On Rating Scales in Performance Appraisals: Performance Effects of an Unused Low Rating Category in Short-Term Interactions

Thomas Vogt
University of Cologne

thomas.vogt@uni-koeln.de

Working Paper May 2021

We investigate unused low rating categories in performance appraisals as an incentive design choice in short-term interactions. The literature proposes that unused low rating categories trigger what we term an *Incentive*, an *Evaluation* and a *Kindness-of-Scale Effect*. We explore how these effects affect performance in short-term interactions in two field experiments on Amazon MTurk. Subjects worked on a real effort task over two periods and received private rank feedback. The computer rated performance using three categories. In the baseline treatment subjects saw the rating scale with the actual three rating categories. In the other treatments subjects saw an additional fourth - but unused - low rating category. Dependent on the treatment, subjects were informed or not that the additional low category was unused. We do not find evidence that an unused low rating category increases performance in short-term interactions, independent of whether individuals are informed or not that this category is unused. Our results indicate that individuals do not only consider their individual incentives and performance ranking but also pay attention to the kindness of a rating scale in short-term interactions.

Keywords: Performance Appraisals, Unused Rating Categories, Incentives, Kindness, Field Experiment

1. Introduction

About 46% of non-managers and 60% of managers receive performance based payments (U.S. Census Bureau 2015). In such payment schemes, companies usually define rating scales to evaluate employees' performance and distribute payments accordingly.

Studies find that employees almost never rank in lower rating categories (see for example Ockenfels et al. 2015, Frederiksen et al. 2017) and thus that lower rating categories are often unused in performance appraisals.

The literature in psychology and economics considers unused rating categories in performance appraisals as rating biases of supervisors who assign too lenient ratings (see for example Landy and Farr 1980 and Prendergast 1999). However, most of the companies do not prevent unused rating categories by for example requiring supervisors to rank pre-defined percentages of employees in

each rating category (Holland 2006). As a result, while low rating categories are often unused, they are not removed from the scale of possible evaluations.

Motivated by the observation that many firms employ scales where the lowest categories are unused, we investigate unused low rating categories as an incentive design choice. We examine whether the presence of an unused low rating category in performance appraisals increases performance in short-term interactions. We refer to an unused low rating category as "dummy category".

When individuals believe that a dummy category is actually used, economic reasoning and tournament theory suggest that it triggers an *Incentive Effect* that raises performance. Employees have higher incentives to perform in the presence of a dummy category as low performance may result in lower ratings and payments. In that light, Berger et al. (2013) show that forcing supervisors to use lower rating categories increases performance. Moreover, following Lazear and Rosen (1981), a dummy category increases incentives when the payment scheme resembles a tournament since it increases the (perceived) prize spread of possible payments.

A broad stream of literature demonstrates that individuals reciprocate behavior of others by rewarding favors and penalizing unkindness (Rabin 1993, Fehr et al. 1993, 1997, Fehr and Rockenbach 2003, Falk et al. 2008). Employees receive more generous ratings in the form of higher relative ratings when they believe that the dummy category is actually used. Accordingly, a dummy category may trigger positively reciprocal reactions that raise performance (see for example Ockenfels et al. 2015, Sebald and Walzl 2014). We refer to this as *Evaluation Effect*. However, such a dummy category may also be seen as unkind and signal bad intentions of the employer as an additional punishment option is introduced (Bowles and Polanía-Reyes 2012). This, in turn, may trigger negatively reciprocal reactions that reduce performance (Levine 1998, Dufwenberg and Kirchsteiger 2004). We refer to this as *negative Kindness-of-the-Scale Effect*.

When employees know that the category is unused, incentives and ratings of employees remain unchanged. However, such a transparent dummy category may signal kindness and good faith of employers transmitting that they intentionally do not use an available punishment option. This, in turn, may induce positively reciprocal reactions and increase performance (*positive Kindness-of-the-Scale Effect*).

We tested how a dummy category affects performance in short-term interactions in two field studies in the online labor market of Amazon Mechanical Turk (MTurk; Horton et al. (2011)). In Study I, we investigated the potential reciprocal reactions to a dummy category when subjects believe that the category is used. More specifically, we tested whether the *Evaluation Effect* raises performance and hence outperforms the *negative Kindness-of-the-Scale Effect*. Therefore, we excluded the *Incentive Effect* of a dummy category by design. In Study II, we examined the total performance effect of a dummy category and hence the joint effect of potential reciprocal reactions

and the *Incentive Effect*. More specifically, we tested whether a dummy category raises performance and hence whether the *Incentive Effect* and *Evaluation Effect* jointly raise performance and thus outperform the *negative Kindness-of-the-Scale Effect*. Moreover, we examined whether a transparent dummy category and thus the *positive Kindness-of-the-Scale Effect* raises performance.

Both studies followed the same base protocol. As a university department, we hired subjects to digitize handwritten class grades but did not disclose that this task was an experiment. Subjects worked twice in two consecutive weeks. We used week one only to rank and provide feedback to subjects in week two. For their work in week one, subjects received a bonus payment based on relative performance. We explained the incentive mechanism of the bonus payment, but did not reveal the rating scale such that week one was identical across treatments. Accordingly, we analyze treatment effects only in week two. In week two, subjects saw their individual performance rating and resulting bonus payment for week one before they worked again on the same task. To evaluate how a dummy category affects subjects' well-being and the perception of rating scales, we asked how satisfied they were with their rating and how kind they perceived their rating scale.

The computer rated performance using three categories but subjects either saw three or four rating categories, dependent on the treatment. In treatment No Dummy (ND), subjects saw the actual three rating categories used by the computer. In treatment Dummy (D), subjects saw an additional fourth category that was never used. We did not inform that the additional category was never used.

In Study I, we tested whether the potential positive *Evaluation Effect* raises performance and hence is stronger than the potential *negative Kindness-of-the-Scale Effect*. To exclude the potential *Incentive Effect*, subjects did not receive a rating or bonus payment but the same fixed payment in both treatments ND and D in week two. Thus, only the rating (scale) shown for week one varied between treatments as either 3 or 4 rating categories were displayed in the performance appraisal. Accordingly, differences between treatments in week two can only be influenced by the rating (scale) seen for week one.

A dummy category did not raise performance in week two of Study I: Average performance and performance across rating categories did not differ significantly between treatments ND and D. Subjects did not report higher satisfaction with their individual rating when seeing an additional rating category in treatment D. They did, however, evaluate the rating scale in treatment D as being less kind.

In Study II, we tested the total performance effect of a dummy category. We analyzed whether a dummy category raises performance when subjects believe that the dummy category is used and hence whether the *Evaluation Effect* and *Incentive Effect* are stronger than the *negative Kindness-of-the-Scale Effect*. Therefore, subjects received a bonus payment based on relative performance and

hence additional relative performance rating also for week two: They learned that the rating scale of week one was also used for week two. Thus, not only the rating (scale) shown for week one but also the incentive scheme in week two varied between treatments D and ND. Subjects in treatment D faced an additional low rating category which increased the (perceived) prize spread of the bonus tournament compared to treatment ND. Accordingly, differences between these treatments in week two may not only be influenced by the rating (scale) seen for week one but also by the anticipation of and hence incentive induced by the rating and payment for week two.

A dummy category did not raise performance in week two of Study II either: Average performance did not differ significantly between treatments ND and D. We do, however, observe opposing effects of a dummy category. Subjects receiving the lowest rating worked significantly more while those receiving higher ratings worked significantly less in treatment D. As in Study I, subjects did not report different levels of individual rating satisfaction between treatments. Moreover, the rating scale in treatment D was perceived as less kind - however, only from those receiving the lowest rating.

In addition, we analyzed whether a dummy category raises performance when subjects are informed that the category is unused (*positive Kindness-of-the-Scale Effect*). Therefore, we ran an additional treatment Transparent Dummy (TD): Subjects saw an additional unused low rating categories but were informed that this rating category was unused. The communicated number of rating categories used and the prize spread were equivalent in treatments ND and TD.

Also a transparent dummy category did not raise performance in week two of Study II: Average performance and performance across rating categories did not differ significantly between treatments ND and TD. However, subjects that did not rank in the lowest rating category evaluated the rating scale in treatment TD as being more kind.

Our results indicate two main insights: (1) A dummy category does not raise performance in short-term interactions. (2) Individuals do not only consider their incentives and individual performance ranking but also pay attention to the design and kindness of a rating scale in short-term interactions. Our work is closest to Vogt et al. (2021). They investigate how a dummy category affects performance in a setting where employees experience multiple ratings and can react dynamically. Consistent with our findings, they do not see significant performance differences in the first working period.

The remainder of this paper is structured as follows. In Section 2, we develop our hypotheses. In Sections 3 and 4, we present Study I and II, respectively. In Section 5, we conclude.

2. Literature and Hypotheses Development

Research on reciprocity in employer-employee interactions, which originated from the theory of gift exchange (Adams 1963, Akerlof 1982), shows that employees reciprocate kind and punish unkind behavior (see for instance Akerlof and Yellen 1988, 1990, Fehr et al. 1993, 1997, Charness 2004, Chung and Narayandas 2017). Moreover, Falk et al. (2008) show that intentions of the gift-giver are crucial to provoke reciprocal reactions (see also Falk and Fischbacher 2006, Kube et al. 2012).

Deploying a dummy category may increase performance by triggering positively reciprocal responses when subjects believe the category is actually used: Ceteris paribus, employees receive more generous feedback in the form of relatively higher ratings: For example, a rating in the category 2 of 3 (top 66%) becomes a rating in the category 2 of 4 (top 75%). This may shift employees' reference point which, in turn, increases positively reciprocal reactions among those who receive high ratings or reduce negatively reciprocal reactions of those whose ratings fall short of their expectations (see for instance Ockenfels et al. (2015) for an analysis of reciprocal reactions to performance ratings). In that view, Bol (2011, p.1555) points out: "The behavioral perspective expects leniency bias to positively affect perceived fairness by increasing the congruence between the rating the employee thinks s/he deserves and the rating s/he actually receives". Following the argument of Ellingsen and Johannesson (2007), more generous ratings positively influence employees if they see them as a sign of employer appreciation. Moreover, receiving higher relative ratings may also make employees happier (Parducci 1965) and consequently motivate higher performance (Oswald et al. 2015). We refer to positive effects of awarding higher relative ratings in the presence of a dummy category as *Evaluation Effect*.

But there may also be negatively reciprocal effects when individuals believe that the dummy category is used. Bowles and Polanía-Reyes (2012, p.388) argue that incentive schemes transmit information about the type and intentions of the incentive designer. If the specific incentive scheme signals an employer's "bad" intentions, employees may punish this. Adding a low rating category to the rating scale may be judged as being unkind and signal "bad news" about the employer's type or intention since an additional punishment option is introduced. In turn, this can induce negative reactions (Fehr and Rockenbach 2003). We refer to negative effects of employing a less kind rating scale in the presence of a dummy category as *negative Kindness-of-the-Scale Effect*.

We expected individuals to focus more on their own rating than on the overall kindness of a rating scale. Therefore, we hypothesized the positive *Evaluation Effect* to be stronger than the *negative Kindness-of-the-Scale Effect* when individuals are not informed that the dummy category is unused. We pre-registered:

Hypothesis 1a: Evaluation Effect *Average performance is higher in the presence of a dummy category if individuals are not informed that the dummy category is unused and incentives are held constant.*

Simple economic reasoning suggests that a dummy category raises performance when employees believe that the category is actually used. Employees have higher incentives to work as the additional low rating category penalizes low performance more. A key result in tournament theory is that higher prize spreads should induce higher performance (see for example Lazear and Rosen (1981)). Since adding a dummy category increases the (perceived) prize spread, a dummy category should raise performance. We refer to positive effects of higher incentives in the presence of a dummy category as *Incentive Effect*.

