferred sequence direction. Chapman studied three health domains: headache pain, facial acne and facial wrinkles. She found that people expect health to decrease over lifetime and thus prefer such a sequence. In cases where decision makers have a strong expectation of decline (e.g. 20-year sequence of facial wrinkles), a preference for declining sequences appears. In contrast people expect wages to rise over their lifetime and correspondingly prefer such a sequence. For short periods, such as up to a year, expectations for health and money are similar, in that preferences are for increasing sequences. Read and Powell

(2002) found that for lifetime sequences, participants preferred rising sequences for lifetime earnings, and falling sequences for health. Hence, both Read and Powell (2002) and Chapman (1996, 2000) argue that it is *expectation* or sense of *appropriateness* that mediates people's preferences. People expect or feel that it is appropriate that their monetary income rise through time, and they expect that their health will deteriorate through time. van der Pol and Cairns (2000) found that while a majority of people demonstrate positive time preference in the health domain, it is still a significant number that do demonstrate negative time preference. Specifically, those that perceived the health state to be more dire, had negative time preference. van der Pol and Cairns (2001) elicited intertemporal preferences for one's own health as compared to preferences for others' future health. They found that individuals exhibited positive time preference. Furthermore, the time-preference rates for others' health were higher (more positive) than the rates for one's own health.

In contrast to the general finding that people generally prefer decreasing sequences of health (Chapman, 1996; Read and Powell, 2002), Guyse et al. (2002) found that business students prefer constant or increasing sequences in qualities of health. Like Read and Powell (2002) and Chapman (2000), in the health domain they found a positive relationship between the shape of the expectation of the most likely sequence and the shape of the ideal sequence found.

The frame matters, in the preference for increasing sequences in any domain only occurs if the outcomes are seen as logically related in a sequence. Loewenstein and Prelec (1991) propose that negative time preference is applied selectively, to those events that are seen as part of a meaningful sequence, having a well-defined starting and ending point. It is the integrity of the sequence which cause rise to negative time preference, the desire to either get unpleasant events out of the way as quickly as possible, coupled with the desire to save the most desirable events to the end. Chapman and Elstein (1995) discovered that implicit discount rates are lower for outcomes embedded in a series as compared with individual outcomes. Loewenstein and Prelec (1993) found that when outcomes were viewed together as a sequence, people prefer utility levels that improve over time. That is, people prefer to postpone better outcomes to the end. Without a proper sequence context, however, people discount outcomes in isolation and in doing so they demonstrate positive time preference with utility levels decreasing over time. The authors suggested that differences between time preferences for individual outcomes versus sequences may be explained as a type of framing effect.

There is also a great deal of individual variation in preferences. Dolen and Gudex (1995) had participants evaluate outcomes of health-related quality of life policies. The researchers found that modal time preference rate is zero but there are a number of responses which imply very high – positive and negative – discount rates. Overall, however, more subjects had negative discount rates than positive discount rates.

Variables such as duration, delay, degree and expectations influence preferences. MacKeigan et al. (1993) propose a theory of intertemporal choice which posits that "anticipation utility" (vivid, fleeting, future events) attenuates positive time preference in the health domain. The authors found that delayed health gains are devalued more than health losses of comparable duration. Furthermore, they found the devaluation predicted by positive time preference was increasingly attenuated as health loss duration decreased.

In summary of the intertemporal choice literature, there is evidence that a majority of people prefer sequences of outcomes that rise in desirability through time, especially in the monetary domain (Loewenstein and Sicherman, 1991; Hsee and Abelson, 1991; Chapman & Elstein, 1995; Read and Powell, 2002), although Guyse et al. (2002) found this not to be the

case. There is evidence of domain effects, that is, while preferring increasing sequences in the monetary domain, people prefer decreasing sequences in the health domain (Chapman, 1996; van de Pol and Cairns, 2000, 2001). Both Read and Powell (2002) and Chapman (2000) argue that it is *expectation* or sense of *appropriateness* that mediates people's preferences. People expect or feel that it is appropriate that their monetary income rise through time, and they expect that their health will deteriorate through time. The frame matters, in the preference for increasing sequences in any domain only occurs if the outcomes are seen as logically related in a sequence (Loewenstein and Prelec, 1991). Variables such as duration influence preferences. For example, for short durations (e.g. 1 year), people prefer increasing sequences in both the money and health domains, (MacKeigan et al, 1993; Chapman, 1996). It is over much longer durations, (e.g. one's lifetime), that people tend to prefer or expect the decreasing sequence of health outcomes, while at the same time preferring that monetary income follow an increasing sequence.

B. *Experienced utility*

In contrast to research on intertemporal choice which typically uses hypothetical choice scenarios to elicit preferences, the majority of research on experienced utility has employed experiments where participants actually underwent the experience they would later assess. Like intertemporal choice the research has focused on few domains. In the case of experienced utility the research focus has been on unpleasant experiences such as listening to uncomfortably loud noises or watching unpleasant film clips, and actual experiences of pain.

Two related theories are proposed to characterize key findings in experienced utility, the peak-end and duration neglect theories. However, research since the theories' proposals has demonstrated that these do not explain all the evidence. The peak-end theory states that in evaluating a past sequence of related experiences the two most salient aspects that characterize the evaluation is the momentary experience contributing to peak affect, and the end experience. The theory of duration neglect states people neglect duration when evaluating an experience. Fredrickson and Kahneman (1993) had subjects view aversive (amputation footage) and pleasant (penguins at play) film clips that varied in duration and intensity. Subjects provided real time ratings during each clip and retrospective evaluations when the clip was over. The authors found that the retrospective evaluations were effectively determined by a weighted average of the peak affect rating and the final rating recorded for each film; the duration of the film did not emerge as an overall predictor of the overall evaluation. Redelmeier et al. (2003) evaluated experienced pain in patients undergoing colonscopy. They found that overall retrospective memory was created by recalling selected moments rather than an exact running total of experience. The duration of the episode had relatively small influence unless it was highly salient, or correlated with intensity.

Ariely and Carmon (2000) found that peak did not play a role in patients with chronic as opposed to acute pain. In a field study of bone-marrow pain transplant patients. they found that the final (end) and the change in the hour-by-hour ratings were predictive of the overall evaluation, but that the most intense (peak) was not. They explain these results by noting that these patients were long term, chronic patients, and that it is likely that they had previous experience as bad or worse pain than their peak pain during the study.

In a review of the literature, Fredrickson (2000), in reflecting on the peak-end utility theory of experiences, suggests that what will become the peak are the moments rich with self-relevant information. The theory most applies with episodes that are clearly bounded, continuous and completed. Specifically, when the end is not known, end affect (more aptly

called recent affect) appears to contribute little to global evaluations, which was a result of her study of anticipated social endings (Frederickson, 1991 as cited in Fredrickson, 2000). Secondly, when episodes are particularly directed toward an end goal, end affect may be all that matters because it comes to symbolize the end of the activity.

As in intertemporal choice (Loewenstein and Prelec, 1991, 1993) framing does matter. Ariely and Zauberman (2000) show that the way subjects summarize an experience depends on whether the experience is composed of single or multiple parts. They show that in the unpleasant experience of listening to grating saw-toothed waveform sounds, subjects' preference for improvement was higher if the experience was perceived as continuous, and substantially reduced if perceived as being composed of discrete parts. They found that once an experience is segmented, its overall evaluation shifts in the direction of the experiences' mean.

Researchers disagree on the value of remembered utility such as that resulting from a peak-end analysis. Kahneman (2000) argues that remembered evaluations are not good measures of an experience and that a moment-based approach, consisting of measurements taken during the experience, provides better measures. Kahneman (2000) argues that experienced utility is best measured by moment-based methods that assess current experience. Moment-based methods record the experience periodically while it is happening (see for example, Redelmeier et al. 2003). He asserts that such methods are more "objective" and that remembered utility deserves "less respect" because it results in illogical choices such as that to extend a painful sequence in order to achieve a lower peak-end rating overall (cf. Kahneman, 1993; Redelmeier and Kahneman, 1996). However a decade earlier, Beese and Morley (1993) argued that memory for acute pain is generally reliable, if not specifically accurate. Patients that had their wisdom teeth removed filled out standard pain questionnaires immediately after the surgery, and two weeks later when they were pain free. In comparing a weighted rank of the two sets of questionnaires, the authors found no shift in the mean. They interpreted this result to mean that memory for pain is generally reliable.

Bolton (1999) argues that on-going measures are impractical in clinical settings regarding chronic pain, and that a single-point pain rating scale that measures "usual" or "on average" pain periodically provides a reliable measure. She states that problematically, the literature on evaluating past experiences of pain has focused on acute pain, and that more research needs to be done on chronic pain. To test her theory, she investigated the pain profiles of chronic musculo-skeletal pain patients over a period of 7 days. She found that the single "on average (usual)" measure on day 8 was an accurate estimate of "actual average" pain intensity over a recording interval of 7 days.

In summary of the experienced utility literature, two related theories are proposed to characterize key findings in experienced utility, the peak-end an duration neglect theories. However, research since the theories' proposals has demonstrated that these theories do not explain all the evidence. The peak-end theory states that in evaluating a past sequence of related experiences the two most salient aspects that characterize the evaluation is the momentary experience contributing to peak affect, and the end experience. The theory of duration neglect states people evaluate the experience neglecting duration as a measure (Fredrickson and Kahneman, 1993; Redelmeier et al. 2003). Fredrickson (2000) found that when the end is not known, then end affect contributes little to global evaluations after the fact. She goes on to discuss that when episodes are directed to an end goal, end affect may be all the matters. In findings paralleling that of Loewenstein and Prelec (1991, 1993) in the intertemporal choice literature, Ariely and Zauberman (2000) found that the way people summarize an experience

depends on whether the experience is perceived as continuous or as disjoint. If the experience is disjoint, the end evaluation tends to shift in the direction of the episodes' mean, rather than reflecting the peak-end episodes. Kahneman (2000) argues that remembered evaluations are not good measures of an experience and that a moment-based approach, consisting of measurements taken during the experience, provides better measures. However a decade earlier, Beese and Morley (1993) argued that memory is a generally reliable measure of experienced acute pain. Bolton (1999) argues that on-going measures are impractical in clinical settings regarding chronic pain, and that a single-point pain rating scale is a good measure. She states that problematically, the literature on evaluating past experiences of pain has focused on acute pain, and that more research needs to be done on chronic pain. Taken together, this literature suggests that while moment-based measures are valuable in many domains, they may not provide the best approach in all domains, for example, in the domain of experienced chronic pain.