We expected individuals to focus more on their individual ratings and (monetary) incentives than on the overall kindness of a rating scale. Therefore, we conjectured the positive *Incentive* and *Evaluation Effect* of a dummy category to be stronger than the *negative Kindness-of-the-Scale Effect* when individuals are not informed that the category is unused. Consequently, we pre-registered:

Hypothesis 1b: Evaluation & Incentive Effect *Average performance is higher in the presence of a dummy category if individuals are not informed that the category is unused.*

When individuals know that the dummy category is unused it may trigger positively reciprocal reactions. Ratings and incentives do not change if individuals know that the dummy category is never used. However, the transparency of not using the lowest category may transmit "good news" about employers as it signals that they refrained from using an available punishment option. Following the reasoning outlined above, this might signal kindness and good faith of employers and in turn increase performance. We refer to positive effects of employing a more kind rating scale in the presence of a transparent dummy category as *positive Kindness-of-the-Scale Effect*. We hence pre-registered:

Hypothesis 2: Positive Kindness-of-the-Scale Effect *Average performance is higher in the presence of a dummy category if individuals are informed that the category is unused.*

We expected that *Evaluation* and *Incentive Effect* are stronger than the *Positive Kindness-of-the-Scale Effect* since we expected that individuals focus more on their individual ranking and (monetary) incentives. Accordingly, we pre-registered:

Hypothesis 3: Evaluation & Incentive Effect II *Average performance is higher if individuals are not informed that the dummy category is unused than if they are informed that the dummy category is unused.*

The literature suggests that those ranking lowest react stronger to an additional low - but unused - rating category. Studies on performance feedback report de-motivating effects of receiving negative feedback in the form of low rankings, see for example Barankay (2011, 2012) or Gill et al. (2019). If individuals are not aware that the category is unused, employing a dummy category avoids giving harsh negative feedback to those ranking lowest. In the same light - if subjects know

that the dummy category is unused - we expect those ranking lowest to perceive the kind act of not using a punishment option stronger. Moreover, one might argue that the proposed *Incentive Effect* is stronger or - even more conservative - only present for those ranking lowest as they have the highest probability to be ranked in that category. We refer to the stronger performance increase of those ranking lowest when seeing a (transparent) dummy category as *Last Place Effect*. We hence pre-registered:

Hypothesis 4: Last Place Effect *The performance increase in the presence of a dummy category is stronger for those ranking lowest.*

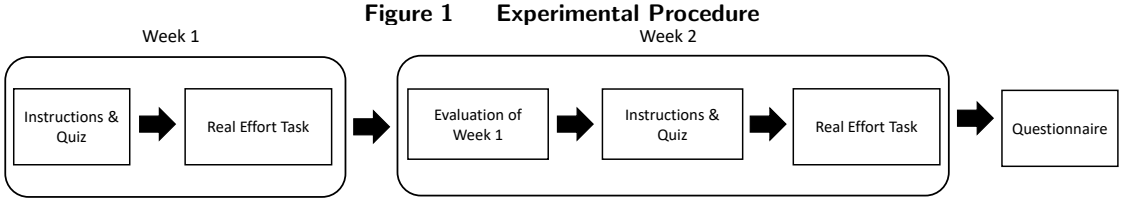
3. Study I: The Effect of a Dummy Category on Performance When Incentives are Held Constant

In Study I, we tested the reciprocal responses to a dummy category when subjects are not informed that the dummy category is unused. We tested Hypothesis 1a and investigated whether the potentially positive effects of giving more generous feedback (*Evaluation Effect*) outweigh the potentially negative effects of employing a potentially less kind rating scale (*negative Kindness-of-the-Scale Effect*). Therefore, we excluded the potential *Incentive Effect* of a dummy category by holding incentives constant between treatments. We also analyzed whether the performance effects were stronger for those ranking lowest (Hypothesis 4).

3.1. Experimental Design

Overview As a university department, we recruited subjects from the online labor market Amazon Mechanical Turk to digitize handwritten grades. Our design is in accordance with standard ethical guidelines but we did not disclose that the task was an experiment. Subjects worked in two consecutive weeks. We paid a bonus payment based on relative performance for week one. We explained that subjects receive higher bonus payments, the higher subjects rank relative to their peers. However, we did not explained details about the rating scale such that week one was identical across treatments. Subjects received private performance feedback for week one before they worked again in week two. In treatment (No) Dummy, we did (not) display a dummy category in the performance evaluation of week one. In week two, subjects did not receive a performance dependent payment in order to prevent the potential *Incentive Effect* from different incentives between treatments. We sent out a questionnaire on the kindness of the rating scales and demographics after week two (see Appendix H). Figure 1 shows the experimental procedure. See Appendix E & F for screenshots of week one and two.

Figure 1 Experimental Procedure



Experimental Details We used week one only to provide performance ratings in week two. Therefore, it was kept the same across treatments to avoid treatment specific performance effects that distort ratings between treatments. Subjects worked for 20 minutes digitizing grades from scanned exam cover sheets. See Appendix A for details on the real effort task. Subjects learned that they receive a bonus based on relative performance in addition to a fixed wage. However, they did not learn the rating scale or respective number of rating categories. Performance was defined as the number of correctly entered cover sheets; a cover sheet was evaluated as entered correctly if all grades were entered correctly. Subjects had to pass a quiz on the task and payment structure to be able to work.

We used week two to test performance effects of a dummy category. Subjects were invited via e-mail to work again. They could work on the task between Monday and Friday. Upon entering the task, subjects received private performance rankings for their work in week one. Dependent on the treatment, subjects saw a dummy category in the rating scale. They were then asked how satisfied they were with their individual rating. The incentive scheme of week two was explained in the instructions afterwards: We paid the same fixed wage in both treatments but no performance dependent bonus to eliminate effects due to different incentive schemes. Hence, subjects did not receive rank feedback for week two. Subjects had to pass a quiz on the task and payment structure to work again. Working time was not restricted.

The computer rated performance in week one using three rating categories in both treatments. Ratings were based on relative performance and followed the same procedure in all treatments such that only categories one to three were actually awarded. Category one was awarded to the highest performing subjects and category three to the lowest performing subjects. Subjects were not informed about the specific details of the rating procedure.

Treatment Variation Subjects either saw three or four rating categories, dependent on the treatment. Figure 2 depicts exemplary the scale subjects saw when receiving the rating "Grade 3" across treatments. In treatment "No Dummy" (ND) subjects saw the actual three-point rating scale used by the computer. In treatment "Dummy" (D), an additional fourth rating category was displayed at the bottom. Subjects were not informed that the additional category was unused. We randomly assigned subjects to either treatment D or ND stratifying assignment based on the performance in

Figure 2 **Performance Ratings Receiving Grade 3 Across Treatments**

No Dummy

Grade	1	2	3
Bonus	$2.00	$1.50	**$1.00**
% of workers	30%	40%	30%

Dummy

Grade	1	2	3	4
Bonus	$2.00	$1.50	**$1.00**	$0.00
% of workers	30%	40%	30%	

week one. In treatment ND, subjects learned that 30% of the ratings were given in the top category - Grade 1 -, 40% in the middle category - Grade 2 -, and 30% in the lowest category - Grade 3. To avoid deception in treatment D, subjects learned that 30% of the ratings were given in category 3 and 4 that is, Grade 3 and 4, respectively.

Experimental Protocol and Subject Pool We recruited subjects on Amazon's Mechanical Turk (MTurk) online labor market using the service of TurkPrime (Litman et al. 2017) to manage our HITs. The experiment was conducted online with Qualtrics and a self-developed Javascript.

Over the past decade, MTurk has received increased attention of researchers as platform to conduct scientific experiments: see for example Horton (2010), Barankay (2011) and the reference in Horton et al. (2011). On Amazon's platform, employers can post job offers, so called Human Intelligence Tasks (HITs), to a workforce of at least eighty-five thousand US workers active during the time of our study (Robinson et al. 2019). For a more detailed description of the marketplace see Ipeirotis (2010) or Paolacci et al. (2010).

We ran our treatments in March 2018. We recruited subjects on Monday and Tuesday in the first week. No subject participated in more than one treatment. Subjects were invited via e-mail to work again on Monday the week after. To increase the likelihood of returning in week two, subjects could work on the task all week (Monday-Friday) as well as pause and return to the task later. We only recruited residents of the United States and required workers to have had completed at least 100 HITs with an approval rate of at least 90% to ensure that subjects were familiar with MTurk, avoid complications arising from difficulties in understanding the English task instructions and to prevent performance noise due to different time-zones. Note that theses sampling restrictions still allow a sufficient total population size (Robinson et al. 2019) and thus worker non-naïveté (Chandler et al. 2014) cannot be a problem in our study.

Selective attrition is not a concern in our study. We avoided selection in the return rate, as treatment details were revealed only after subjects returned in week two. However, when returned, subjects could drop-out after receiving their performance rating. Moreover, we excluded subjects from the experiment that failed the quiz or worked on a device without sufficient screen resolution. Additionally, subjects had the choice to answer the questionnaire as it was sent out after working

in week two. We check selective attrition for the afore mentioned cases. There are no statistically significant differences between treatments neither for the drop-out rates ($\chi^2(1) = 0.27$ $p = .60$), the screen-out rates ($\chi^2(1) = 0.47$ $p = .49$) or the questionnaire-return rates ($\chi^2(1) = 0.05$ $p = .82$). Our study hence does not suffer from selective attrition.

946 subjects completed week two. Of those who answered the questionnaire, 58% were female, the median age was 35. The median educational level was a bachelor's degree and the median income class ranged from \$30,001 to \$40,000. See Table A1 in the Appendix B for detailed sample demographics. Earnings ranged between \$5.50 and \$8.50 depending on the bonus payment. The median experiment duration was 46.67 minutes (sum of week one and two). This results in a median hourly wage of \$7.55, which is substantially above median earnings on MTurk and above the federal minimum wage in the United States.

3.2. Results

We first analyze how a dummy category affected performance. We then examine questionnaire data to explore how it affected rating satisfaction and the perceived kindness of a rating scale.

3.2.1. Performance Across Treatments We hypothesized that average performance is higher in treatment Dummy (D) as compared to treatment No Dummy (ND) since we expected individuals to focus more on their own rating than on the kindness of a rating scale (Hypothesis 1a: *Evaluation Effect*). We analyze performance in week two as week one was the same across treatments. Subjects did not receive a performance rating for week two and incentives did not differ between treatments in week two. Thereby, we excluded any potential *Incentive Effect* of a dummy category. Hence, performance can only be affected by reciprocal responses induced by the additional rating category shown.

Contrary to our hypothesis, a dummy category did not increase average performance: The coefficient of the treatment indicator "Dummy Category" is insignificant and negative in our regression analysis shown in column (1) of Table 1. We report OLS regressions and control for week one performance to capture individual performance differences that can still occur despite our sampling procedure. The results are robust to using tobit regressions and controlling for the day, as well as time of day subjects worked (Appendix C). Thus, we do not find support that the *Evaluation Effect* is stronger than the *Kindness-of-the-Scale Effect* when subjects receive one performance rating. It seems that the two opposing reciprocal effects offset each other.

We next analyze how a dummy category affected the performance of those ranking lowest. We hypothesized that the effects are stronger for these subjects since being rated second last in

Table 1 Impact of a Dummy Category on Performance if Incentives are Held Constant

Dependent Variable:	**ND vs. D**	
Number of Cover Sheets Entered Correctly	(1)	(2)
Dummy Category	-3.77	-4.57
	(2.78)	(3.55)
Dummy Category#Grade 3 in t-1		2.98
		(5.19)
Grade 3 in t-1		-2.26
		(5.18)
Pre-round Performance	0.57***	0.57***
	(0.05)	(0.08)
Constant	15.86***	16.96***
	(3.38)	(6.51)
Observations	946	946

Note: Ordinary least squares regressions on individual output are performed. D := Dummy Treatment; ND := No Dummy Treatment.

* p < 0.10, ** p < 0.05, *** p <0.01. Robust standard errors are in parentheses, clustered at the individual level.

treatment D as compared to last in treatment ND may trigger a stronger reaction (Hypothesis 4: *Last Place Effect*).

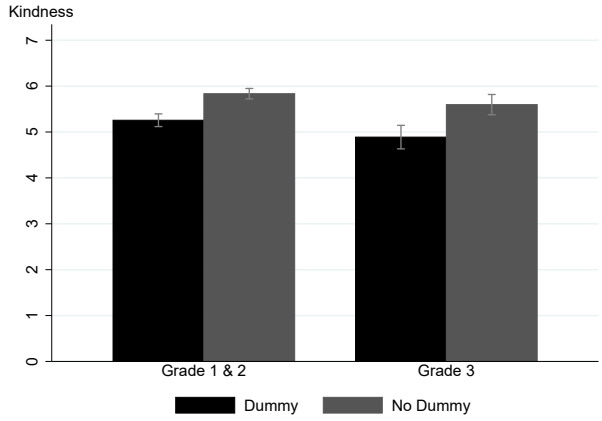
Grade 3 was the lowest possible ranking. In our setting, grades were equivalent to performance ranks in the pre-round: Subjects receiving grade 3 belonged to the lowest 30% of the performance distribution, grade 1 and 2 belonged to the top 70%.

Contrary to our hypothesis, also low performer did not work more when seeing a dummy category: The interaction term of receiving grade 3 with the treatment indicator "Dummy Category" in column (2) of Table 1 is insignificant. The results are robust to using tobit regressions and controlling for day and time of day (see Appendix C). Hence, we also do not observe a *Last Place Effect* when subjects receive only one performance rating.

3.2.2. Rating Satisfaction Across Treatments We test whether individuals in treatment D were more satisfied with their rating (grade) than those awarded with the same category in treatment ND.

When receiving their performance rating in week two, we asked subjects - on the same screen - how satisfied they were with their rating. The scale ranged from 1 (not satisfied at all) to 7 (extremely satisfied), increasing values of the score thus reflect higher satisfaction levels. The payment scheme of the second part could not influence satisfaction as incentives did not differ between treatments and were communicated after the ranking was shown.

Figure 3 Kindness of Rating Scale Across Treatments



Subjects across treatments did not report different levels of satisfaction when receiving grade 1, 2 or 3 (see Table A5 in Appendix D for the p-values of Wilcoxon tests comparing the satisfaction across grades). Hence, in case of one evaluation, seeing a dummy category did not increase rating satisfaction - in line with the results comparing performance between treatments.

3.2.3. Kindness of Rating Scale Across Treatments To investigate whether a dummy category conveys "bad news" about an employer, we analyze survey data obtained in the post-trial questionnaire (see Appendix H). We showed subjects the rating scale of their treatment again and asked how kind they perceived it. The scale ranged from 1 (very unkind) to 7 (very kind), such that increasing values of the score reflect higher kindness levels.

If a dummy category was interpreted as bad news, subjects should have evaluated the rating scales of treatment Dummy (D) as less kind. We test this by comparing subjects' kindness evaluations between treatments ND and D. Figure 3 shows the mean kindness evaluations between treatments. We differentiate between subjects receiving the top rating categories grade 1 and 2 (on the left) and the lowest rating category grade 3 (on the right hand side) to analyze whether effects differ across performance classes.

The additional low rating category was interpreted as "bad" news. Across performance classes, subjects evaluated the rating scale in treatment D - when they did not know that the additional rating category was unused - as being less kind (Wilcoxon test, two-sided, $p = .000$ and $p = .000$, respectively).

4. Study II: The Effect of a Dummy Category on Performance

In Study II, we investigated the total performance effect of a dummy category. We tested whether a dummy category raises performance when individuals believe that the category is used (Hypothesis

Figure 4 **Performance Ratings Receiving Grade 3 Across Treatments**

No Dummy

Grade	1	2	3
Bonus	$2.00	$1.50	**$1.00**
% of workers	30%	40%	**30%**

Dummy

Grade	1	2	3	4
Bonus	$2.00	$1.50	**$1.00**	$0.00
% of workers	30%	40%	30%	

Transparent Dummy

Grade	1	2	3	4
Bonus	$2.00	$1.50	**$1.00**	$0.00
% of workers	30%	40%	**30%**	0%

1b) and hence whether the potentially positive effects of higher incentives (*Incentive Effect*) and more generous feedback (*Evaluation Effect*) outweigh the potentially negative effects of employing a potentially less kind rating scale (*negative Kindness-of-the-Scale Effect*). In addition, we tested whether a dummy category raises performance when individuals are informed that the dummy category is unused (Hypothesis 2) and hence whether a more kind rating scale raises performance (*positive Kindness-of-the-Scale Effect*). We also examined whether performance was higher in the treatment where individuals were not informed that the dummy category is unused than in the treatment where they were informed (Hypothesis 3). In all three analyses, we tested if the performance effects were stronger for those ranking lowest (Hypothesis 4).

4.1. Experimental Design

Compared to Study I, subjects received a bonus payment and hence performance rating also for their performance in week two. As a result, Study II followed the protocol of Study I and everything else was the same except for the payment scheme in week two. Subjects learned that the rating scale of week one was also used for determining the rating and payment in week two. Accordingly, not only the rating (scale) shown for week one but also the anticipation of and the incentives induced by the rating in week two can affect performance in week two. See Appendix E & G for screenshots of week one and week two.

To test the effect of a dummy category when individuals know that the category is unused, we ran an additional treatment "Transparent Dummy" (TD). In the new treatment subjects also saw four rating categories but learned that the fourth rating category was unused. Note that treatment ND and treatment TD have the same prize spread and the communicated number of rating categories in use is equivalent. Figure 4 shows exemplary the scale subjects saw when receiving the rating "Grade 3" across treatments. We randomly assigned subjects to either treatment D, ND or TD stratifying assignment based on the performance in week one.

We recruited subjects on Amazon's Mechanical Turk (MTurk) in June 2018. We excluded participants of Study I and no subject took part in more than one treatment. A questionnaire was sent out to all subjects two weeks after week two.

Selective attrition is not a concern in our study. We avoided selection in the return rate as treatment details were revealed only after subjects returned in week two. We check selective attrition for the drop-out, screen-out and questionnaire return rates. There are no statistically significant differences across treatments neither for the drop-out rates ($\chi^2(2) = 0.21$ $p = .90$), the screen-out rates ($\chi^2(2) = 1.78$ $p = .41$) or the questionnaire-return rates ($\chi^2(2) = 4.63$ $p = .10$).

1,389 subjects completed week two. Of those who answered the questionnaire 61% were female, the median age was 34. The median educational level was a bachelor's degree and the median income class ranged from \$30,001 to \$40,000. See Table A2 in the Appendix B for detailed sample demographics. Earnings ranged between \$5.50 and \$8.50 depending on the bonus payment. The median experiment duration was 51.47 minutes resulting in a median hourly wage of \$8.74, which is substantially above median earnings on MTurk and above the federal minimum wage in the United States.

4.2. Results

We first present performance effects of a dummy category. We then analyze how a dummy category affected individual rating satisfaction and the perceived kindness of a rating scale.

4.2.1. Performance Across Treatments In all treatments, incentive schemes in week two resembled a tournament in which participants competed for bonus payments. Subjects in treatment ND and TD faced a three prize tournament while those in treatment D entered a (perceived) four prize tournament. A conjecture of the incentive literature is that the dummy category in treatment D - where subjects did not know that it is unused - increases performance (*Incentive Effect*). Compared to Study I, behavior in these experimental conditions may hence not only be influenced by a different rating (scale) shown (*Evaluation & negative Kindness-of-the-Scale Effect*) but also by different communicated incentives (*Incentive Effect*).

Table 2 shows our regression results. We investigate performance effects in week two as week one was the same across treatments. We report OLS regressions and control for week one performance to capture individual performance differences that can still occur within the degrees of freedom of our sampling procedure. The results are robust to controlling for the day and time of day subjects worked as well as performing tobit regressions (see Appendix C).

First, we compare treatment D - where individuals were not informed that the dummy category was unused - to treatment ND. We expected the positive effects of higher relative ratings (*Evaluation Effect*) and higher incentives (*Incentive Effect*) to be stronger than the potentially negative effects of a less kind rating scale in the presence of a dummy category (*negative Kindness-of-the-Scale Effect*). We thus hypothesized that average performance in treatment D is higher than in treatment ND (Hypothesis 1b: *Evaluation & Incentive Effect*).

Table 2 Impact of a Dummy Category on Performance

Dependent Variable:	ND vs. D		ND vs. TD		D vs. TD	
Number of Cover Sheets Entered Correctly	(1)	(2)	(3)	(4)	(5)	(6)
Dummy Category	-3.69	-7.51*			-4.54	-5.50
	(3.31)	(4.05)			(2.93)	(3.51)
Dummy Category#Grade 3 in t-1		15.32**				3.68
		(6.62)				(6.30)
Transparent Dummy Category			0.90	-2.05		
			(3.35)	(4.01)		
Transparent Dummy Category#Grade 3 in t-1				11.56		
				(7.07)		
Grade 3 in t-1		-14.48**		-15.11**		-0.14
		(6.83)		(6.64)		(6.08)
Pre-round Performance	0.95***	0.88***	0.96***	0.87***	0.92***	0.94***
	(0.05)	(0.08)	(0.05)	(0.08)	(0.05)	(0.07)
Constant	26.05***	34.47***	25.32***	35.49***	29.15***	28.02***
	(4.26)	(7.59)	(4.15)	(7.14)	(3.92)	(6.39)
Observations	928	928	934	934	934	934

Note: Ordinary least squares regressions on individual output are performed. D, Dummy Treatment; ND, No Dummy Treatment; TD, Transparent Dummy Treatment.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors are in parentheses, clustered at the individual level.

The results do not support our hypothesis. Average performance in treatment D was not higher than in treatment ND: The coefficient of the treatment indicator "Dummy Category" is insignificant and negative in column (1) of Table 2. Thus, we do not find support for a stronger *Evaluation* and *Incentive Effect* when subjects receive two ratings. Instead, it seems that the negative effects of a less kind rating scale offset the positive effects of higher rankings and higher incentives.

We next analyze the effects of the dummy category in treatment D on subjects receiving the lowest ranking (grade 3). The underlying hypothesis is that both, *Incentive* and *Evaluation Effect* are stronger for those ranking lowest (Hypothesis 4: *Last Place Effect*).

The results support the hypothesis of a *Last Place Effect* in the presence of a dummy category. Subjects who received grade 3 in treatment D worked significantly more than subjects with the same rating in treatment ND: The interaction term of the treatment indicator "Dummy Category" with receiving a grade 3 for performance in week one is significant in column (2) of Table 2. Interestingly, subjects receiving grade 1 or 2 worked significantly less in treatment D indicated by the significant negative treatment indicator "Dummy Category" in column (2) of Table 2. This indicates that - when individuals receive two ratings - a dummy category has positive performance effects only on those ranking lowest.

Second, we compare treatment TD - where individuals did know that the dummy category was unused - to treatment ND. Incentives did not differ between these treatments. However, a

transparent dummy category might signal kindness of the employer and induce positively reciprocal reactions that increase performance. We hence, hypothesized that average performance in treatment TD is higher than in treatment ND (Hypothesis 2: *Positive Kindness-of-the-Scale Effect*).

Contrary to our hypothesis, we do not find support for a *positive Kindness-of-the-Scale Effect* when subjects receive two ratings. Average performance in treatment TD was not higher than in treatment ND: The coefficient of the treatment indicator "Transparent Dummy Category" is statistically and economically insignificant in column (3) of Table 2.

Comparing performance between those receiving the lowest ranking, we do not find support for a *Last Place Effect* in presence of a transparent dummy category (Hypothesis 4: *Last Place Effect*). The interaction term of the treatment indicator "Transparent Dummy Category" with grade 3 is insignificant in column (4) of Table 2.

Third, we compare performance in treatment D with performance in treatment TD. We hypothesized that average performance is higher in treatment D than in treatment TD (Hypothesis 3: *Evaluation & Incentive Effect II*) since we expected individuals to focus more on their own rating and incentives than on the kindness of a rating scale.

We do not find support that performance is higher in the presence of a dummy category as compared to a transparent dummy category. Performance was not significantly different between treatments D and TD: The coefficient of the treatment indicator "Transparent Dummy Category" is insignificant and negative in column (5) of Table 2.

We do not find support for a *Last Place Effect* (Hypothesis 4) since there are also no significant differences comparing performance between those ranking lowest. The interaction term of the treatment indicator "Dummy Category" with grade 3 is insignificant in column (6) of Table 2.

4.3. Rating Satisfaction Across Treatments

We analyze questionnaire data to investigate whether higher relative ratings in treatment Dummy increase subjects' rating satisfaction. When subjects received their rating in week two, we asked them how satisfied they were with their rating. The scale ranged from 1 (not satisfied at all) to7 (extremely satisfied), increasing values of the score reflect higher satisfaction levels.

As in Study I, higher relative ratings did not increase rating satisfaction in Study II which is also in line with the results of the performance analysis. Subjects across treatments did not report different levels of satisfaction when receiving grade 1, 2 or 3 (see Table A5 in Appendix D for the p-values of Wilcoxon tests comparing the satisfaction across grades). Note that the different payment schemes across treatments can not influence responses as they were communicated afterwards.

4.3.1. Kindness of Rating Scale Across Treatments To investigate whether the presence of a dummy category affects the perceived kindness of a rating scale, we report survey data obtained in the post-trial questionnaire (see Appendix H). We showed subjects the rating scale of their treatment again and asked how kind they perceived it on a scale from 1 (very unkind) to 7 (very kind). Increasing values of the score reflect higher kindness levels. Figure 5 shows the kindness evaluations of subjects receiving grade 1 and 2 on the left and grade 3 on the right hand side. See Table A6 in Appendix D for the p-values of Wilcoxon tests comparing the kindness across treatments.

If individuals do not know that the additional rating category is unused, a dummy category introduces an additional punishment option and thereby may signal an employers unkindness and bad intentions. The literature suggests - and we also observed in Study I - that individuals perceive the rating scale in treatment Dummy (D) as being less kind compared to the rating scale used in treatment No Dummy (ND).

We find again support that a dummy category is interpreted as "bad" news. Subjects receiving grade 3 - the lowest rating - evaluated the rating scale in treatment D as being less kind than the scale in treatment ND (Wilcoxon test, two-sided, $p = .000$). Subjects receiving grade 1 and 2, however, did not perceive the rating scale in treatment D as being less kind than in treatment ND (Wilcoxon test, two-sided, $p = .544$).

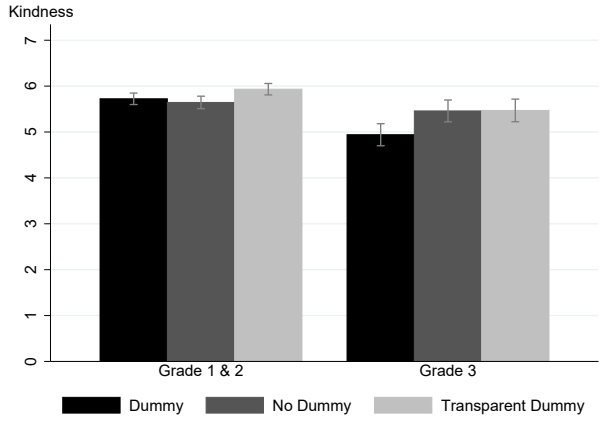
If individuals do know that the additional rating category is unused, employing a transparent dummy category may signal kindness and good intentions of employers as they refrain from using an available punishment option. Therefore, the literature suggests that subjects perceive the scale in treatment Transparent Dummy (TD) as being more kind compared to the rating scale used in treatment D and ND.

We find support that employing a transparent dummy category is interpreted as being kind. Compared to treatment D, subjects across performance classes evaluated the scale in treatment TD as more kind (Wilcoxon test, two-sided, $p = .011$ $p = .002$ comparing grade 1&2 and grade 3, respectively). Compared to treatment ND, only subjects receiving grade 1 and 2 evaluated the scale in treatment TD as more kind (Wilcoxon test, two-sided, $p = .003$ $p = .782$ comparing grade 1&2 and grade 3, respectively).

5. Conclusion

We studied performance effects of a dummy category in short-term interactions. Contrary to our hypotheses, a dummy category did not increase performance in our setting - independent of whether subjects were informed or not that the additional category is unused.

Figure 5 Kindness of Rating Scale Across Treatments



We expected that more generous ratings (*Evaluation Effect*) and higher incentives (*Incentive Effect*) in the presence of a dummy category increase performance and thus offset potential negative effects arising from employing a less kind rating scale (*negative Kindness-of-the-Scale Effect*).