C. Summary: Intertemporal choice and experienced utility

Intertemporal choice research typically uses hypothetical choice scenarios, whereas experienced utility research has primarily had subjects actually undergo the experience which they are to assess. The intertemporal choice literature has focused largely on the domains of money and health, whereas the experienced utility literature has assessed evaluations of unpleasant experiences and pain. Even within the realm of pain, far less research has been done on chronic pain than on acute pain, the difference which has implications on the assessment tools of experienced utility (moment-based approaches versus retrospective or single-point evaluations). Key findings in intertemporal choice are that sequence trend direction matters, as does duration, and that there are many times when people prefer negative time discounting (improving sequences), a finding which stands in stark contrast to traditional economic discount utility theory (DUT) which in turn assumes that people prefer positive time discounting (declining sequences) in all life domains. Key findings in experienced utility are that in many instances neither trend nor duration matters in final assessments, but select moments, notably the peak experience (the most personally salient, or the most painful in an experience of pain) and the end experience wholly determine the final assessment. Furthermore, much research has illuminated circumstances when the peak-end hypotheses and duration neglect do not apply. A key issue in experienced utility is which assessment tools provide the 'best" way to measure experienced utility. Clearly the literature on intertemporal choice and experienced utility are related in significant ways, but they differ to varying degrees in the domains researched, the methods used, and in key results.

D. *Fairness of assessment mechanisms*

Fairness of assessment mechanisms in education

Research has found that students expect and believe they deserve better grades than they actually get. Murstein (1965) conducted an experiment that extended the duration of an entire course. Students predicted their final grade at the beginning of the course, after two mid-term exams were given, and right before the final exam was given. All students started out expecting high grades (A or B). In general students showed no significant change in their predictions as a function of experience during the school semester. The only significant difference was between the predicted grades and the actual final grade; with the latter invariably lower than the predictions in all cases where the latter was not the highest grade possible. Unless the highest

grade was received, it was always below both students' expectations and the grade they thought they deserved. Several decades later, Wendorf (2002) examined students' grade expectations three different times during a semester: within the first week of the semester, midway through, and within the week just prior to the final exam. Wendorf found that relative to actual grade point, expected grades were about a grade higher. Although expected grades decreased over the course of the semester, they remained higher than the actual grade received.

Research has also found that the grading process is more salient when students' grades are lower than expected. Hull (1980) found that students who perceived grades to be unfair also questioned the grading process. Furthermore, students who perceived teachers to be unfair in grading tended to receive lower grades, and students that indicated stronger agreement that teachers were fair in their grading practices tended to receive higher grades. Tata (1999) found the grading process to be more salient when the grade was lower than expected, as compared to when the grade is as good as or better than expected.

Some research lends support to the idea that in at least some cases, the grading process is *more* important than the grade itself. Wendorf and Alexander (2005) examined the relative influence of perceptions of fairness and expected grades on students' satisfaction with their instructors and their grades. They found that perceptions of fairness explained a larger part of the variance of student satisfaction with their grade; than did the actual grade itself.

Certainly process and outcome are intricately related. Rodabaugh and Kravitz (1994) found that, as determined by teacher ratings, people are more accepting of outcomes that do not benefit them if they perceive that the procedures used to determine the outcomes are fair.

Finally, Lackey (1980) and Kaufman (1981) in their discovery that differences in fairness perceptions exist between different academic disciplines underscore the fact that while overall theories of fairness may provide a valuable evaluating framework, domain effects do occur. Lackey (1980) found that fairness in grading had greater impact on the explanation of students' evaluations of their teachers in mathematics than in biology and sociology. Kaufman (1981) found that art majors indicated a lower level of fairness as an important quality than did computer science majors and psychology majors. Art students also considered the ideal teacher to be one who graded less accurately than did psychology or computer science students.

Given that in many cases students' grades received are a full letter grade below their expectations and what they think they deserve, (Murstein, 1965; Wendorf, 2002), and that the grading process is more salient when students' grades are lower than expected (Hull, 1980; Tata, 1999), it is clear that the grading process is as important as the actual grade. Furthermore, the fact that perceptions of fairness explain a larger part of the variance of student satisfaction than the actual grade itself (Wendorf and Alexander, 2005), lends support to the idea that perhaps in at least some cases the grading process is *more* important than the grade itself. Clearly, process and outcome are intricately related, to measure one without the other will not provide the full picture: Radabaugh and Kravitz, (1994) found that people are more accepting of outcomes that do not benefit them if they perceive that the procedures used to determine the outcomes are fair. Finally, Lackey (1980) and Kaufman (1981) in their discovery that differences in fairness perceptions exist between different academic disciplines underscore the fact that while overall theories of fairness may provide a valuable evaluating framework, domain effects do occur.

Fairness of assessment mechanisms in other domains

When one's outcome of an assessment experience can be compared to a standard or to comparison others' outcomes, then the process leading to that assessment is less salient in the affective feelings toward the overall experience. Messe and Watts (1983) conducted studies to

determine the role of established standards, versus social comparison, in subjects' judgments of fairness and satisfaction. They found that social comparison was found to affect subjects' fairness judgments when their pay was low compared to an agreed upon standard. Social comparison did not appreciably affect fairness judgments when subjects were paid the standard amount. Their results also found that absolute amount of pay had a greater impact on evaluations of satisfaction than on fairness judgments. van den Bos et al. (1997a) found that when people do not have information about others' outcomes they use procedural fairness to assess how to react to their outcome, but they rely less on procedure information when they are informed about the outcome of another person. Also, when outcomes are worse than that of comparison others, procedure has no effect on judgments.

However, when outcomes are not as expected (either better or worse), people use process fairness as a heuristic to inform them on how to react. van den Bos et al. (1998) found that when people receive outcomes that are better or worse than expected, they use procedural fairness as a heuristic substitute to assess how to react to their outcome but that they rely less on procedure information when they receive outcomes that are equal to, better than, or worse than those of comparison others.

If the process is inconsistent, but nonetheless benefits them, it is not considered any less fair than a consistent process. Ployhart and Ryan (1998) found that positive inconsistency as compared with consistency generally resulted in similar perceptions of fairness. They also found that the most negative reactions occurred when the process was unfair but the outcome was fair.

However, Tyler (1994) found process to be more significant in influencing affect than the outcome in legal experiences such as appearances in court, calls to the police for help, or being stopped by police; and in work experiences with supervisors Tyler found procedural justice to be the primary justice judgment influencing affect, although distributive justice influence also occurred.

van den Bos et al. (1997b) found that whether people found procedural justice to be more important than the actual outcome, or vice versa, depended on the order in which people were told information about the two. Specifically, when people were told about the procedure first and then the outcome, the procedure weighed more heavily in their perceptions of fairness. In contrast, when people were told about the outcome prior to the procedure use, the outcome weighed more heavily in their perceptions of fairness.

Finally, there exists support in industry fields for the aforementioned theories of procedural and distributive fairness. Greenberg (1986) surveyed managers in the cable TV, wholesale pharmaceutical and credit union industries, as part of their participation in management training seminars. Participants were asked to think of an incident in which they received either a particularly fair or unfair performance evaluation. In the end, the participants classified the groups into two broad categories. The first category included attributes such as consistent application of standards. The second category included attributes such as receipt of rating based on performance achieved and recommendation for salary and/or promotion based on the rating. Greenberg draws the parallel between the content of these derived categories and theories of procedural and distributive fairness.

In summary, when one's outcome of an assessment experience can be compared to a standard or to comparison others' outcomes, then the process leading to that assessment is less salient in the affective feelings toward the overall experience (Messe and Watts, 1983; van den Bos et al., 1997a; van den Bos et al., 1998). However, when outcomes are not as expected (either better or worse), people use process fairness as a heuristic to inform them on how to react (van

den Bos et al. 1998). If the process is inconsistent, but nonetheless benefits them, it is not considered any less fair than a consistent process (Ployhart and Ryan, 1998). However, Tyler (1994) found process to be more significant in influencing affect than the outcome, and van den Bos et al. (1997b) found that whether people found the process to be more important than the actual outcome, or vice-versa depended on the order in which they were told information about the two. Which ever they were told about first mattered more in their subsequent perceptions of fairness. Furthermore, industry field support exists for the theories of procedural and distributive fairness, as found in input from middle managers in the cable TV, wholesale pharmaceutical and credit union industries (Greenberg, 1986).

E. *Implications for the present research*

In the behavioral decision theory literature, there is evidence that people prefer increasing sequences of related outcomes in several domains, at the same time preferring decreasing sequences of outcomes in others. Research suggests that what compels people to prefer an increasing versus decreasing sequence of outcomes is guided by what they think is appropriate, or what they expect. In general people think that it is appropriate for salaries to rise throughout their lifetime, and given the choice, would prefer an increasing salary sequence. People expect that their health will decline over the years and thus given the choice, prefer a decreasing sequence of health. However, once a sequence of outcomes has past, the trend of the sequence has not been found to be a significant predictor of peoples' experienced utility. Rather it is two other aspects of the sequence that determine the experienced utility: the outcome or momentary experience that provided the peak affect, and the outcome at the end of the sequence.

In the present research, I consider both decision utility and experienced utility theories and apply them in a new domain not addressed previously by either: education. This is significant, because domain effects have been shown to exist with respect to both theories: neither theory explains all the evidence. Specifically I consider the sequence of grade outcomes in the context of an academic course. I consider the possibility that people prefer a sequence of grades that increase throughout the course and that this drives their perceptions of fairness, and I consider the possibility that the peak-end hypothesis explains their perceptions of fairness.

Consistent with the methods of the experienced utility literature, I conduct an experiment in which participants actually experience the sequence of outcomes they are to rate. However, unlike the acute experiences researched in the experienced utility literature, one's experience in education can be long term, and thus may exhibit features associated with chronic experiences.

The education literature has not approached evaluating fairness of assessment processes in terms of decision or experienced utility. Rather it has evaluated fairness in the context of procedural and distributive justice. Findings in the education literature and other related literature on fairness of assessment processes, are that both process and outcome matter. In some cases process may even matter more, that is, may be a larger determining factor of fairness perceptions than the outcome itself.

The theory, experiment and analytical methods used in the present research are framed in terms of decision and experienced utility, yet results are consistent with findings in the education and related literature on the fairness of assessment processes, that is, the outcome matters, but sometimes the process matters even more.

III. Theory, Hypotheses and Exploratory Analysis

The experiment presented in this paper tests the theory of preferences for increasing sequences of outcomes from the intertemporal choice literature, using the key method from the experienced utility literature of having participants actually experience the sequence of outcomes for purposes of subsequently rating that experience. In rating the experience subjects rated their perceptions of fairness, which is a key theoretical focus of the education literature.

The experiment was designed to emulate the salient aspects of a school course, while enabling increased experimental control, relative to an actual course, on variables that could have relevance in the retrospective judgments of fairness. Three grading schemes where devised, one which effected a decreasing sequence of outcomes (decreasing grades over time), a second which effected an increasing sequence of outcomes (increasing grades over time) with the same expected average overall outcome as the first, and a third which also effected an increasing sequence of outcomes, but with a lower average overall outcome than the first two. Each grading scheme was tested in multiple, separate experimental sessions. Participants entered a classroom setting. The experimenter acted as the teacher figure who explained the activities with reference to a syllabus which detailed the grade weighting of each assignment and final. Participants were informed of the grading criteria for the assignments and final prior to undertaking each, and of their grade after each were completed. After all the assignments and final were completed, participants were informed of their overall course grade, after which they filled out a questionnaire to assess their perceptions of fairness about the process and the outcome.

The theory of preference for increasing sequences suggests that participants who experience a decreasing trend in the sequence of course outcomes should be less satisfied than those who experience an increasing trend of course outcomes even in the case that the average overall course grade is the same in the two grading schemes:

> ***Hypothesis 1***: Given two grade sequences in which the first sequence is decreasing and the second increasing, but with equivalent average difficulty, students will rate the second sequence significantly fairer than the first sequence.

Furthermore, it is possible that an increasing trend is so valued, that it compensates for an otherwise lower outcome. That is, there exist two sequences such that one would rate a sequence with a decreasing trend and higher final outcome less fair than a sequence with an increasing trend and lower final outcome:

> ***Hypothesis 2***: Students will rate an increasing grade sequence as being more fair than a decreasing grade sequence, even when the increasing-grade sequence is more difficult on average and even when it leads to a significantly lower grade.

The results here are also considered in terms of theories of the experienced utility literature. As previously noted, this literature has as the key method, having participants undergo an experience, and then rate that experience. The rating, in a word, constitutes their satisfaction or utility with that experience. In this experiment however, subjects rated whether or not they perceive the experience to have been *fair* in process and in outcome. Subjects did not rate *satisfaction* with their grade outcome. The focus on fairness perceptions makes sense given that the experimental domain is education, and in the educational domain, perceptions of fairness surrounding grading procedures are of significant theoretical importance at the same time satisfaction with such procedures is relatively not. Given that fairness and not satisfaction measures were taken of the grading experience in this experiment, it is not possible to apply

experienced utility theories - which theorize about satisfaction - directly. Nonetheless, it is instructive to explore just how the experienced utility theories might apply in the educational domain and to the theoretically important issues of perceived fairness of grades and grading processes in that domain. Such an exploration is done in the results section after Hypothesis 1 and 2 have been discussed.

IV. Methods

Participants. Sixty-four students at Carnegie Mellon University and the University of Pittsburgh and other individuals signed up to participate in a cartoon experiment for pay. They were told they would have the opportunity to test their skill in identifying differences between two versions of the same cartoon in a fixed amount of time, they would repeat the exercise for a series of five cartoons, and they would be paid according to the number of differences found across all five. Participants received up to $12 for participating.

Materials. I created the stimuli from five different cartoons I obtained the academic right to use from cartoonstock.com. I created second versions of each cartoon containing twenty-five differences from the original using Adobe Photoshop®. The stimuli are given in Appendix A. The criteria for choosing the stimuli was that 1) the content be stimulating to such individuals who choose to participate, 2) participants be motivated to give their best performance, 2) participants view their performance as a measure of their ability, 3) performance be clearly measurable, and 4) grading be fully objective across all conditions. The cartoon stimuli and grading methods chosen meets these criteria. Other materials included a syllabus which listed the cartoon names and their respective weight contribution to the final grade, and instructions for the task; description sheets in which participants listed the differences found (one description sheet per cartoon pair); grade scale charts to inform participants the grading criteria for a given cartoon pair; a questionnaire to assess perceptions of fairness after all cartoon trials were completed; and a grade sheet which reported participants' grades.

Procedure. Experimental sessions were conducted approximately every two hours over the course of four days with between two and six people in each session. The experimental design had as the goal to set up a classroom-like environment in which there was a teacher-figure (the experimenter), a syllabus, several students (the participants), graded assignments (the first four cartoons), a graded final (the last cartoon), and a course grade. Prior to evaluating the cartoons, participants were given the syllabus. The first four cartoons in the sequence were worth 12.5% each and the last cartoon was worth 50%. The evaluation and grading process was fully explained and all questions were answered prior to commencing the experiment trials.

Respondents were then given the five cartoon pairs to evaluate in succession. Prior to each cartoon trial, participants were told the grading criteria for that trial, that is, how many differences would have to be found to receive an A+, A, A-, B+, etc. Participants had six minutes per cartoon to identify differences between the two versions. Unknown to the participants, there were twenty-five differences between the two versions of each cartoon. A stopwatch was used to insure all participants in a given session started and stopped at the same time. After each cartoon trial, individual results were tallied and grades reported. After all five cartoons were evaluated, the grades were tallied to an overall course grade for the session. The overall course grade was then reported to each participant. The participants then filled out a

questionnaire to evaluate their perception of their overall course grade and the grading process. They where then paid for their time at the rate of $0.65 per grade-point-squared[1].

Subjects participated in one of the three conditions (between subjects design across sessions). In the *lenient-hard* condition, the first four cartoons were graded leniently and the final cartoon was graded hard. In the *hard-lenient* condition, the grading was reversed: the first four cartoons were graded hard and the fifth cartoon graded leniently. In the *really hard-relatively lenient* condition, the first four cartoons were graded very hard, and the fifth cartoon graded relatively leniently. Recall that participants had six minutes per cartoon to identify differences between the two versions. Pilot testing indicated that an average of 11 differences would be found in six minutes. The *lenient-hard* condition was designed to elicit a decreasing sequence of outcomes. The *hard-lenient* condition was designed to elicit the opposite: an increasing sequence of outcomes. The *really hard-relatively lenient* condition was designed to elicit an increasing sequence but with an lower average overall grade than the other two conditions. Table 1 summarizes the grading criteria. Recall that prior to each cartoon trial, participants were told the grading criteria for that trial. This information was given to them in charts similar to that in Table 1. Importantly, the expected grade-point average was the same in the first two conditions and lower in the third condition. For example, if participants were to average 11 in each cartoon, the expected grade-point averages in the first two conditions would be 3.5[2,3], and the expected grade-point average in the third condition would be 3.0[4].

Condition					
1) Lenient-hard condition		Cartoons 1-4	Cartoon 5		
2) Hard-lenient condition		Cartoon 5	Cartoons 1-4		
3) Really hard-relatively lenient condition				Cartoons 1-4	Cartoon 5
Grade	Grade Point	# found	# found	# found	# found
A+	4.3	**10+**	20+	25	14
A	4.0	9	18-19	22-24	12-13
A-	3.7	8	16-17	20-21	**11**
B+	3.3	7	14-15	18-19	10
B	3.0	6	12-13	15-17	8-9
B-	2.7	5	**10-11**	13-14	7
C+	2.3	4	8-9	**11-12**	6
C	2.0	3	6-7	8-10	4-5
C-	1.7	2	4-5	6-7	3
D	1.0	1	2-3	3-5	1-2
F	0	0	0-1	0-2	0

Table 1

[1] I chose the formula to insure a compensation spread between $0 and $12, with average expected compensation in the *lenient-hard* and *hard-lenient* conditions of approximately $8, and an average expected compensation in the *really hard-relatively lenient* condition of approximately $6, that is, so that subjects would receive compensation for their time consistent with typical compensation in comparable experiments.

[2] Lenient-hard condition: 4x4.3x0.125+1x2.7x0.5 = 3.5

[3] Hard-lenient condition: 4x2.7x0.125+1x4.3x0.5 = 3.5

[4] Really hard-relatively lenient condition: 4x2.3x0.125+1x3.7x0.5 = 3.0

V. Findings

A. Tabulated results

Out of the sixty-four participants, I dropped one participants' result as an outlier[5]. There remained twenty participants in the *lenient-hard* condition, twenty-one participants in the *hard-lenient* condition, and twenty-two participants in the *really hard-relatively lenient* condition, for a total of sixty-three participants. The average course grade point in the *lenient-hard* condition was 3.68, and the average course grade point in the *hard-lenient* condition was 3.84, which are significantly different from each other ($t=-2.46$, $p<0.02$). The average course grade point in the *really hard-relatively lenient* condition was 3.44, which is significantly different from both grade point averages in the *lenient-hard* and *hard-lenient* conditions (respectively: $t=2.72$, $p<0.00$; $t=5.32$, $p<0.00$). See Table 2.