We did, however, not see a performance increase in the presence of a dummy category. It seems that the opposing effects of a category outweigh each other. Subjects ranking lowest increased performance in the presence of a dummy category while those ranking higher reduced performance: the *Incentive Effect* of a dummy category seems only present for those ranking lowest. Moreover, we found indication that subjects perceive the rating scale in treatment Dummy as being less kind.

We expected that employing a transparent dummy category increases performance by triggering positively reciprocal reactions to employing a more kind rating scale (*positive Kindness-of-the-Scale Effect*).

We did, however, not see a performance increase in the presence of a transparent dummy category either. Interestingly, we found indication that subjects perceive the rating scale in treatment Transparent Dummy as being more kind. However, these kindness perceptions did not translate into performance effects.

To sum up, we do not find evidence that a dummy category increases performance in short-term interactions but we find that individuals also pay attention to the kindness of rating scales. The kindness of a rating scale seems to be as important as the individual rating and incentives when subjects receive one or two performance ratings.

In practice, employees usually work over multiple periods and receive multiple consecutive performance ratings over a longer time period. It is thus a very important question which of the effects - *Incentive, Rating* or *Kindness-of-the-Scale Effect* - prevails or whether they cancel out in the long run. We leave this study to future research.

Acknowledgments

We thank the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) for financial support through the research unit "Design & Behavior" [FOR 1371; projects TP3 and TP5]. The experiment is registered under AEARCTR-0002736 and AEARCTR-0003029 (Study I and Study II, respectively). Please note that the change in the authorship is due to the fact that most of the work was done by Thomas Vogt.

Appendix A: A Novel Real Effort Task

The instructions in Appendix E-G show the working screen in our field experiment. We provided sets of artificially created exam cover sheets that contained a table of six handwritten grades. Employees' task was to enter the grades displayed at the top into the entry fields on the bottom of the screen.

We chose the appearance of the real effort task in line with a job (1.) MTurk workers are used to, (2.) a university department actually does and (3.) which is reasonably outsourced. The task is very similar to the jobs otherwise found in the Mechanical Turk marketplace in terms of the type of work, difficulty and time required. It is also common practice for university departments to digitize class results that were initially marked using pen and paper. To make it plausible that our class results had not been digitized yet, we chose exam from the year 2004.

Our real effort task is very tedious, non cognitively-demanding that primarily requires attention and can be solved without prior knowledge. It induces positive effort costs and potential learning effects can be neglected. Due to the nature of the task, we assume that intrinsic motivation does not play a role in our setting.

We assured that the individual grading on a cover sheet as well as the overall class grading followed a reasonable distribution. Furthermore, grades were written from the same person to eliminate difficulties due to different handwriting. We created an initial set of 200 cover sheets. Out of the initial set we created new sets by changing information on the cover sheet - the exam date and number of pages - leaving the handwritten grades untouched. Note that thereby the actual grades on the exam cover sheets were identical across sets. We randomly varied which set was displayed and assured that every subject saw a specific version only once. In addition, the display order of individual cover sheets within one set was randomly varied for each subject. In the Study I, subjects could enter up to 200, in Study II up to 400 cover sheets in week two.

We also paid attention to reducing noise in performance due to different technical skills and varying technical conditions. We did so by restricting the input of grades to capital letters and did not allow any special character other than "+" and "-". Moreover, we required a screen resolution of at least 1200 (width) x 700 (height) such that none of the subjects had to scroll while transferring the grades.

Appendix B: Sample Demographics

Table A1 Sample Demographics Study I (N=838)

Demographics	Percentage
Age[a]	38.21 (12.05)
Gender (female)	58.00
Highest level of education	
Less than High school degree	0.00
High school graduate	6.80
Vocational/technical school	6.56
Some college	31.74
Bachelor's degree	41.89
Master's degree	10.38
Doctoral degree	0.84
Advanced Professional Degree (JD, MD, MBA, etc.)	1.79
Employment status	
Working (paid employee)	67.30
Working (self-employed)	16.23
Not working	14.44
Other	2.03
Annual income from all sources before taxes	
$10,000 or less	15.39
$10,001 to $20,000	10.38
$20,001 to $30,000	9.90
$30,001 to $40,000	14.08
$40,001 to $50,000	12.41
$50,001 to $60,000	12.05
$60,001 to $70,000	5.85
$70,001 to $80,000	7.76
Over $80,000	12.17

Note: [a] Mean (Standard Deviation)

Table A2 Sample Demographics Study II (N=1,339)

Demographics	Percentage
Age[a]	36.16 (11.36)
Gender (female)	61.24
Highest level of education	
Less than High school degree	0.60
High school graduate	8.51
Vocational/technical school	5.15
Some college	28.23
Bachelor's degree	42.49
Master's degree	12.40
Doctoral degree	1.12
Advanced Professional Degree (JD, MD, MBA, etc.)	1.49
Employment status	
Working (paid employee)	65.42
Working (self-employed)	16.36
Not working	15.46
Other	2.76
Annual income from all sources before taxes	
$10,000 or less	12.85
$10,001 to $20,000	10.68
$20,001 to $30,000	15.24
$30,001 to $40,000	12.70
$40,001 to $50,000	11.73
$50,001 to $60,000	11.80
$60,001 to $70,000	8.14
$70,001 to $80,000	7.54
Over $80,000	9.33

Note: [a] Mean (Standard Deviation)

Appendix C: Robustness Checks of Regressions in Table 1 & Table 2

Table A3 Impact of a Dummy Category on Individual Performance II

Dependent Variable:	Study I		Study II					
Number of Cover Sheets Entered Correctly	ND vs. D		ND vs. D		ND vs. TD		D vs. TD	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dummy Category	-4.17	-4.37	-3.44	-7.57*			-4.62	-5.61
	(2.92)	(3.67)	(3.35)	(4.07)			(2.95)	(3.52)
Dummy Category#Grade 3 in t-1		0.75		16.70**				3.78
		(5.61)		(6.81)				(6.37)
Transparent Dummy Category					1.27	-1.98		
					(3.37)	(4.02)		
Transparent Dummy Category#Grade 3 in t-1						12.86*		
						(7.26)		
Grade 3 in t-1		-1.50		-15.54**		-16.21**		-0.13
		(5.42)		(6.97)		(6.79)		(6.10)
Pre-round Performance	0.61***	0.59***	0.97***	0.90***	0.98***	0.88***	0.93***	0.95***
	(0.05)	(0.09)	(0.06)	(0.09)	(0.05)	(0.08)	(0.05)	(0.07)
Constant	12.57***	13.73**	24.38***	33.28***	23.62***	34.33***	28.49***	27.32***
	(3.63)	(6.80)	(4.39)	(7.64)	(4.27)	(7.19)	(3.97)	(6.42)
Observations	946	946	928	928	934	934	934	934

Note: Tobit regressions on individual output are performed. D, Dummy Treatment; ND, No Dummy Treatment; TD, Transparent Dummy Treatment

* p < 0.10, ** p < 0.05, *** p <0.01. Robust standard errors are in parentheses, clustered at the individual level.

Table A4 Impact of a Dummy Category on Individual Performance III

Dependent Variable:	Study I		Study II					
Number of Cover Sheets Entered Correctly	ND vs. D		ND vs. D		ND vs. TD		D vs. TD	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dummy Category	-3.73	-4.67	-3.16	-6.70			-3.46	-4.45
	(2.81)	(3.59)	(3.38)	(4.12)			(2.98)	(3.57)
Dummy Category#Grade 3 in t-1		3.50		14.41**				3.79
		(5.32)		(6.74)				(6.36)
Transparent Dummy Category					1.34	-0.59		
					(3.32)	(3.97)		
Transparent Dummy Category#Grade 3 in t-1						7.56		
						(7.10)		
Grade 3 in t-1		-2.40		-14.26**		-12.71*		-1.22
		(5.31)		(6.78)		(6.55)		(6.13)
Pre-round Performance	0.56***	0.56***	0.94***	0.87***	0.95***	0.86***	0.92***	0.93***
	(0.05)	(0.08)	(0.06)	(0.08)	(0.05)	(0.08)	(0.05)	(0.07)
Constant	-9.48***	-6.88	75.21	84.42*	87.54**	98.11***	26.78**	26.63**
	(1.63)	(7.35)	(46.69)	(46.34)	(36.72)	(36.77)	(11.78)	(12.81)
Observations	946	946	928	928	934	934	934	934
Session Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note: OLS regressions on individual output are performed. Session dummies are included. D, Dummy Treatment; ND, No Dummy Treatment; TD, Transparent Dummy Treatment

* p < 0.10, ** p < 0.05, *** p <0.01. Robust standard errors are in parentheses, clustered at the individual level.

Appendix D: Tests of Satisfaction Levels and Perceived Kindness Across Treatments

Table A5 **P-values of Tests Comparing Satisfactions Levels Across Treatments**

	Satisfaction in Study I			Satisfaction in Study II		
	Grade 1	Grade 2	Grade 3	Grade 1	Grade 2	Grade 3
D vs. ND	.607	.332	.296	.752	.574	.862
D vs. TD	-	-	-	.511	.821	.315
ND vs. TD	-	-	-	.734	.756	.232

Note: We report p-values of two-sided Wilcoxon ranksum tests.

Table A6 **P-values of Tests Comparing the Perceived Kindness Across Treatments**

	Scale Kindness in Study I		Scale Kindness in Study II	
	Grade 1 & 2	Grade 3	Grade 1 & 2	Grade 3
D vs. ND	.000	.000	.544	.004
D vs. TD	-	-	.011	.002
ND vs. TD	-	-	.003	.782

Note: We report p-values of two-sided Wilcoxon ranksum tests.

Appendix E: Instructions Week One

Instructions | Quiz | Task | End

Welcome to our task

We are academics who value your work and always pay as promised. To participate in this HIT you must answer a short quiz correctly. Your compensation will consist of two components, the **fixed amount** that you earn for this HIT **plus a bonus payment** on Mechanical Turk**.**

You will receive a validation code for this HIT. **You must enter this validation code into the Mechanical Turk HIT in order to receive your payment.**

Please type "yes" into the field below to indicate that you have read the text above carefully and understood that you must enter the validation code into the Mechanical Turk HIT to receive your payment.

yes

Please click Continue for further instructions.

Continue

Survey Powered By Qualtrics

Instructions | Quiz | Task | End

Instructions

Your task is to update a database on class grades. We provide scanned cover sheets of exam papers. **Your task is to enter the handwritten grades into our database.** You can **navigate** through the entry fields **using the tab key**  .

We have two sets of cover sheets that are assigned to two HITs. That is, one set of cover sheets for each HIT. Today, **you can work on the first HIT for up to 20 minutes**. We will e-mail you a link via Mechanical Turk to **the second HIT on Monday next week**. Once you get the link, you will have **four days to work on the second HIT**.

To make sure that we can pay you once you started working, we will provide the validation code before you work on this task.

So, this job comprises two HITs, this first HIT today and the second HIT available on Monday next week.

Continue

Survey Powered By Qualtrics

Instructions | Quiz | Task | End

Payment

You will receive a **fixed payment of $2.25 and an additional bonus payment based on your relative performance.**
We will screen your work and assess your performance. Your **performance is assessed based on how many cover sheets you enter correctly** in this HIT. A cover sheet is evaluated as entered correctly **only if all grades, i.e. letters and "+" or "-" are entered correctly.**
The bonus payments are assigned based on your relative performance compared to all workers who work on the first HIT. **The higher you rank** compared to the other workers, **the higher will be your bonus payment**.

We will e-mail you a link via Mechanical Turk to the second HIT on Monday next week. Please follow the instructions in the e-mail to see and receive your bonus for the first HIT and to start the second HIT.

Please click Continue to start with the quiz.

Back | Continue

Survey Powered By Qualtrics

Instructions | Quiz | Task | End

Quiz

Please answer the following quiz. If you do not answer all questions correctly in the first attempt, **you can correct your answers once**. **If you fail to answer** all questions **correctly in the second attempt, you cannot work on this task**.

How many HITs comprises this job?

- ○ 1
- ◉ 2
- ○ 3

You can **navigate** through the entry fields **using**

- ◉ The **tab key** 
- ○ The **enter key** 
- ○ The **shift key** 

After you worked on this HIT **we will**

- ◉ **assess your performance based on** how many **cover sheets** you **entered correctly** in this HIT. A **cover sheet is** evaluated as **entered correctly** only **if all grades, i.e. letters and "+" or "-" are entered correctly.**
- ○ **assess your performance based on** how many **cover sheets** you **entered correctly** in this HIT. A **cover sheet is** evaluated as **entered correctly** only **if all grades, i.e. letters without "+" or "-" are entered correctly.**
- ○ **assess your performance based on** how many **grades** you **entered** in this HIT. It **does not matter whether all grades** on a cover sheet **are entered correctly**.

What is your **payment for this HIT?**

- ○ You will receive a **fixed payment of $2.25** for this HIT. There will be **no bonus payment**.
- ◉ You will receive a **fixed payment of $2.25** for this HIT. Additionally, you will receive a **bonus payment** that **depends on your relative performance** on this HIT. **The higher you rank** compared to the other workers, **the higher will be your bonus payment**.
- ○ You will receive a **fixed payment of $2.25** for this HIT. Additionally, you will receive a **bonus payment** that **depends on your relative performance** on this HIT. **The lower you rank** compared to the other workers, **the higher will be your bonus payment**.

Back | Continue

Survey Powered By Qualtrics

Instructions | Quiz | **Task** | End

Validation Code

Please **enter the validation code below into the Mechanical Turk HIT now** in order to receive your payment**.**

Validation code: 8957275

Do not click Continue before you entered the validation code into the Mechanical Turk HIT. Otherwise, you cannot be paid.

Please indicate that you understood how you can get paid by choosing the right answer:

- ◉ I must **enter** the above **validation code** into the Mechanical Turk HIT **now**. If I continue without having entered the validation code, I will not be paid.
- ○ I can **enter** the above **validation code** into the Mechanical Turk HIT **later**. If I continue without having entered the validation code, I will be paid later.
- ○ I do **not** have to **enter** the above **validation code** into the Mechanical Turk HIT.

Continue

Instructions | Quiz | **Task** | End

Validation Code

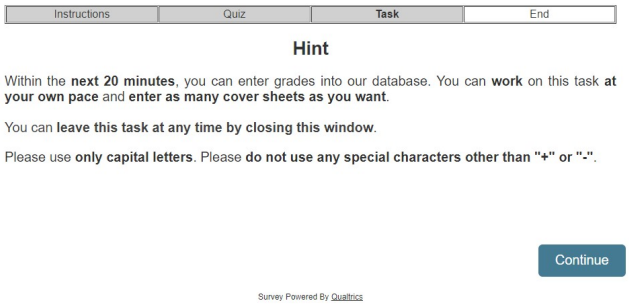
Did you **enter the validation code 8957275 into the Mechanical Turk HIT?**

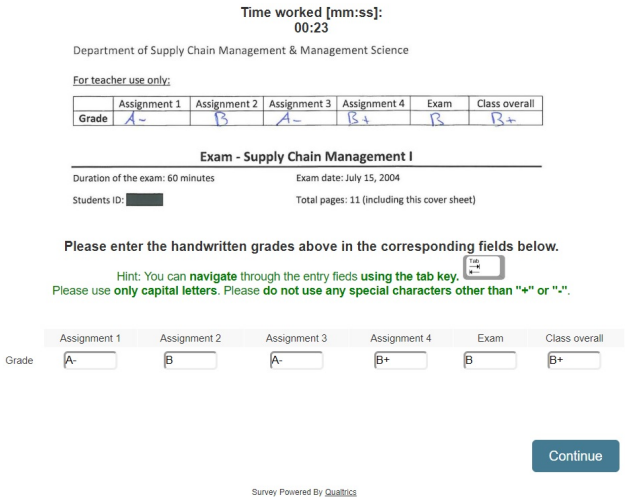
- ◉ Yes
- ○ No

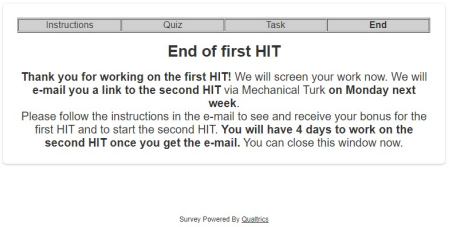
Do not click Continue before you entered the validation code into the Mechanical Turk HIT. Otherwise, you cannot be paid.

Back | Continue

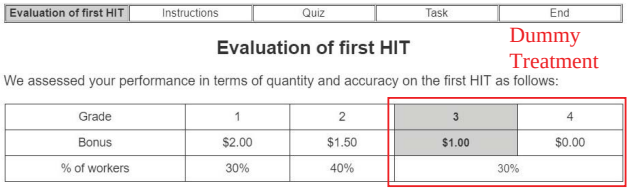
Survey Powered By Qualtrics


Instructions
Quiz
Task
End
Hint
Within the next 20 minutes, you can enter grades into our database. You can work on this task at your own pace and enter as many cover sheets as you want.
You can leave this task at any time by closing this window.
Please use only capital letters. Please do not use any special characters other than "+" or "-".
Continue
Survey Powered By Qualtrics


Time worked [mm:ss]:
00:23
Department of Supply Chain Management & Management Science
For teacher use only:
Assignment 1 Assignment 2 Assignment 3 Assignment 4 Exam Class overall
Grade A- B A- B+ B B+
Exam - Supply Chain Management I
Duration of the exam: 60 minutes
Exam date: July 15, 2004
Students ID:
Total pages: 11 (including this cover sheet)
Please enter the handwritten grades above in the corresponding fields below.
Hint: You can navigate through the entry fieds using the tab key.
Please use only capital letters. Please do not use any special characters other than "+" or "-".
Assignment 1 Assignment 2 Assignment 3 Assignment 4 Exam Class overall
Grade A- B A- B+ B B+
Continue
Survey Powered By Qualtrics


Instructions
Quiz
Task
End
End of first HIT
Thank you for working on the first HIT! We will screen your work now. We will e-mail you a link to the second HIT via Mechanical Turk on Monday next week.
Please follow the instructions in the e-mail to see and receive your bonus for the first HIT and to start the second HIT. You will have 4 days to work on the second HIT once you get the e-mail. You can close this window now.
Survey Powered By Qualtrics

Appendix F: Instructions Week Two Study I



Evaluation of first HIT

Dummy Treatment

We assessed your performance in terms of quantity and accuracy on the first HIT as follows:

Grade	1	2	3	4
Bonus	$2.00	$1.50	$1.00	$0.00
% of workers	30%	40%	30%	

We will pay you the bonus of $1.00 for the first HIT on Mechanical Turk irrespective of whether you work on the second HIT or not.

Please enter your grade for the first HIT.

3

Please indicate how satisfied you are with the evaluation.

Not satisfied at all ○ ○ ○ ○ ○ ◉ ○ Extremely satisfied

Please click Continue to start the second HIT. If you don't want to work on the second HIT, please click Leave Task.

Continue

Leave Task

Are you sure you do not want to work on the second HIT?

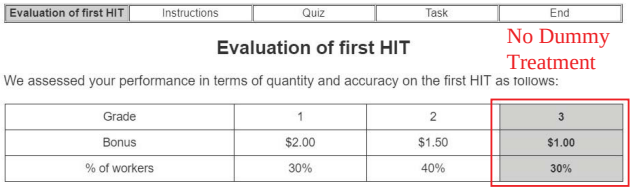
If you want to work on the second HIT, please click "No, Go Back".

No, Go Back

Yes

Survey Powered By Qualtrics

screen seen only if "Leave Task" was selected on the previous screen



Evaluation of first HIT

No Dummy Treatment

We assessed your performance in terms of quantity and accuracy on the first HIT as follows:

Grade	1	2	3
Bonus	$2.00	$1.50	$1.00
% of workers	30%	40%	30%

We will pay you the bonus of $1.00 for the first HIT on Mechanical Turk irrespective of whether you work on the second HIT or not.

Please enter your grade for the first HIT.

3

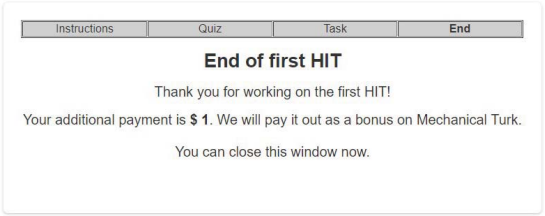
Please indicate how satisfied you are with the evaluation.

Not satisfied at all ○ ○ ○ ○ ○ ◉ ○ Extremely satisfied

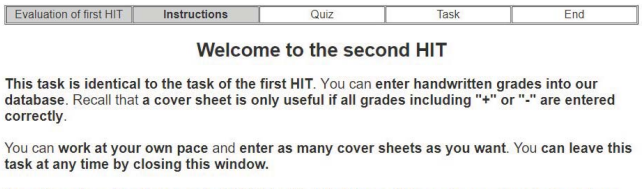
Please click Continue to start the second HIT. If you don't want to work on the second HIT, please click Leave Task.

Continue

Leave Task



screen seen only if "Yes" was selected on the previous screen; these subjects did not participate in week two



Welcome to the second HIT

This task is identical to the task of the first HIT. You can **enter handwritten grades into our database**. Recall that **a cover sheet is only useful if all grades including "+" or "-" are entered correctly**.

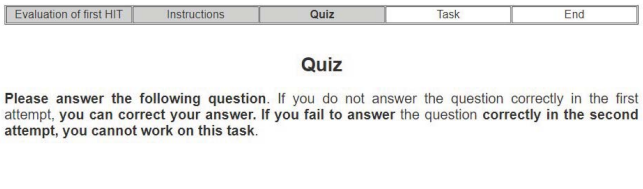
You can **work at your own pace** and **enter as many cover sheets as you want**. You **can leave this task at any time by closing this window.**

You will receive a **fixed payment of $2.25** for this HIT. **There will be no bonus payments and no grading.**

Please click Continue to start with the Quiz.

Continue

Survey Powered By Qualtrics



Quiz

Please answer the following question. If you do not answer the question correctly in the first attempt, **you can correct your answer. If you fail to answer** the question **correctly in the second attempt, you cannot work on this task**.

What is your payment for this HIT?

◉ You will receive a **fixed payment of $2.25** for this HIT. There will be **no bonus payment and no grading**.

○ You will receive a **fixed payment of $2.25** for this HIT. Additionally, you will receive a **bonus payment** that **depends on your relative performance** on this HIT. **The higher you rank** compared to the other workers, **the higher will be your bonus payment**.

○ You will receive a **fixed payment of $2.25** for this HIT. Additionally, you will receive a **bonus payment** that **depends on your relative performance** on this HIT. **The lower you rank** compared to the other workers, **the higher will be your bonus payment**.

Back

Continue

Survey Powered By Qualtrics

Evaluation of first HIT | Instructions | Quiz | **Task** | End

Validation Code

Please **enter the validation code below into the Mechanical Turk HIT now** in order to receive your fixed payment of $2.25.

Validation code: 2610728

Do not click Continue before you entered the validation code into the Mechanical Turk HIT. Otherwise, you cannot be paid.

Please indicate what you have to do to get paid by choosing the right answer:

- (•) I must **enter** the above **validation code** into the Mechanical Turk HIT **now**. If I continue without having entered the validation code, I will not be paid.
- ( ) I can **enter** the above **validation code** into the Mechanical Turk HIT **later**. If I continue without having entered the validation code, I will be paid later.
- ( ) I do **not** have to **enter** the above **validation code** into the Mechanical Turk HIT.

Continue

Survey Powered By Qualtrics

Evaluation of first HIT | Instructions | Quiz | **Task** | End

Validation Code

Did you **enter the validation code 2610728 into the Mechanical Turk HIT?**

- (•) yes
- ( ) no

Do not click Continue before you entered the validation code into the Mechanical Turk HIT. Otherwise, you cannot be paid.

Back | Continue

Survey Powered By Qualtrics

Evaluation of first HIT | Instructions | Quiz | **Task** | End

Hint

You can **work** on this task **at your own pace and as long as you want. You can enter as many cover sheets as you like**.

You can **leave this task at any time by closing this window**.

You can **navigate** though the entry fields **using the tab key**.

Please use **only capital letters**. Please **do not use any special characters other than "+" or "-"**.

Please click Continue to start working.

Continue

Survey Powered By Qualtrics

Time worked [mm:ss]:
00:26

Department of Supply Chain Management & Management Science

For teacher use only:

	Assignment 1	Assignment 2	Assignment 3	Assignment 4	Exam	Class overall
Grade	B	D	B+	C+	B+	B-

Exam - Supply Chain Management I

Duration of the exam: 60 minutes — Exam date: July 15, 2004

Students ID: [redacted] — Total pages: 11 (including this cover sheet)

Please enter the handwritten grades above in the corresponding fields below.

Hint: You can **navigate** through the entry fieds **using the tab key.**
Please use **only capital letters**. Please **do not use any special characters other than "+" or "-"**.

	Assignment 1	Assignment 2	Assignment 3	Assignment 4	Exam	Class overall
Grade	B	D	B+	C+	B+	B-

Next

Survey Powered By Qualtrics

You worked for more than 20 minutes. Feel free to continue working. If you want to leave this HIT, just close this window.

Time worked [mm:ss]:
21:15

Department of Supply Chain Management & Management Science

For teacher use only:

	Assignment 1	Assignment 2	Assignment 3	Assignment 4	Exam	Class overall
Grade	A-	B	B-	B	B+	B+

Exam - Supply Chain Management I

Duration of the exam: 60 minutes — Exam date: July 15, 2004

Students ID: [redacted] — Total pages: 11 (including this cover sheet)

Please enter the handwritten grades above in the corresponding fields below.

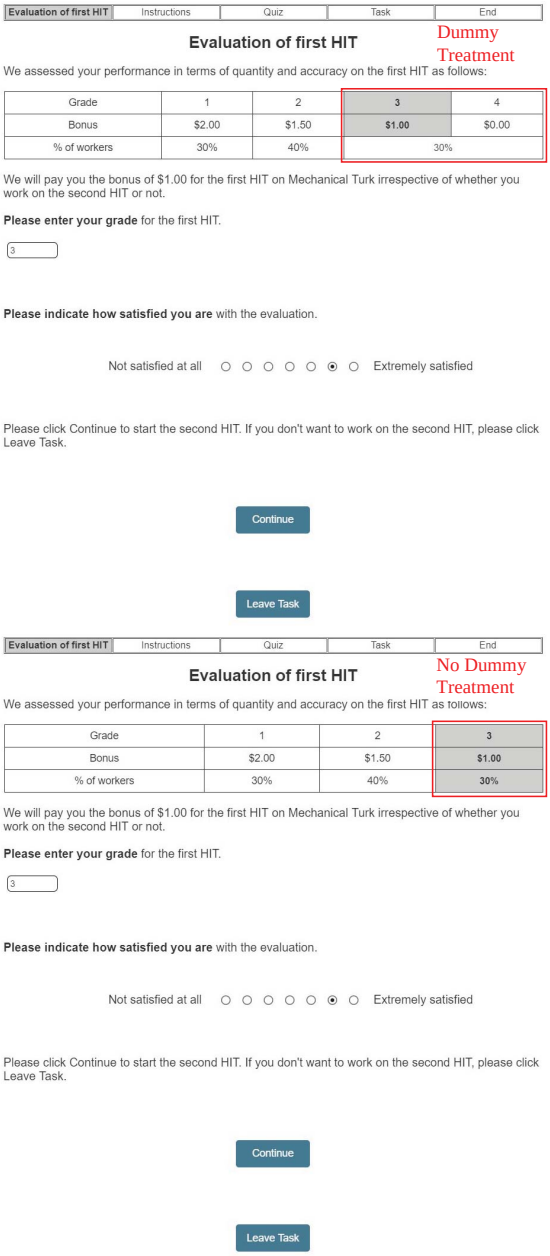
Hint: You can **navigate** through the entry fieds **using the tab key.**
Please use **only capital letters**. Please **do not use any special characters other than "+" or "-"**.