Condition	**Cartoons 1-4 Grades**		**Cartoon 5 Grade**		**Course Grade**	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
1) Lenient-hard	4.28 *	0.07	3.07 *	0.47	3.68 *	0.26
2) Hard-lenient	3.38 *	0.33	4.30 *	0.00 [6]	3.84 *	0.17
3) Really hard–relatively lenient	2.77 *	0.32	4.10 *	0.35	3.44 *	0.31
* between groups: pairs within the same column that are significantly different at the 5% level [7]						

Table 2

Consider that course grade is correlated with condition by design: the *really hard-relatively lenient* condition had a lower expected course grade than both the *lenient-hard* and *hard-lenient* conditions even though the stimuli used were identical across all three conditions. On the other hand, precisely because identical stimuli were used across all three conditions, performance is not correlated with condition by design. Performance per cartoon trial is the number of differences found between the two versions of the cartoon. Overall raw performance is the average of the performance on each of the five cartoon trials. Overall course performance, henceforth *performance*, is the weighted average of the performance on each of the five cartoons, with performance on the first four worth 12.5% each, and performance on the last cartoon worth 50%. Overall raw performance and course performance by condition is given in Table 3.

[5] The outlier z-score =-2.86. When included in the analysis, it significantly changed the results of hypothesis 2. I will elaborate further at the point hypothesis 2 is discussed.

[6] All participants received the highest grade of 4.3 in this condition, on the last cartoon.

[7] ttests assuming unequal variances used when both samples were approximately normally distributed but the Std. Dev. differed by a factor of 2 or more (cartoons 1-4 conditions 1&2 and 1&3), the kryskal-wallis equality of populations rank test and two-sample wilcoxon rank-sum test were used where at least one sample was not normally distributed (cartoon 5, conditions 1&2 and 2&3), ttests assuming equal variances were used otherwise (cartoon 1-4 conditions 1&3; cartoon 5 conditions 1&3; and course grade in all conditions).

Condition	Cartoons 1-4 Performance		Cartoon 5 Performance		Raw Performance		Performance	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
1) Lenient-hard	9.94 * #	0.23	12.90 *	2.80	10.53 * #	0.68	11.42 *	1.47
2) Hard-lenient	14.73 *	1.95	10.00 * #	0.00 [8]	13.79 *	1.56	12.37 *	0.97
3) Really hard–relatively lenient	14.36 #	2.28	13.09 #	1.52	14.10 #	2.04	13.72 *	1.74
*, # between groups: pairs within the same column that are significantly different at the 5% level[9]								

Table 3

The fact that performance differs significantly between conditions is particularly interesting. Since the stimuli and experimental procedures used across all three conditions were identical – only the grading scheme changed – random assignment of participants should have insured that ability, hence performance based on ability, not significantly differ between the conditions. I argue that the differential standards set for a given grade elicited differential motivation thus performance across the conditions. Although reflecting in hindsight it is reasonable that differential motivation would occur during the sessions, I did not expect any actual difference to be significant hence did not account for it in the experimental design. Anecdotally however, I observed very differential motivation thus performance[10]. Since the differences did indeed turn out to be significant, this phenomena can be scientifically explored via experiments controlling for performance in the analysis of fairness, in future research.

Participants' fairness perceptions of the grading process and course grade were reported on a one page questionnaire of four five-point scale statements to which the participant specified level of agreement (-2 strongly disagree to 2 strongly agree). The four questions were "My final grade was fair", "The grading process was fair", "The final grade accurately reflects my ability", and "The grade scale on the last cartoon was fair", henceforth referred to as *gradefair,*

[8] Performance measures were estimated from the final grade considering the condition. In this condition on this last cartoon, all final grades were the highest possible, with no variation, hence estimated performance has no variance.

[9] ttests assuming unequal variances used when both samples were approximately normally distributed but the Std. Dev. differed by a factor of 2 or more (cartoons 1-4 conditions 1&2 and 1&3; and rawperformance conditions 1&2 and 1&3), the kryskal-wallis equality of populations rank test and two-sample wilcoxon rank-sum test were used where at least one sample was not normally distributed (cartoon 5, conditions 1&2 and 2&3), ttests assuming equal variances were used otherwise (cartoon 1-4 conditions 1&3; cartoon 5 conditions 1&3; rawperformance conditions 2&3, and performance in all conditions).

[10] During the sessions, differential motivation, thus performance was obvious. It appeared that participants strived to identify only the number of differences necessary to receive an A+. In the *lenient-hard* condition, in which only 10 differences were necessary to earn an A+ for the first four cartoons (each worth 12.5%), I observed that participants did not rush to start with the stop-watch, they switched pens or pencils, they wrote longer descriptions of the differences, and were otherwise more "laid-back" during the whole process. They thus did not have the experience of working faster during the fifth cartoon (worth 50%), in which 20 differences were necessary to earn an A+. The opposite occurred in the *hard-lenient* condition, in which 20 differences were required to earn an A+ for the first cartoons and only 10 was necessary to earn an A+ for the fifth cartoon. For each successive cartoon I observed improved performance (more differences found), say for example, 10 for the first, 13 for the second, 15 for the third, and 17 for the fourth cartoon. By the fifth cartoon, participants in the *hard-lenient* condition easily found the 10 differences to earn the A+. In the *really hard-relatively lenient* condition, I observed a competitive and achievement oriented atmosphere. Participants poised to quickly start with the stop watch, they worked more quickly, and only wrote the minimum necessary to specify the difference. I observed that participants in the harder conditions exhibited more stream-lined personal working processes, egged on by motivation to do what was necessary to earn an A+.

processfair, gradeaccurate and *lastcartoonfair*, respectively. The means and standard deviations of the four assessment statements are given in Table 4.

Condition	**gradefair**		**processfair**		**gradeaccurate**		**lastcartoonfair**		***fair***	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
1) Lenient-hard	0.55 *	0.94	0.65 *	1.14	-0.15	1.31	-0.25 * #	1.25	0.20 * #	0.95
2) Hard-lenient	1.33 *	0.97	1.33 *	1.02	0.62	1.20	0.67 *	1.28	0.99 *	0.81
3) Really hard–relatively lenient	1.00	0.69	1.18	0.73	0.09	1.06	0.86 #	1.13	0.78 #	0.65
*, # between groups: pairs within the same column that are significantly different at the 5% level in 2-sample t-tests										
Scale										
Strongly Disagree				Neither Agree Nor Disagree					Strongly Agree	
-2	-1			0			1		2	

Table 4

Across all three conditions no one **strongly** disagreed with the statements that their grade was fair or that the process was fair. It is interesting to note that on average, the last cartoon which figured 50% of the course grade was considered to be fairer in the *really hard-relatively lenient* condition than in the *hard-lenient* condition, when in fact it was graded more leniently in the latter and they both had increasing sequences; however the difference is not significant.

Cronbach's alpha reliability coefficient of the four-item scale is 0.77, indicating that the items comprise a reasonably reliable assessment scale. I conducted a factor analysis to determine whether the items were univocal in their assessment of the fairness construct. A scree test indicated a large break between the first and second factors. Considering two factors, rotated loadings for all four items on the first factor ranged between 0.54 and 0.85, while loadings on the second factor ranged between 0.07 and 0.32. Hence I retained one factor, and correlated it with a simple average of the four items: the correlation was high at 0.97. Therefore I used the simple average of the four items as a single measure of fairness, and gave it the name *fair*. The last column in Table 4 presents the means and standard deviations of *fair*.

B. *Key relationships between condition, performance, grade, and perceptions of fairness*

To consider key relationships between the conditions, performance, grade, and perceptions of fairness, I conducted several linear regressions with condition represented by two dummy variables (hardlenient = 1 for the *hard-lenient* condition and 0 for the other two conditions, and rhardrlenient = 1 for the *really hard-relatively lenient* condition and 0 for the other two conditions), grades represented by overall grade point average, performance as discussed in the paragraph preceding Table 3 [11], and perceptions of fairness represented by the *fair* variable defined above.

[11] The analysis here uses weighted performance as discussed in the paragraph preceding Table 3. However regressions using raw performance were similarly significant.

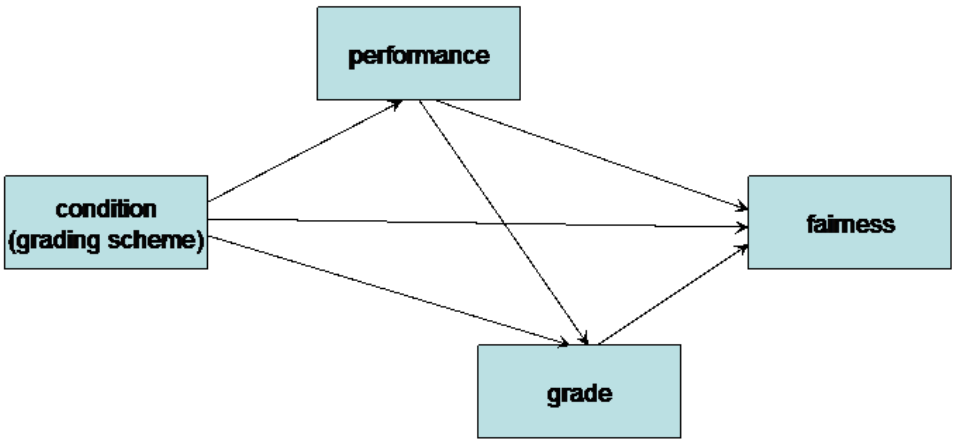


Figure I

The key relationships are graphically summarized in Figure I: condition predicts performance. This is discussed below with respect to equation [1]; condition and performance together predict grade (discussed below with respect to equation [2]); condition also predicts perceptions of fairness (discussed with respect to equations [3], [4], and [5]); condition together with grade predicts fairness (equations [6] and [7]); and condition and performance together predict fairness (equation [8]).

Performance and Grades

First, condition significantly predicts performance in [1]. Performance in the *hard-lenient* versus *lenient-hard* conditions, and between the *really hard-relatively lenient* versus *lenient-hard* conditions is significantly different. This result replicates the results shown in the last column of Table 3 and discussed in the subsequent paragraph.

Condition predicts performance

	performance	*constant*	*hardlenient*	*rhardrlenient*	
[1]		11.42	0.95	2.31	$F_{(2,60)}$=13.77,
		(0.32)	(0.45)	(0.44)	p<0.00, R^2 =0.31
		t=35.61	t=2.11	t=5.20	
		p<0.00	p<0.04	p<0.00	

Given that grades are based on performance, it is not at all surprising that performance and condition together significantly predict overall course grade in [2]. In fact the correlation between grade and performance by condition is greater than 99%. Grade is predicted to increase by 0.17 points for every additional difference found between the two versions of each cartoon.

Condition and performance predict grades

	grade	*constant*	*performance*	*hardlenient*	*rhardrlenient*	
[2]		1.69	0.17	-0.00	-0.64	$F_{(3,59)}$=2785.03
		(0.03)	(0.00)	(0.01)	(0.01)	p<0.00, R^2 =0.99
		t=62.67	t=75.29	t=-0.02	t=-66.95	
		p<0.00	p<0.00	p<0.98	p<0.00	

Hypothesis 1

Condition alone predicts perceptions of fairness. Average *fair* is significantly different in both equation [3], which evaluates the subset of data from the *hard-lenient* and *lenient-hard* conditions, and equation [4], which evaluates the subset of data from the *really hard-relatively lenient* and *lenient-hard* conditions. Furthermore the results still hold when considering the data from all three conditions together in [5]. In short, participants subjected to either the *hard-lenient* or *really hard-relatively lenient* grading schemes had significantly higher fairness perceptions (0.79 and 0.58 points respectively) than participants subjected to the *lenient-hard* grading scheme. These results replicate the results given in the last column of Table 4.

Condition predicts perceptions of fairness

	fair	*constant*	*hardlenient*	*rhardrlenient*	
[3]		0.20 (0.20)	0.79 (0.28)		$F_{(1,39)}$=8.17 p<0.01, R^2 =0.17
[4]		0.20 (0.18)		0.58 (0.25)	$F_{(1,40)}$=5.50 p<0.02, R^2 =0.12
[5]		0.20 (0.18)	0.79 (0.25)	0.58 (0.25)	$F_{(2,60)}$=5.23 p<0.01, R^2 =0.15
		t=1.11 p<0.27	t=3.12 p<0.00	t=2.34 p<0.02	

It is also the case that condition together with overall course grade predicts perceptions of fairness. In equation [6], which considers the subset of data from the *hard-lenient* versus *lenient-hard* conditions, grade is not a significant predictor of fairness perceptions, but condition is: the difference in fairness perceptions is significantly 0.59 points higher in the *hard-lenient* versus *lenient-hard* condition. Equation [7], which considers data from all three conditions, is also significant. Grade continues to not be a significant predictor, and when controlling for grade, condition continues to predict fairness. Perceptions of fairness are significantly higher in the *hard-lenient* versus *lenient-hard* condition, and in the *really-hard relatively lenient* versus *lenient-hard* condition. It is very interesting that grade is not a significant predictor but condition is. Simply stated, the grading scheme influences perceptions of fairness more so than the actual grade.

Condition and grade predict perceptions of fairness

	fair	*constant*	*grade*	*hardlenient*	*rhardrlenient*	
[6]		-4.25 (2.36)	1.21 (0.64)	0.59 (0.29)		$F_{(2,38)}$=6.14 p<0.00, R^2 =0.24
		t=-1.80 p<0.08	t=1.89 p<0.07	t=2.05, p<0.05		
[7]		-1.52 (1.54)	0.47 (0.42)	0.71 (0.26)	0.70 (0.27)	$F_{(3,59)}$=3.92 p<0.01, R^2 =0.17
		t=-0.99 p<0.33	t=1.12 p<0.27	t=2.73 p<0.01	t=2.59 p<0.01	

Considering [3] and [5] which demonstrate that condition alone significantly predicts perceptions of fairness, and [6] and [7] which demonstrate that condition together with grade significantly predicts perceptions of fairness, hypothesis 1 - that given two grade sequences in which the first sequence is decreasing and the second increasing, but with equivalent average

difficulty, students will rate the second sequence significantly fairer than the first sequence - is supported. It is the case that the *lenient-hard* condition has a decreasing sequence of outcomes and the *hard-lenient* condition has an increasing sequence of outcomes. It is also the case that when controlling for grade there is a significant difference in perceptions of fairness between individuals subject to one versus the other condition.

Hypothesis 2

Since better performance is required in the *really hard-relatively lenient* condition versus the *lenient-hard* condition to get a given grade, it is not sufficient to control for grade to test hypothesis 2, which compares perceptions of fairness between these two conditions. Rather, to test hypothesis 2, I control for productivity differences across groups. Specifically I consider performance in lieu of grade as a predictor of perceptions of fairness. Equation [8], which evaluates the subset of data from the *really hard-relatively lenient* and *lenient-hard* conditions, is significant. However neither performance nor condition is independently predictive of perceptions of fairness. That is, differential performance between the two conditions is not a significant predictor, and when controlling for performance, participants subjected to the *really hard-relatively lenient* versus *lenient-hard* grading scheme did not have significantly different perceptions of fairness.

Condition and performance predict perceptions of fairness

	fair	*constant*	*performance*	*rhardrlenient*	
[8]		-0.70	0.08	0.40	$F_{(2,39)}$=3.25,
		(0.92)	(0.08)	(0.31)	p<0.05, R^2 =0.14
		t=-0.76	t=1.00	t=1.30	
		p<0.45	p<0.32	p<0.20	

Considering [4] and [5] that demonstrate condition alone significantly predicts perceptions of fairness, but that in [8] the *really hard-relatively lenient* versus *lenient-hard* conditions together with performance are not significant in predicting fairness, hypothesis 2 – that students will rate an increasing grade sequence as being more fair than a decreasing grade sequence, even when the increasing-grade sequence is more difficult on average and even when it leads to a significantly lower grade - receives mixed support. It is the case that the *lenient-hard* condition is a decreasing sequence of outcomes, and the *really hard-relatively lenient* condition is an increasing sequence of outcomes, and that the average grade of the former is significantly higher at 3.68 than the latter at 3.44 (from Table 2), yet students rated the latter significantly fairer at average 0.78 than the former at average 0.20 (from [4], [5] and Table 4)[12]. However, when controlling for performance, the difference in perceptions of fairness between the *lenient-hard* and the *really hard-relatively lenient* conditions is not significant, t=1.25, p<0.22.

C. *Comparison of the "assignment" grades with the "final" grade*

Now I consider the results in terms of the experienced utility literature which has as a prevailing method, having participants undergo an experience, and then rate that experience. This prevailing method is similar to that in this experiment, with the sole difference being that in

[12] If the outlier is included then [4] loses its significance: $F_{(1,41)}$=2.82, p<0.10. There is no longer a significant difference in perceptions of fairness between the *lenient-hard* and the *really hard-relatively lenient* conditions, and hypothesis 2 is not supported even when not controlling for performance.

this experiment participants rated the fairness of the experience versus their satisfaction with the experience. In satisfaction ratings of experienced utility, the peak-end theory predicts behavior. It is instructive to examine whether fairness ratings of an experience will manifest similar peak-end behavior.

Consider that according to Fredrickson and Kahneman (1993), the peak is defined as the experience with the peak affect, and the end is defined as the last experience of the sequence. In this experiment one might define peak and end several ways.

In an analogy to pain, peak could be defined as the lowest grade received on the first four cartoons while end could be defined as the grade received on the last cartoon. While it turns out this is a significantly predictive model [9], only the last cartoon grade is a significant predictor variable. Or peak could be defined as the *experience* of receiving the grade on the last cartoon (again since it alone is weighted the most at 50% of the course grade, it is likely to elicit the strongest affect). Similarly, end could be defined as the *experience* of receiving the overall course grade. In this case however, since overall course grade is just the average of the first four cartoon grades, averaged with the last cartoon grade, a regression with the last cartoon grade and the course grade as predictor variables is more transparently expressed as the regression with the average of the first four cartoon grades, and the last cartoon grade as the two predictor variables. This model in [10] is likewise significant, but once again, only the last cartoon grade is a significant predictor variable. Finally, the analysis thus far suggests that the last cartoon grade alone is the significant predictor of perceptions of fairness. Therefore consider the model in [11]. It is predicatively significant and the most parsimonious of the three potential models.

Some combinations of the four"assignment" grades and the "final" grade that predict perceptions of fairness

	fair	*constant*	*lowestfirst4grade*	*first4gradeaverage*	*lastcartoongrade*	
[9]		-.151 (1.17)	0.01 (0.16)		0.56 (0.20)	$F_{(2,60)}$=5.93, p<0.00, R^2 =0.17
		t=-1.28 p<0.20	t=0.04 p<0.97		t=2.74 p<0.01	
[10]		-1.27 (1.26)		-0.03 (0.19)	0.53 (0.20)	$F_{(2,60)}$=5.95, p<0.00, R^2 =0.17
		t=-1.01 p<0.32		t=-0.18 p<0.86	t=2.69 p<0.01	
[11]		-1.47 (0.62)			0.56 (0.16)	$F_{(1,61)}$=12.06, p<0.00, R^2 =0.17

Certainly there are other possible mappings of course rating data onto the peak-end theory, however, based on this brief exploratory analysis, the peak-end theory does not explain the evidence resulting from this research. Rather, in educational experiences, where the final assignment receives the largest weight in the overall course grade, the final assignment is the sole significant predictor of perceptions of fairness of process and of outcome.