	Assignment 1	Assignment 2	Assignment 3	Assignment 4	Exam	Class overall
Grade	A-	B	B-	B	B+	B+

Next

Survey Powered By Qualtrics

Appendix G: Instructions Week Two Study II



Evaluation of first HIT | Instructions | Quiz | Task | End

Evaluation of first HIT

Dummy Treatment

We assessed your performance in terms of quantity and accuracy on the first HIT as follows:

Grade	1	2	3	4
Bonus	$2.00	$1.50	$1.00	$0.00
% of workers	30%	40%	30%	

We will pay you the bonus of $1.00 for the first HIT on Mechanical Turk irrespective of whether you work on the second HIT or not.

Please enter your grade for the first HIT.

3

Please indicate how satisfied you are with the evaluation.

Not satisfied at all ○ ○ ○ ○ ○ ◉ ○ Extremely satisfied

Please click Continue to start the second HIT. If you don't want to work on the second HIT, please click Leave Task.

Continue

Leave Task

Evaluation of first HIT | Instructions | Quiz | Task | End

Evaluation of first HIT

No Dummy Treatment

We assessed your performance in terms of quantity and accuracy on the first HIT as follows:

Grade	1	2	3
Bonus	$2.00	$1.50	$1.00
% of workers	30%	40%	30%

We will pay you the bonus of $1.00 for the first HIT on Mechanical Turk irrespective of whether you work on the second HIT or not.

Please enter your grade for the first HIT.

3

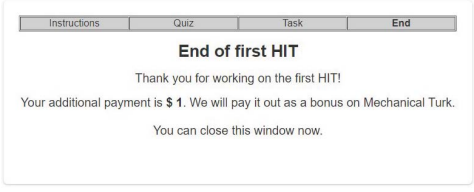
Please indicate how satisfied you are with the evaluation.

Not satisfied at all ○ ○ ○ ○ ○ ◉ ○ Extremely satisfied

Please click Continue to start the second HIT. If you don't want to work on the second HIT, please click Leave Task.

Continue

Leave Task



Instructions | Quiz | Task | End

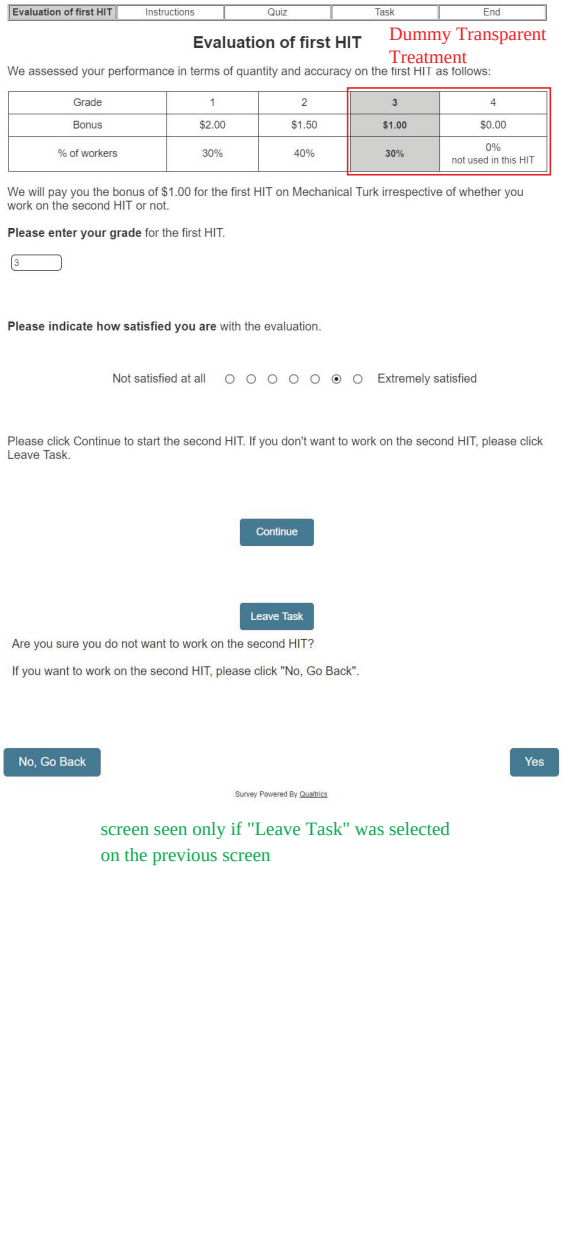
End of first HIT

Thank you for working on the first HIT!

Your additional payment is **$ 1**. We will pay it out as a bonus on Mechanical Turk.

You can close this window now.

screen seen only if "Yes" was selected on the previous screen; Survey Powered By Qualtrics these subjects did not participate in week two



Evaluation of first HIT | Instructions | Quiz | Task | End

Evaluation of first HIT

Dummy Transparent Treatment

We assessed your performance in terms of quantity and accuracy on the first HIT as follows:

Grade	1	2	3	4
Bonus	$2.00	$1.50	$1.00	$0.00
% of workers	30%	40%	30%	0% not used in this HIT

We will pay you the bonus of $1.00 for the first HIT on Mechanical Turk irrespective of whether you work on the second HIT or not.

Please enter your grade for the first HIT.

3

Please indicate how satisfied you are with the evaluation.

Not satisfied at all ○ ○ ○ ○ ○ ◉ ○ Extremely satisfied

Please click Continue to start the second HIT. If you don't want to work on the second HIT, please click Leave Task.

Continue

Leave Task

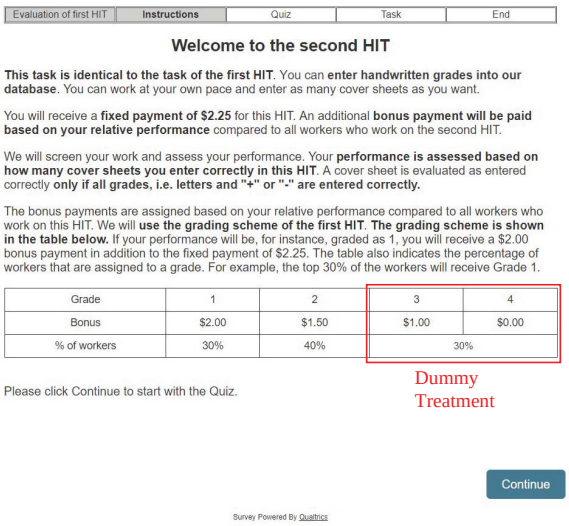
Are you sure you do not want to work on the second HIT?

If you want to work on the second HIT, please click "No, Go Back".

No, Go Back Yes

Survey Powered By Qualtrics

screen seen only if "Leave Task" was selected on the previous screen



Evaluation of first HIT | **Instructions** | Quiz | Task | End

Welcome to the second HIT

This task is identical to the task of the first HIT. You can **enter handwritten grades into our database**. You can work at your own pace and enter as many cover sheets as you want.

You will receive a **fixed payment of $2.25** for this HIT. An additional **bonus payment will be paid based on your relative performance** compared to all workers who work on the second HIT.

We will screen your work and assess your performance. Your **performance is assessed based on how many cover sheets you enter correctly in this HIT**. A cover sheet is evaluated as entered correctly **only if all grades, i.e. letters and "+" or "-" are entered correctly.**

The bonus payments are assigned based on your relative performance compared to all workers who work on this HIT. We will **use the grading scheme of the first HIT**. **The grading scheme is shown in the table below.** If your performance will be, for instance, graded as 1, you will receive a $2.00 bonus payment in addition to the fixed payment of $2.25. The table also indicates the percentage of workers that are assigned to a grade. For example, the top 30% of the workers will receive Grade 1.

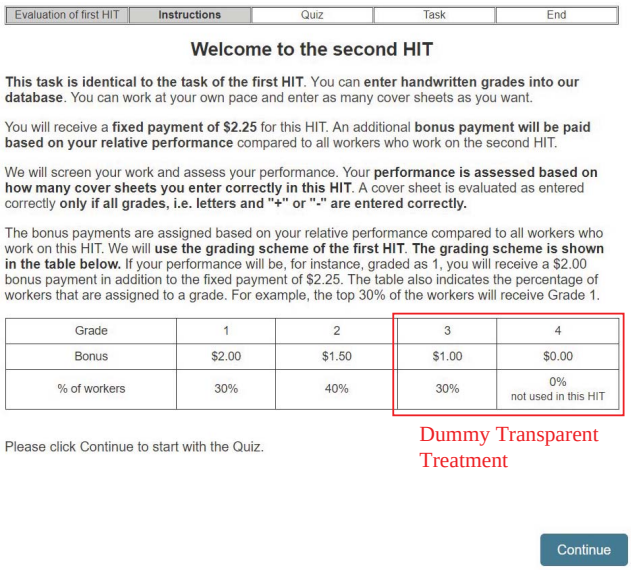
Grade	1	2	3	4
Bonus	$2.00	$1.50	$1.00	$0.00
% of workers	30%	40%	30%	

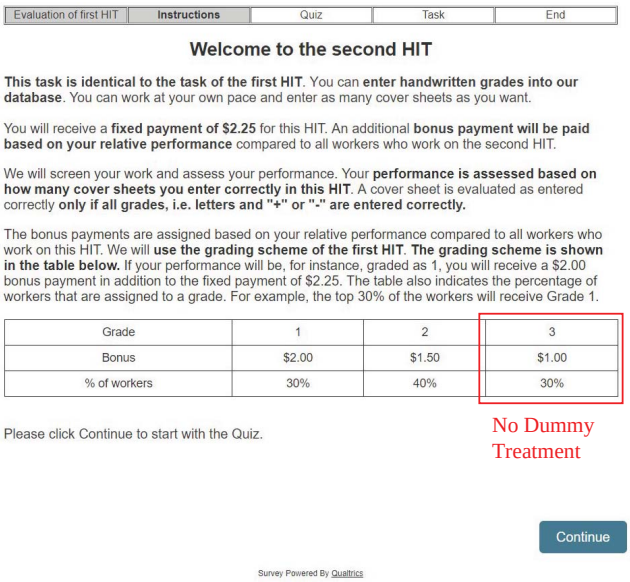
Dummy Treatment

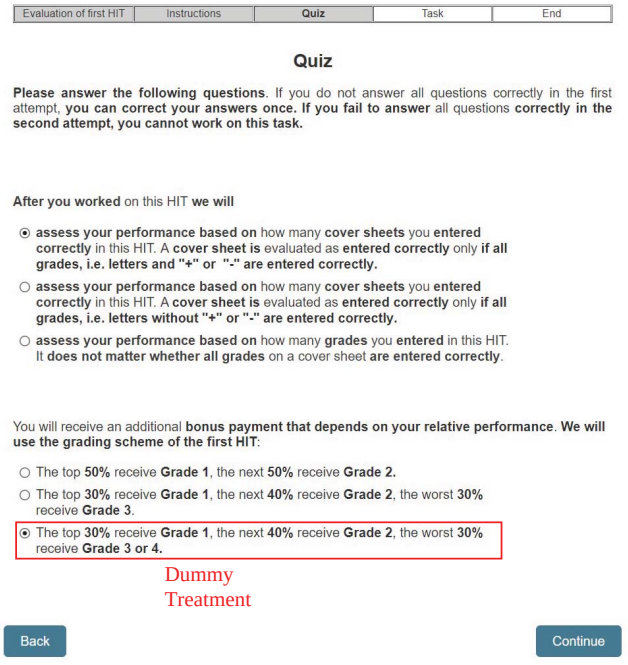
Please click Continue to start with the Quiz.

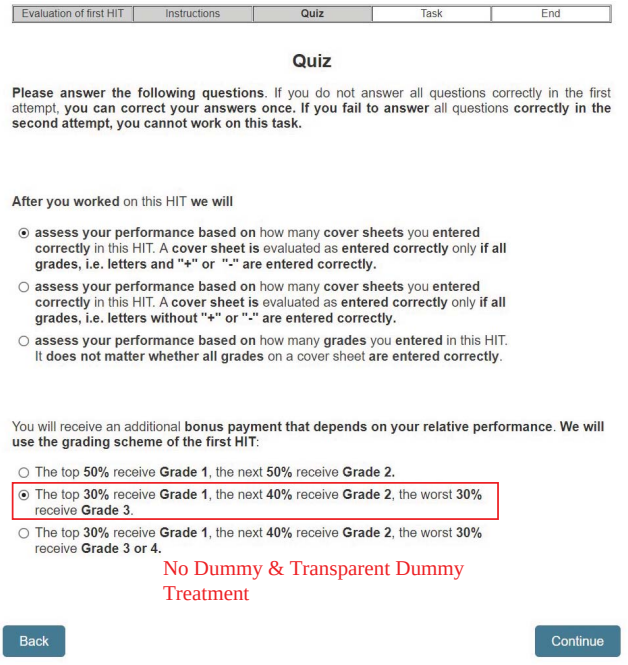
Continue

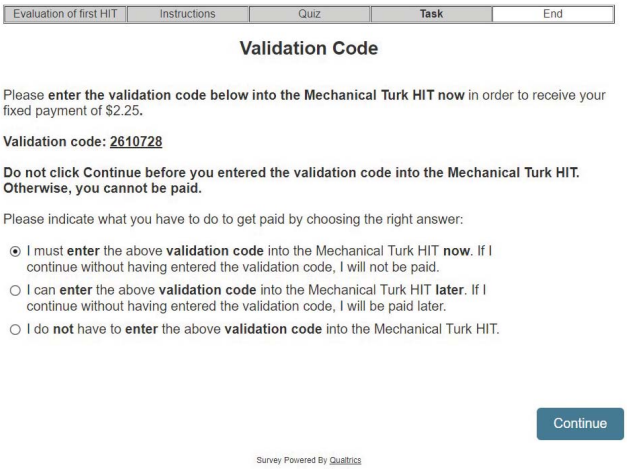
Survey Powered By Qualtrics

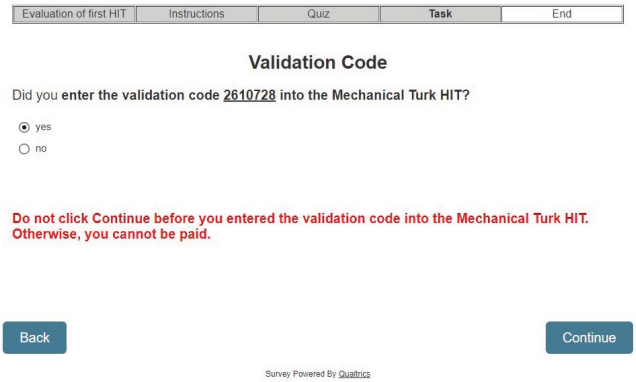

Evaluation of first HIT
Instructions
Quiz
Task
End
Welcome to the second HIT
This task is identical to the task of the first HIT. You can enter handwritten grades into our database. You can work at your own pace and enter as many cover sheets as you want.
You will receive a fixed payment of $2.25 for this HIT. An additional bonus payment will be paid based on your relative performance compared to all workers who work on the second HIT.
We will screen your work and assess your performance. Your performance is assessed based on how many cover sheets you enter correctly in this HIT. A cover sheet is evaluated as entered correctly only if all grades, i.e. letters and "+" or "-" are entered correctly.
The bonus payments are assigned based on your relative performance compared to all workers who work on this HIT. We will use the grading scheme of the first HIT. The grading scheme is shown in the table below. If your performance will be, for instance, graded as 1, you will receive a $2.00 bonus payment in addition to the fixed payment of $2.25. The table also indicates the percentage of workers that are assigned to a grade. For example, the top 30% of the workers will receive Grade 1.
Grade 1 2 3 4
Bonus $2.00 $1.50 $1.00 $0.00
% of workers 30% 40% 30% 0% not used in this HIT
Please click Continue to start with the Quiz.
Dummy Transparent Treatment
Continue
Survey Powered By Qualtrics

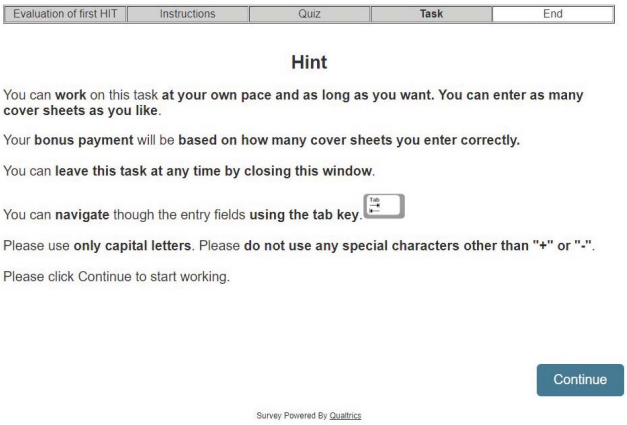

Evaluation of first HIT
Instructions
Quiz
Task
End
Welcome to the second HIT
This task is identical to the task of the first HIT. You can enter handwritten grades into our database. You can work at your own pace and enter as many cover sheets as you want.
You will receive a fixed payment of $2.25 for this HIT. An additional bonus payment will be paid based on your relative performance compared to all workers who work on the second HIT.
We will screen your work and assess your performance. Your performance is assessed based on how many cover sheets you enter correctly in this HIT. A cover sheet is evaluated as entered correctly only if all grades, i.e. letters and "+" or "-" are entered correctly.
The bonus payments are assigned based on your relative performance compared to all workers who work on this HIT. We will use the grading scheme of the first HIT. The grading scheme is shown in the table below. If your performance will be, for instance, graded as 1, you will receive a $2.00 bonus payment in addition to the fixed payment of $2.25. The table also indicates the percentage of workers that are assigned to a grade. For example, the top 30% of the workers will receive Grade 1.
Grade 1 2 3
Bonus $2.00 $1.50 $1.00
% of workers 30% 40% 30%
Please click Continue to start with the Quiz.
No Dummy Treatment
Continue
Survey Powered By Qualtrics

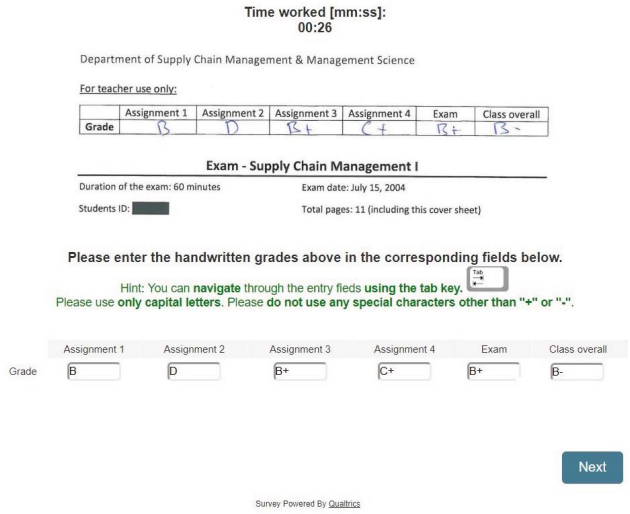

Evaluation of first HIT
Instructions
Quiz
Task
End
Quiz
Please answer the following questions. If you do not answer all questions correctly in the first attempt, you can correct your answers once. If you fail to answer all questions correctly in the second attempt, you cannot work on this task.
After you worked on this HIT we will
assess your performance based on how many cover sheets you entered correctly in this HIT. A cover sheet is evaluated as entered correctly only if all grades, i.e. letters and "+" or "-" are entered correctly.
assess your performance based on how many cover sheets you entered correctly in this HIT. A cover sheet is evaluated as entered correctly only if all grades, i.e. letters without "+" or "-" are entered correctly.
assess your performance based on how many grades you entered in this HIT. It does not matter whether all grades on a cover sheet are entered correctly.
You will receive an additional bonus payment that depends on your relative performance. We will use the grading scheme of the first HIT:
The top 50% receive Grade 1, the next 50% receive Grade 2.
The top 30% receive Grade 1, the next 40% receive Grade 2, the worst 30% receive Grade 3.
The top 30% receive Grade 1, the next 40% receive Grade 2, the worst 30% receive Grade 3 or 4.
Dummy Treatment
Back
Continue
Survey Powered By Qualtrics

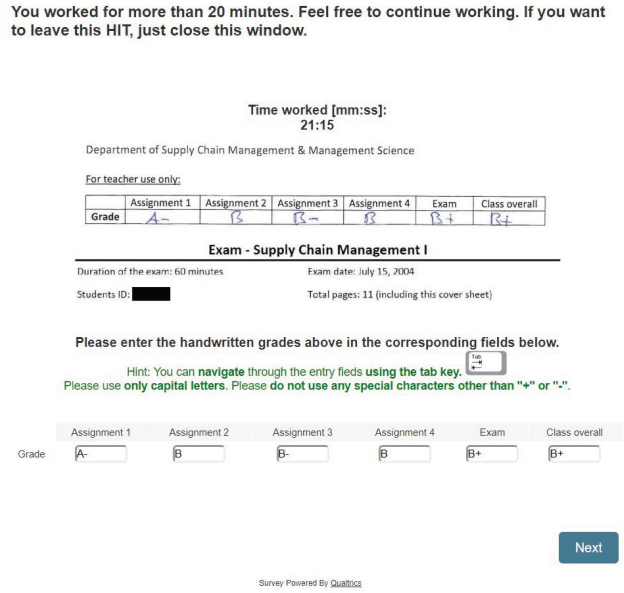

Evaluation of first HIT
Instructions
Quiz
Task
End
Quiz
Please answer the following questions. If you do not answer all questions correctly in the first attempt, you can correct your answers once. If you fail to answer all questions correctly in the second attempt, you cannot work on this task.
After you worked on this HIT we will
assess your performance based on how many cover sheets you entered correctly in this HIT. A cover sheet is evaluated as entered correctly only if all grades, i.e. letters and "+" or "-" are entered correctly.
assess your performance based on how many cover sheets you entered correctly in this HIT. A cover sheet is evaluated as entered correctly only if all grades, i.e. letters without "+" or "-" are entered correctly.
assess your performance based on how many grades you entered in this HIT. It does not matter whether all grades on a cover sheet are entered correctly.
You will receive an additional bonus payment that depends on your relative performance. We will use the grading scheme of the first HIT:
The top 50% receive Grade 1, the next 50% receive Grade 2.
The top 30% receive Grade 1, the next 40% receive Grade 2, the worst 30% receive Grade 3.
The top 30% receive Grade 1, the next 40% receive Grade 2, the worst 30% receive Grade 3 or 4.
No Dummy & Transparent Dummy Treatment
Back
Continue
Survey Powered By Qualtrics


Evaluation of first HIT
Instructions
Quiz
Task
End
Validation Code
Please enter the validation code below into the Mechanical Turk HIT now in order to receive your fixed payment of $2.25.
Validation code: 2610728
Do not click Continue before you entered the validation code into the Mechanical Turk HIT. Otherwise, you cannot be paid.
Please indicate what you have to do to get paid by choosing the right answer:
I must enter the above validation code into the Mechanical Turk HIT now. If I continue without having entered the validation code, I will not be paid.
I can enter the above validation code into the Mechanical Turk HIT later. If I continue without having entered the validation code, I will be paid later.
I do not have to enter the above validation code into the Mechanical Turk HIT.
Continue
Survey Powered By Qualtrics


Evaluation of first HIT
Instructions
Quiz
Task
End
Validation Code
Did you enter the validation code 2610728 into the Mechanical Turk HIT?
yes
no
Do not click Continue before you entered the validation code into the Mechanical Turk HIT. Otherwise, you cannot be paid.
Back
Continue
Survey Powered By Qualtrics


Evaluation of first HIT
Instructions
Quiz
Task
End
Hint
You can work on this task at your own pace and as long as you want. You can enter as many cover sheets as you like.
Your bonus payment will be based on how many cover sheets you enter correctly.
You can leave this task at any time by closing this window.
You can navigate though the entry fields using the tab key.
Please use only capital letters. Please do not use any special characters other than "+" or "-".
Please click Continue to start working.
Continue
Survey Powered By Qualtrics


Time worked [mm:ss]:
00:26
Department of Supply Chain Management & Management Science
For teacher use only:
Assignment 1 Assignment 2 Assignment 3 Assignment 4 Exam Class overall
Grade B D B+ C+ B+ B-
Exam - Supply Chain Management I
Duration of the exam: 60 minutes
Exam date: July 15, 2004
Students ID:
Total pages: 11 (including this cover sheet)
Please enter the handwritten grades above in the corresponding fields below.
Hint: You can navigate through the entry fieds using the tab key.
Please use only capital letters. Please do not use any special characters other than "+" or "-".
Assignment 1 Assignment 2 Assignment 3 Assignment 4 Exam Class overall
Grade B D B+ C+ B+ B-
Next
Survey Powered By Qualtrics


You worked for more than 20 minutes. Feel free to continue working. If you want to leave this HIT, just close this window.
Time worked [mm:ss]:
21:15
Department of Supply Chain Management & Management Science
For teacher use only:
Assignment 1 Assignment 2 Assignment 3 Assignment 4 Exam Class overall
Grade A- B B- B B+ B+
Exam - Supply Chain Management I
Duration of the exam: 60 minutes
Exam date: July 15, 2004
Students ID:
Total pages: 11 (including this cover sheet)
Please enter the handwritten grades above in the corresponding fields below.
Hint: You can navigate through the entry fieds using the tab key.
Please use only capital letters. Please do not use any special characters other than "+" or "-".
Assignment 1 Assignment 2 Assignment 3 Assignment 4 Exam Class overall
Grade A- B B- B B+ B+
Next
Survey Powered By Qualtrics

Appendix H: Post Trial Questionnaire

Welcome to our Survey

We are **academics who value your participation**. Your **responses** will be kept **confidential** and **anonymous**.

This survey will take about **3 minutes**.

You recently worked on our database update HITs - **Thank you for your effort!**
This is a **follow-up questinnaire** on our HITs. We **really appreciate your honest answers**.