VI. Discussion

It is the case that given two grade sequences in which the first sequence is decreasing and the second increasing, but with equivalent average difficulty, students will rate the second sequence significantly fairer than the first sequence. Furthermore, when performance is

controlled for, the difference is still significant. Students believe that the second sequence is significantly fairer than the first sequence.

The second hypothesis, that students will rate an increasing grade sequence as being more fair than a decreasing one, even when the former leads to a significantly lower grade, received mixed results. The hypothesis is supported when performance is not controlled in analysis; the latter which should have been unnecessary given that performance based on ability should have been statistically non-significant across all conditions. The fact that performance did significantly differ between condition leads to the idea that different course processes elicit differential performance. This finding provides an opportunity for further study. However, with respect to the present research, the second hypothesis is not supported once performance is controlled in analysis for the particular course sequences used.

Interestingly, the peak-end theory, which has been found to be a significant predictor of an individuals' feelings about experiences, was not a significant predictor of the participants' perceptions of fairness here. Significantly, perceptions of fairness are not the same as satisfaction, which may provide a better measure of affect. The reason perceptions of fairness and not satisfaction were measured in this research is that between fairness of and satisfaction with a grade or grading process, educators are more concerned with the former, and this research is concerned with modeling an educational course. In education, it not possible to have all students satisfied with their grades. In fact, considering the findings of Murstein (1965) and Wendorf (2002) that actual grade received is predictably found to be below grade expected and grade believed deserved, it is likely that all except those students that receive the very highest grades will be dissatisfied. Finally, it is interesting to consider Fredrickson's (2000) suggestion that when episodes are directed to an end goal, end affect may be all that matters. It is the case that all assignments were directed to the end goal of an overall course grade. Yet with respect to fairness perceptions, the course grade was not a significant driver, whereas the grade on the last cartoon was. Again the reasons for why this is the case, provide ample opportunity for further research.

VII. Contributions and Future Research

The present research considered foundational aspects from both the intertemporal choice literature and the experienced utility literature in theory and design. The theory behind the research has its roots in findings of the intertemporal choice literature: the hypothesis that people generally prefer increasing as opposed to decreasing sequences of outcomes. The experimental design has its roots in the methods of the experienced utility literature: participants in the experiment actually underwent the experience they later assessed. This research explored the domain of education, which has not previously been studied in the context of either intertemporal choice or experienced utility. Rather it has previously been studied in terms of the theories of procedural and distributive justice and fairness. An opportunity for future research is to evaluate, in the intertemporal choice / experienced utility framework, intra-educational domain differences that have been suggested to exist in the educational literature. In general, the domain differences appear to be between quantitative versus qualitative (or non-quantitative) academic disciplines. In terms of this dichotomy, the present research modeled the former in that results were easily objectively measurable. Future research could explore the intertemporal choice / experienced utility frameworks as applied to qualitative (or non-quantitative) academic disciplines.

Future research could explore the role of expectedness as introduced in the intertemporal choice literature (cf. Chapman, 2000; Read and Powell, 2002), in the preference for increasing sequences in the educational course context. In a similar vein, exploring the relationship of the grade expected and/or grade felt deserved to the grade received, would enable the opportunity to explore the differential salience of process as suggested in the education and related literature (cf. Hull, 1980; Rodabaugh and Kravitz, 1994; Tata, 1999; van den Bos et al., 1998), from within the intertemporal choice / experienced utility framework.

The present research did not tease apart the differential effect of process versus outcome on perceptions of fairness. The 4-item questionnaire used was found to measure a single underlying factor of overall fairness. Future research could lengthen the questionnaire with the goal of discriminating two different fairness factors, one for process and one for outcome. The relationship of these differential factors to affect could be explored.

One final comment on measuring experienced utility. Recall that Kahneman (2000) suggested that remembered utility deserves "less respect" as a measure of experienced utility than do ongoing momentary utility measurements during the actual experience. Bolton (1999) on the other hand pointed out that ongoing measurements during experiences may be impractical and may not lead to assessments significantly better than a single point rating taken at the end. With respect to the present research and the educational context: experienced utility as measured through remembered utility is what matters in terms of assessing students' perceptions of the course, fairness or otherwise. It is the remembered utility that will determine students' feelings about course, first when the course is completed, later when recalling that experience to make decisions about the future, and even later, when simply reflecting on the experience as part of one's past.

VIII. Acknowledgments

I would like to thank George Loewenstein, with whom discussions led to the critical aspects of the experimental design, especially the decision to explore the possibility that students might prefer an increasing grade sequence even at the expense of an overall better grade. Thanks also to Linda Babcock and Mike DeKay whose ideas for and guidance on analysis was invaluable, and to Don Moore for valuable comments received.

IX. References

Amin, M.E., "Gender as a Discriminating Factor in the Evaluation of Teaching", *Assessment & Evaluation in Higher Education*, 1994, 19(2), 135-143.

Ariely, D., "Combining Experiences over Time: The Effects of Duration, Intensity Changes and On-Line measurements on Retrospective Pain Evaluations", *Journal of Behavioral Decision Making,* 1998, 11, 19-45.

Ariely, D., Carmon, Z., "Gestalt Characteristics of Experiences: The Defining Features of Summarized Events", *Journal of Behavioral Decision Making,* 2000. 13(2). 191-201.

Ariely, D., Kahneman, D., and Loewenstein, G., "Joint Commentary on the Importance of Duration in Ratings of, and Choices Between, Sequences of Outcomes", *Journal of Experimental Psychology: General,* 2000. 129(4). 524-529.

Ariely, D. and Loewenstein, G., "When Does Duration Matter in Judgment and Decision Making?". *Journal of Experimental Psychology; General,* 2000, 129(4), 508-523.

Ariely, D., Loewenstein, G., & Prelec, D. "Coherent arbitrariness: Duration-sensitive pricing of hedonic stimuli around an arbitrary anchor", (Tech. Rep.), 1999, Cambridge, MA: Massachusetts Institute of Technology, Sloan School of Management.

Austin, W., McGinn, N. and Susmilch, C. "Internal standards revisited: eŁects of social comparisons and expectancies on judgments of fairness and satisfaction", *Journal of Experimental Social Psychology*, 1980, 16(5), 426-441.

Ariely, D. and Zauberman, G., "On the Making of an Experience: The Effects of Breaking and Combining Experineces on Their Overall Evaluation", *Journal of Behavioral Decision Making*, 2000, 13(2), 219-232.

Beese, A. and Morley, S. "Memory for acute pain experience is specifically inaccurate but generally reliable", Pain, 1993, 53, 183-189.

Bolton, J. E., "Accuracy of recall of usual pain intensity in back pain patients", *Pain*, 1999, 83, 533-539.

Chapman, G. B. "Expectations and Preferences for Sequences of Health and Money", *Organizational Behavior and Human Decision Processes,* 1996, 67, 59-75.

Chapman, G. B., "Preferences for Improving and Declining Sequences of Health Outcomes", *Journal if Behavioral Decision Making*, 2000, 13(2), 203-215.

Chapman, G. B., and Elstein, E. S., "Valuing the future: Temporal discounting of health and money", *Medical Decision Making*, 1995, 15(4), 373–386.

Cherry, B., Ordonez, L.D., and Gilliland, S.W., "Grade Expectations: The Effects of

Expectations on Fairness and Satisfaction Perceptions", *Journal of Behavioral Decision Making,* 2003, 16, 375-395.

Dolan, P. and Gudex, C., "Time Preference. Duration and Health State Valuations", *Heath Economics,* 1995, 4, 289-299.

Feather, N. T. "Effects of prior success and failure on expectations of success and failure", *Journal of Personality a nd Social Psychology*, 1966, 3(3), 287-298.

Frederick, S., "Time Preference and Personal Identity", in *Time and Decision: Economic and Psychological Perspectives in Intertemporal Choice,* edited by Loewenstein, G., Read, D. and Baumeister, R., 2002, NY: Russell Sage.

Fredrickson, B.L., "Anticipated endings: An explanation for selective social interaction" (Doctoral dissertation, Stanford University, 1990). Dissertation Abstracts International, 3, 1991, AAD91± 00818.

Fredrickson, B.L., "Exracting meaning from past affective experiences: The importance of peaks, ends, and specific emotions", *Cognition and Emotion",* 2000, 14(4), 577-606.

Fredrickson, B. L., and Kahneman, D., "Duration Neglect in Retrospective Evaluations of Affective Episodes", *Journal of Personality and Social Psychology,* 1993, 65, 45-55.

Greenberg, J., "Determinants of Perceived Fairness of Performance Evaluations", *Journal of Applied Psychology,* 1986, 71(2), 340-342.

Guyse, J. L., Keller, L. R., & Eppel, T., 'Valuing environmental outcomes: preferences for constant or improving sequences', *Organizational Behavior and Human Decision Processes,* 2002, 87(2), 253–277.

Hsee, C. K., and Abelson, R. P., "Velocity Relation: Satisfaction as a Function of the First Derivative over Time", *Journal of Personality and Social Psychology,* 1991, 60(1), 341-347.

Hsee, C. K., Salovey, P., and Abelson, R. P, "The Quasi-Acceleration Relation: Satisfaction as a Function of the Change of Velocity of Outcome over Time", *Journal of Experimental Social Psychology,* 1994, 30(1), 96-111.