At the end of the survey you will receive an individual validation code. **You must enter this validation code into the Mechanical Turk HIT and submit the HIT in order to receive your payment.**

Please click "Start Survey" to continue.

Start Survey

Survey Powered By Qualtrics

Questionnaire

No Dummy Treatment

We used the following evaluation scale in our HIT:

Grade	1	2	3
Bonus	$2.00	$1.50	$1.00
% of workers	30%	40%	30%

	very unkind	unkind	somewhat unkind	neutral	somewhat kind	kind	very kind
How kind is the evaluation scale:	○	○	○	○	○	○	○

Continue

Survey Powered By Qualtrics

Questionnaire

Dummy Treatment

We used the following evaluation scale in our HIT:

Grade	1	2	3	4
Bonus	$2.00	$1.50	$1.00	$0.00
% of workers	30%	40%	30	

	very unkind	unkind	somewhat unkind	neutral	somewhat kind	kind	very kind
How kind is the evaluation scale:	○	○	○	○	○	○	○

Continue

Survey Powered By Qualtrics

Questionnaire

Transparent Dummy Treatment

We used the following evaluation scale in our HIT:

Grade	1	2	3	4
Bonus	$2.00	$1.50	$1.00	$0.00
% of workers	30%	40%	30%	0% not used in this HIT

	very unkind	unkind	somewhat unkind	neutral	somewhat kind	kind	very kind
How kind is the evaluation scale:	○	○	○	○	○	○	○

Continue

Survey Powered By Qualtrics

Questionnaire

	Does not apply to me at all	Does not apply to me	Somewhat not applies to me	neutral	Somewhat applies to me	Applies to me	Applies to me perfectly
If someone does me a favor, I am prepared to return it	○	○	○	○	○	○	○
I go out of my way to help somebody who has been kind to me before	○	○	○	○	○	○	○
I am ready to undergo personal costs to help somebody who helped me before	○	○	○	○	○	○	○

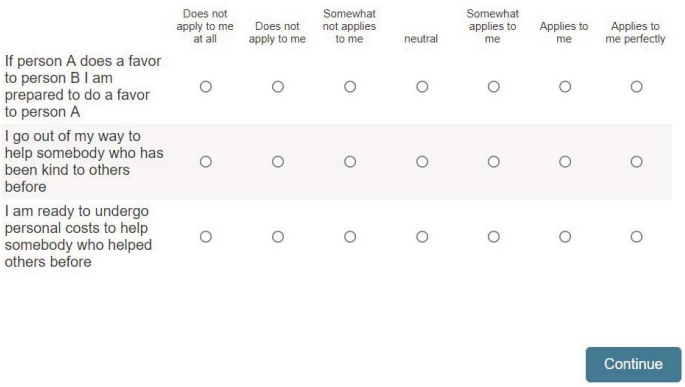
Continue

Survey Powered By Qualtrics

	Does not apply to me at all	Does not apply to me	Somewhat not applies to me	neutral	Somewhat applies to me	Applies to me	Applies to me perfectly
If I suffer a serious wrong, I will take revenge as soon as possible, no matter what the cost	○	○	○	○	○	○	○
If somebody puts me in a difficult position, I will do the same to him/her	○	○	○	○	○	○	○
If somebody offends me, I will offend him/her back	○	○	○	○	○	○	○

Continue

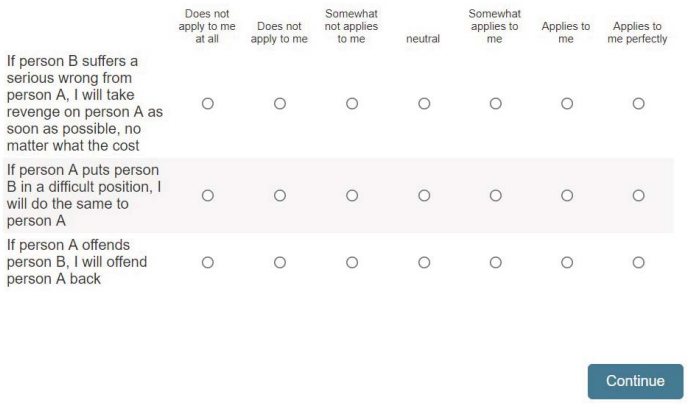
Survey Powered By Qualtrics



	Does not apply to me at all	Does not apply to me	Somewhat not applies to me	neutral	Somewhat applies to me	Applies to me	Applies to me perfectly
If person A does a favor to person B I am prepared to do a favor to person A	○	○	○	○	○	○	○
I go out of my way to help somebody who has been kind to others before	○	○	○	○	○	○	○
I am ready to undergo personal costs to help somebody who helped others before	○	○	○	○	○	○	○

Continue

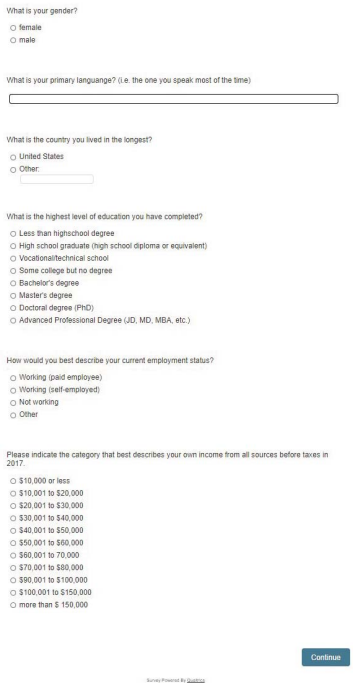
Survey Powered By Qualtrics



	Does not apply to me at all	Does not apply to me	Somewhat not applies to me	neutral	Somewhat applies to me	Applies to me	Applies to me perfectly
If person B suffers a serious wrong from person A, I will take revenge on person A as soon as possible, no matter what the cost	○	○	○	○	○	○	○
If person A puts person B in a difficult position, I will do the same to person A	○	○	○	○	○	○	○
If person A offends person B, I will offend person A back	○	○	○	○	○	○	○

Continue

Survey Powered By Qualtrics



What is your gender?

○ female
○ male

What is your primary languange? (i.e. the one you speak most of the time)

What is the country you lived in the longest?

○ United States
○ Other:

What is the highest level of education you have completed?

○ Less than highschool degree
○ High school graduate (high school diploma or equivalent)
○ Vocational/technical school
○ Some college but no degree
○ Bachelor's degree
○ Master's degree
○ Doctoral degree (PhD)
○ Advanced Professional Degree (JD, MD, MBA, etc.)

How would you best describe your current employment status?

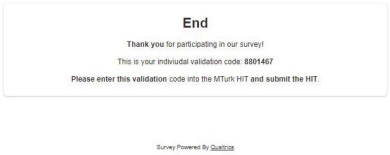
○ Working (paid employee)
○ Working (self-employed)
○ Not working
○ Other

Please indicate the category that best describes your own income from all sources before taxes in 2017.

○ $10,000 or less
○ $10,001 to $20,000
○ $20,001 to $30,000
○ $30,001 to $40,000
○ $40,001 to $50,000
○ $50,001 to $60,000
○ $60,001 to 70,000
○ $70,001 to $80,000
○ $90,001 to $100,000
○ $100,001 to $150,000
○ more than $ 150,000

Continue

Survey Powered By Qualtrics



End

Thank you for participating in our survey!

This is your indiviudal validation code: **8801467**

Please enter this validation code into the MTurk HIT **and submit the HIT.**

Survey Powered By Qualtrics

References

Adams JS (1963) Towards an Understanding of Inequity. *The Journal of Abnormal and Social Psychology* 67(5):422–436.

Akerlof GA (1982) Labor Contracts as Partial Gift Exchange. *The Quarterly Journal of Economics* 97(4):543.

Akerlof GA, Yellen JL (1988) Fairness and Unemployment. *The American Economic Review* 78(2):44–49.

Akerlof GA, Yellen JL (1990) The Fair Wage-Effort Hypothesis and Unemployment. *The Quarterly Journal of Economics* 105(2):255–283.

Barankay I (2011) Rankings and Social Tournaments: Evidence From a Crowd-Sourcing Experiment. *Working Paper, University of Pennsylvania, Philadelphia* .

Barankay I (2012) Rank Incentives: Evidence From a Randomized Workplace Experiment. *Working Paper, University of Pennsylvania, Philadelphia* .

Berger J, Harbring C, Sliwka D (2013) Performance Appraisals and the Impact of Forced Distribution–An Experimental Investigation. *Management Science* 59(1):54–68.

Bol JC (2011) The Determinants and Performance Effects of Managers' Performance Evaluation Biases. *Accounting Review* 86(5):1549–1575.

Bowles S, Polanía-Reyes S (2012) Economic Incentives and Social Preferences: Substitutes or Complements? *Journal of Economic Literature* 50(2):368–425.

Chandler J, Mueller P, Paolacci G (2014) Nonnaïveté Among Amazon Mechanical Turk Workers: Consequences and Solutions for Behavioral Researchers. *Behavior Research Methods* 46(1):112–130.

Charness G (2004) Attribution and Reciprocity in an Experimental Labor Market. *Journal of Labor Economics* 22(3):665–688.

Chung DJ, Narayandas D (2017) Incentives Versus Reciprocity: Insights from a Field Experiment. *Journal of Marketing Research* 54(4):511–524.

Dufwenberg M, Kirchsteiger G (2004) A theory of sequential reciprocity. *Games and Economic Behavior* 47(2):268–298, ISSN 08998256, URL `http://dx.doi.org/10.1016/j.geb.2003.06.003`.

Ellingsen T, Johannesson M (2007) Paying Respect. *Journal of Economic Perspectives* 21(4):135–149.

Falk A, Fehr E, Fischbacher U (2008) Testing Theories of Fairness-Intentions Matter. *Games and Economic Behavior* 62(1):287–303.

Falk A, Fischbacher U (2006) A Theory of Reciprocity. *Games and Economic Behavior* 54(2):293–315.

Fehr E, Gachter S, Kirchsteiger G (1997) Reciprocity as a Contract Enforcement Device: Experimental Evidence. *Econometrica* 65(4):833.

Fehr E, Kirchsteiger G, Riedl A (1993) Does Fairness Prevent Market Clearing? An Experimental Investigation. *The Quarterly Journal of Economics* 108(2):437–459.

Fehr E, Rockenbach B (2003) Detrimental Effects of Sanctions on Human Altruism. *Nature* 422(6928):137–140.

Frederiksen A, Lange F, Kriechel B (2017) Subjective performance evaluations and employee careers. *Journal of Economic Behavior and Organization* 134:408–429.

Gill D, Kissova Z, Lee J, Prowse VL (2019) First-Place Loving and Last-Place Loathing: How Rank in the Distribution of Performance Affects Effort Provision. *Management Science* 65(2):494–507.

Holland K (2006) Performance Reviews: Many Need Improvement. URL `http://www.nytimes.com/2006/09/10/business/yourmoney/10mgmt.html`.

Horton JJ (2010) Employer Expectations, Peer Effects and Productivity: Evidence from a Series of Field Experiments. Technical report.

Horton JJ, Rand DG, Zeckhauser RJ (2011) The Online Laboratory: Conducting Experiments in a Real Labor Market. *Experimental Economics* 14(3):399–425.

Ipeirotis PG (2010) Analyzing the Amazon Mechanical Turk Marketplace. *The ACM Magazine for Students* 17(2):16.

Kube S, Maréchal MA, Puppe C (2012) The Currency of Reciprocity - Gift-Exchange in the Workplace. *The American Economic Review* 102(4):1644–1662.

Landy FJ, Farr JL (1980) Performance Rating. *Psychological Bulletin* 87(1):72–107.

Lazear EP, Rosen S (1981) Rank-Order Tournaments as Optimum Labor Contracts. *Journal of Political Economy* 89(5):841–864.

Levine DK (1998) Modeling Altruism and Spitefulness in Experiments. *Review of Economic Dynamics* 1(3):593–622.

Litman L, Robinson J, Abberbock T (2017) TurkPrime.com: A Versatile Crowdsourcing Data Acquisition Platform for the Behavioral Sciences. *Behavior Research Methods* 49(2):433–442.

Ockenfels A, Sliwka D, Werner P (2015) Bonus Payments and Reference Point Violations. *Management Science* 61(7):1496–1513.

Oswald AJ, Proto E, Sgroi D (2015) Happiness and Productivity. *Journal of Labor Economics* 33(4):789–822.

Paolacci G, Chandler J, Ipeirotis P (2010) Running Experiments on Amazon Mechanical Turk. *Judgment and Decision making* 5(5):411–419.

Parducci A (1965) Category Judgment: A Range-Frequency Model. *Psychological Review* 72(6):407–418.

Prendergast CJ (1999) The Provision of Incentives in Firms. *Journal of Economic Literature* 37(1):7–63.

Rabin M (1993) Incorporating fairness into game theory and economics. *The American Economic Review* 83(5):1281–1302.

Robinson J, Rosenzweig C, Moss AJ, Litman L (2019) Tapped out or Barely Tapped? Recommendations for how to Harness the Vast and Largely Unused Potential of the Mechanical Turk Participant Pool. *PLoS ONE* 14(12):1–29.

Sebald A, Walzl M (2014) Subjective performance evaluations and reciprocity in principal-agent relations. *Scandinavian Journal of Economics* 116(2):570–590.

US Census Bureau (2015) 2015 Management and Organizational Practices Survey. URL `https://www.census.gov/data/tables/2015/econ/mops/2015-survey-release.html`.

Vogt T, Sliwka D, Thonemann UW (2021) On Rating Scales in Performance Appraisals - Performance Effects of a Dummy Category. *Working Paper* .