Huber, J., Lynch Jr., J. G., Corfman, K., Feldman, J., Holbrook, M., Lehmann, D., Munier, B., Schkade, D. and Simonson, I., "Thinking about Values in Prospect and in Retrospect: Maximizing Experienced Utility", *Marketing Letters,* 1997, 8(3), 323-334.

Hull, R., "Fairness in grading: perceptions of junior high school students", *The Clearing House,* 1980, 53(7), 340-343.

Kahneman, D., "Reference points, anchors, norms, and mixed feelings", *Organizational Behavior and Human Decision Processes,* 1992, 51, 296-312.

Kahneman, D., "Evaluation by moments: Past and future", In *Choices, Values and Frames,*

edited by Kahneman, D., and Tversky, A., 2000, New York: Cambridge University Press.

Kahneman, D., "Experienced utility and objective happiness: A moment-based approach", In *Choices, Values and Frames,* edited by Kahneman, D., and Tversky, A., 2000, New York: Cambridge University Press.

Kahneman, D., Fredrickson, B. L., Schrieber, C. A., and Redelmeier, D. A., When More Pain is Preferred to Less: Adding a Better End", *Psychological Science,* 1993, 4(6), 401-405.

Kahneman D. and Tversky, A. "Prospect theory: an analysis of decision under risk", *Econometrica,* 1979, 47, 263-291.

Kahneman, D., Wakker, P. and Sarin, R., "Back to Bentham? Explorations of Experienced Utility", *Quarterly Journal of Economics,* 1997, 112(2), 375-405.

Kaufman, B.J., "Departmental Differences in Student Perceptions of "Ideal" Teaching", *Paper presented at the Annual Meeting of the Southeastern Psychological Association*, Atlanta, GA, March 26, 1981.

Lackey, P.N., "Comparison of the structure of students' evaluations of teaching in biology, mathematics, and sociology", *College Student Journal,* 1980, 14(1), 27-36.

Loewenstein, G. and Prelec, D., "Negative Time Preference", *The American Economic Review,* May, 1991, 347-352.

Loewenstein, G. and Prelec, D., "Preferences for Sequences of Outcomes", *Psychological Review,* 1993, 100(1), 91-108.

Loewenstein, G. and Sicherman, N., "Do Workers Prefer Increasing Profiles", *Journal of Labor Economics,* 1991, 9(1), 67-84.

Mackeigan, L.D., Larson, L.N., Draugalis, J.R., Bootman, J.L., and Burns, L.R., "Time Preference for Health Gains vs. Health Losses", *Pharmacoccon,* 1993, 3(5), 374-386.

Matsumoto, D., Peecher, M. E., & Rich, J. S., "Evaluations of outcome sequences", *Organizational Behavior and Human Decision Processes*, 2000, 83(2), 331–352.

Messe´, L. A. and Watts, B. L. "Complex nature of the sense of fairness: internal standards and social comparison as bases for reward evaluations", *Journal of Personality and Social Psychology*, 1983, 45(1), 84-93.

Messick, D. and Sentis, K., "Fairness and preference", *Journal of Experimental Social Psychology*, 1979, 15(4), 418-434.

Murstein, B. I, "The relationship of grade expectations and grades believed to be deserved to actual grades received", *The Journal of Experimental Education*, 1965, 33(4), 357–362.

Oliver, R. L., "A cognitive model of the antecedents and consequences of satisfaction

decisions", *Journal of Marketing Research*, 1980, 17(4), 460–469.

Ordondez, L.D., Connolly, T. and Coughlan, R., "Multiple Reference Points in Satisfaction and Fairness Assessment", *Journal of Behavioral Decision Making,* 2000, 13(3), 329-344.

Ployhart, R.E. and Ryan, A.M., "Applicants' reactions to the fairness of selection procedures: The effects of positive rule violations and time of measurement", *Journal of Applied Psychology,* 1998, 83, 3-16.

Read, D. and Loewenstein, G., "Enduring Pain for Money: Decisions Based on the Percpetion and Memory of Pain", *Journal of Behavioral Decision Making,* 1999, 12, 1-17.

Read, D. and Powell, M., "Reasons for Sequence Preferences", *Journal of Behavioral Decision Making,* 2002, 15, 433-460.

Redelmeier, D. A. and Kahneman, D. "Patients' Memories of Painful Medical Treatments: Real-Time and Retrospective Evaluations of Two Minimally Invasive Procedures", *Pain*, 1996, 66, 3-8.

Redelmeier, D. A., Katz. J., and Kahneman, D., "Memories of colonoscopy: a randomized trial", *Pain*, 2003, 104, 187-194.

Rodabaugh, R.C., and Kravitz, D.A., "Effects of procedural fairness on student judgments of professors", *Journal on Excellence in College Teaching,* 1994, 5(2), 67-83.

Schaffner, M., Burry-Stock J. A. Cho, G-P., Boney, T., and Hamilton, G., "What do kids think when their teachers grade?", *Paper presented at annual meeting of the American Educational Research Association*, (New Orleans, LA, April 24-28, 2000).

Schreiber, C.A., & Kahneman, D., Determinants of the remembered utility of aversive sounds. Journal of Experimental Psychology: General, 2000, 129(1), 27-42.

Stapleton, R.J., and Murkison, G., "Optimizing the fairness of student evaluations: A study of correlations between instructor excellence, study production, learning production, and expected grades", *Journal of Management Education,* 2001, 25(3), 269-291.

Tata, J., "Grade distributions, grading procedures, and students' evaluations of instructors: a justice perspective", *The Journal of Psychology*, 1999, 133(3), 263–271.

Tversky, A., and Kahneman, D. "Advances in prospect theory: Cumulative representation of uncertainty", *Journal of Risk and Uncertainty,* 1992, 5, 297-323.

Tyler, T.R., "Psychological models of the justice motive: Antecedents of distributive and procedural justice", *Journal of Personality and Social Psychology,* 1994, 67(5), 850-863.

van den Bos, K., Lind, E. A., Vermunt, R., & Wilke, H. A. M., "How do I judge my outcome when I do not know the outcome of others? The psychology of the fair process effect", *Journal of Personality and Social Psychology*, 1997, 72(5), 1034–1046.

van den Bos, K., Vermunt, R., and Wilke, H. A. M., "Procedural and distributive justice: What is fair depends more on what comes first than on what comes next", *Journal of Personality and Social Psychology,* 1997, 72, 1034-1046.

van den Bos, K., Wilke, H. A. M., Lind, E. A., & Vermunt, R., "Evaluating outcomes by means of the fair process effect: evidence for different processes in fairness and satisfaction judgments", *Journal of Personality and Social Psychology*, 1998, 74(6), 1493–1503.

van der Pol, M. M., and Cairns, J. A., "Negative and zero time preference for health", *Health Economic Letters*, 2000, 9(2), 171–175.

van der Pol. M. M., and Cairns, J. A., "Estimating Time Preferences for Health Using Discrete Choice Experiments", *Social Science Medicine,* 2001, 52, 1459-1470.

Varey, C. A. and Kahneman, D., "Experiences Extended Across time: Evaluation of Moments and Episodes", *Journal of Behavioral Decision Making,* 1992, 5, 169-185.

Wendorf, C. A. "Grade point average and changes in (great) grade expectations", *Teaching of Psychology,* 2002, 29(2), 136-138.

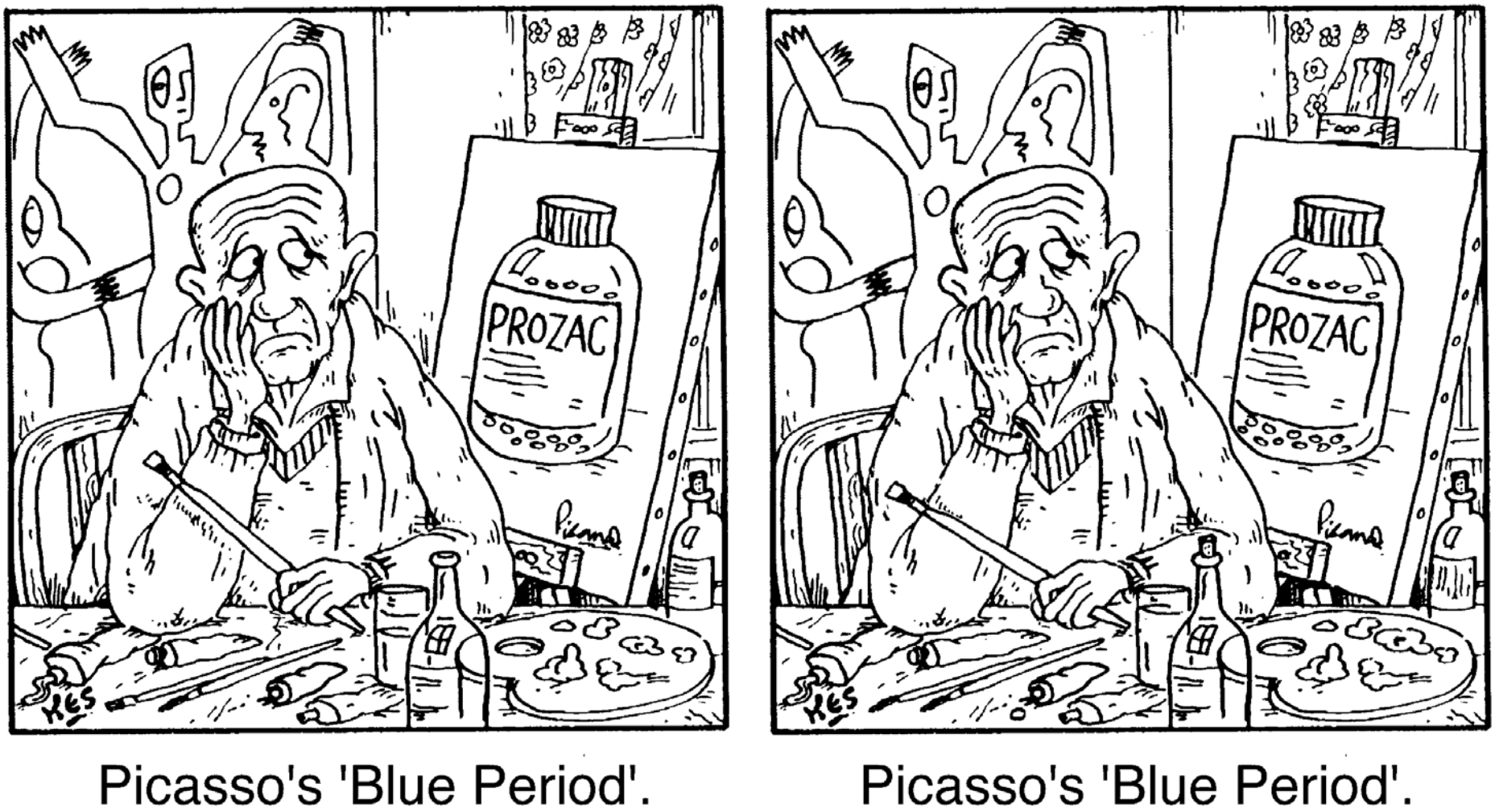
Wendorf, C.A., and Alexander, S., "The influence of individual- and class-level fairness-related perceptions on student satisfaction", *Contemporary Educational Psychology,* 2005, 30(2), 190-206.

X. Appendix A

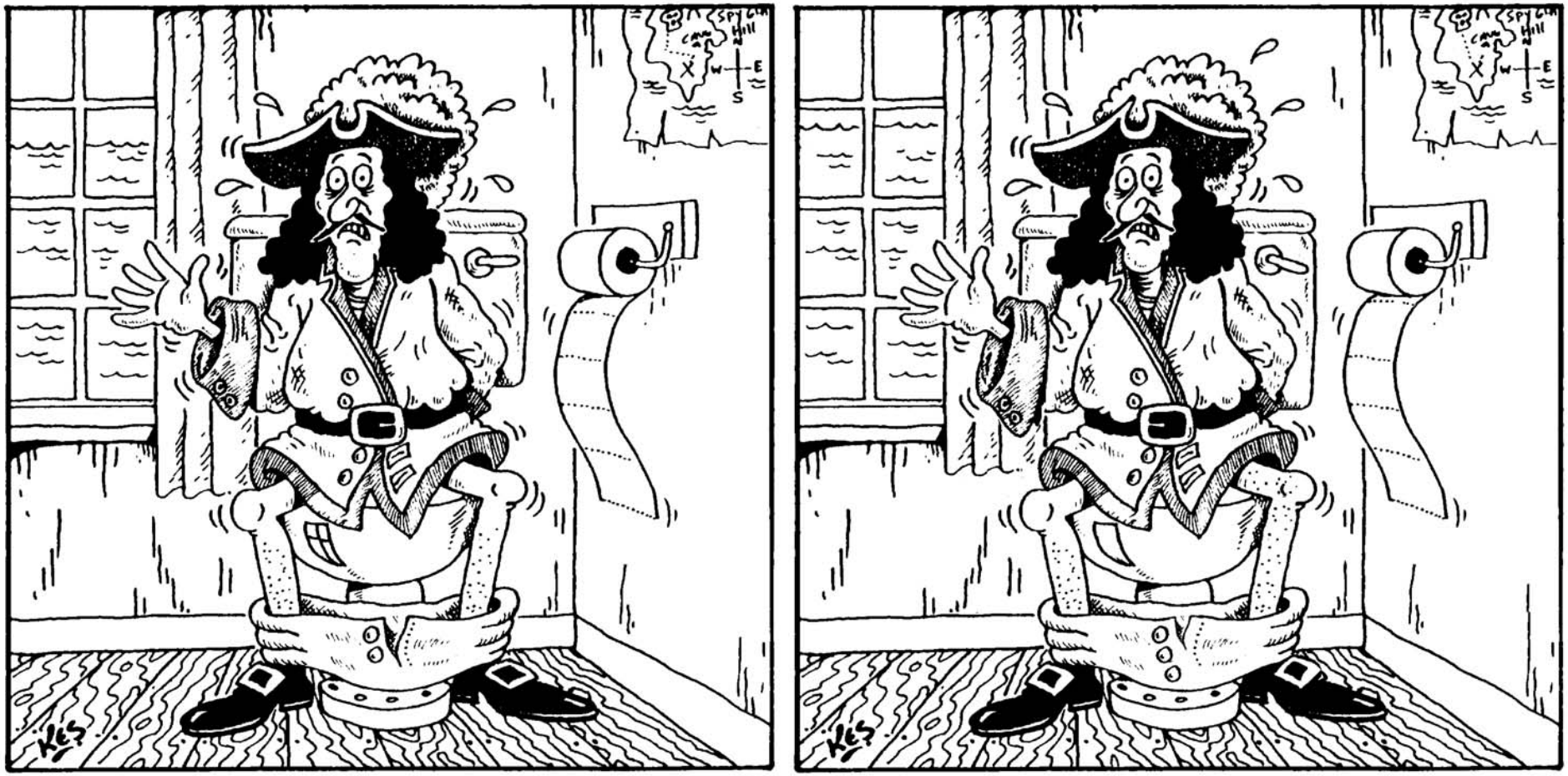
Cartoons and Grade Scale

Cartoons	% of Final Grade
Sniffer Dogs	12.5%
Picasso's Blue Period	12.5%
Captain Hook	12.5%
Oxford Don	12.5%
Clark Kent	50%

Cartoon 1: Sniffer Dogs



Cartoon 2: Picasso's Blue Period

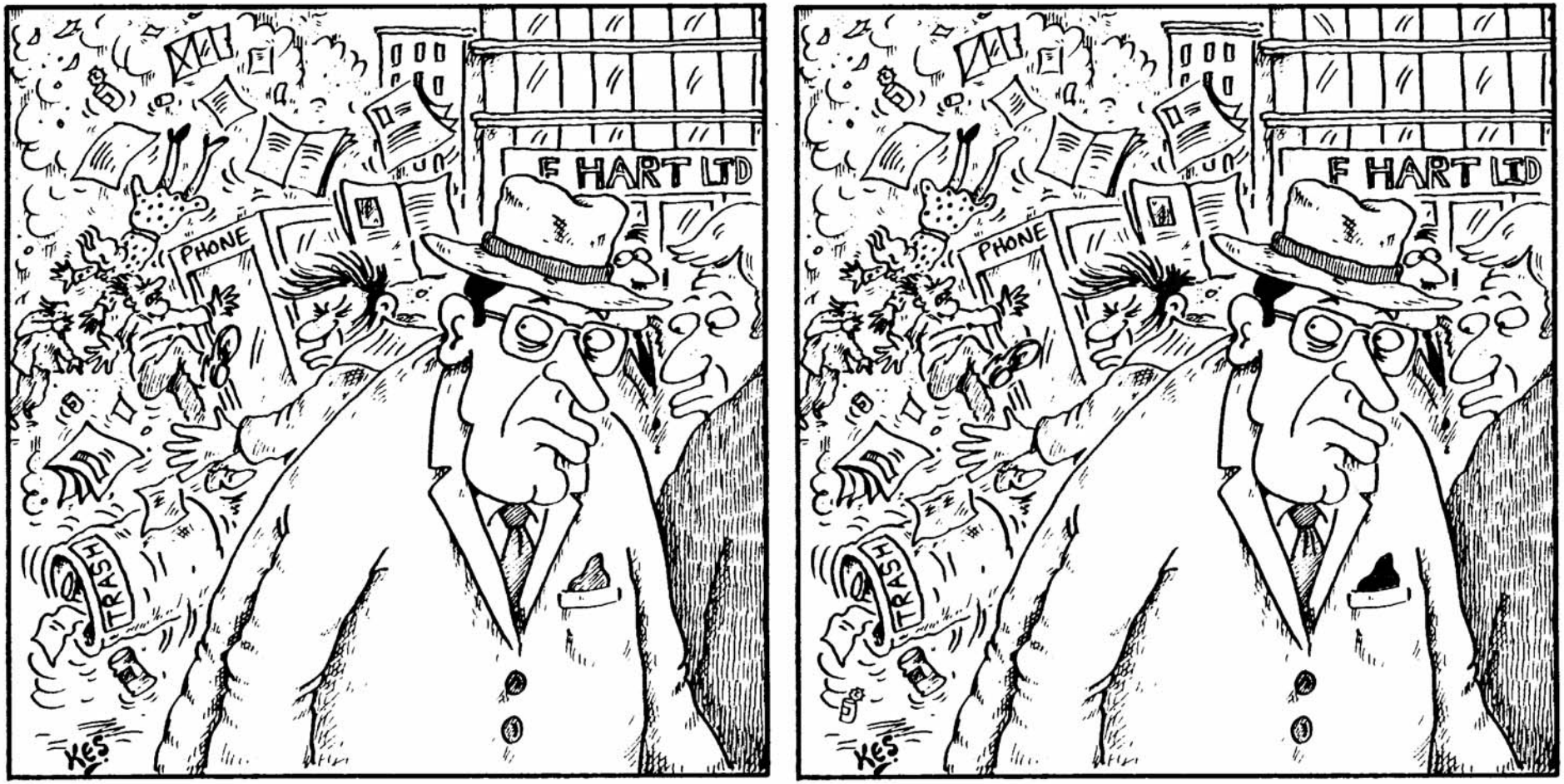


Cartoon 3: Captain Hook

An Oxford Don

An Oxford Don

Cartoon 4: Oxford Don



Cartoon 5: Clark Kent