The effects of performance report layout on managers' subjective evaluation judgments

Maas, V.S.; Verdoorn, N.

DOI
10.1080/00014788.2017.1324756

Publication date
2017

Document Version
Final published version

Published in
Accounting and Business Research

Accounting and Business Research, 2017
Vol. 47, No. 7, 731–751, https://doi.org/10.1080/00014788.2017.1324756

The effects of performance report layout on managers' subjective evaluation judgments**

VICTOR S. MAAS [a*] and NIELS VERDOORN[b]

[a]*Amsterdam Business School, University of Amsterdam, Amsterdam, Netherlands;*
[b]*Cure+, Rotterdam, Netherlands*

Managers tend to provide subjective performance evaluations that are relatively high (leniency) and not very dispersed (compression). This paper reports on an experiment that investigates whether the layout of performance reports affects the leniency and compression of managers' subjective evaluations. Relying on psychology theory, we predict that subjective ratings will be higher and more compressed if performance reports contain alphabetically listed indicators rather than categorically listed indicators (as in a balanced scorecard). Moreover, we predict that ratings will be higher and more compressed if performance reports present indicator target and actual values in tables than when this information is presented in graphs. The results from the experiment provide support for the hypothesis that performance ratings are higher if measures are listed in alphabetical order as opposed to presented in a four-category balanced scorecard format. However, there is no support for the other hypotheses. We discuss the implications of the study for accounting research and practice.

Keywords: subjective performance evaluation; rating leniency; rating compression; balanced scorecard; graphs

1. Introduction

Organizations often allow managers some discretion in evaluating the performance of their subordinates. There are different ways in which organizations can introduce discretion in the performance evaluation process (Bol 2008, Höppe and Moers 2011, Van Rinsum and Verbeeten 2012). One common procedure is that managers subjectively determine an overall performance rating based on a report that contains target and actual values for a number of selected performance measures (e.g. Lipe and Salterio 2000, Ittner et al. 2003, Liedtka et al. 2008). The media selection literature (e.g. Daft and Lengel 1986, Russ et al. 1990) posits that written and numeric documents such as these reports have relatively low information carrying capacity, compared to 'richer' media such as face-to-face meetings. Requiring managers to capture their evaluation of a subordinate in one key figure, and to base this figure on a collection of figures in a

*Corresponding author. Email: vmaas@uva.nl
**The data and research instrument from this study are available from the first author upon request.
Paper accepted by David Marginson

formal report, on the one hand reduces the potential for impression management and ambiguity in the evaluation process (Marginson 2006, Brutus 2010). On the other hand, it can also have negative consequences to the extent that less tangible aspects of an employee's performance remain unobserved, or more complex and equivocal issues are not given due weight in the evaluation (Russ et al. 1990, Brutus 2010).

In this study, we investigate whether – and if so, how – the layout of performance reports affects subjective evaluation judgments in settings in which managers determine an overall performance rating based on the signals in these reports. We specifically look at the effects of performance report layout on the leniency and compression of performance ratings. Leniency refers to the level of the performance ratings. Existing research shows that managers exhibit a tendency to provide subordinates with subjective performance ratings that are above the average of the rating scale, and that subjective indicators are often more favorable than objective indicators of subordinate performance (i.e. indicators that cannot be influenced by the evaluator) (e.g. Saal and Landy 1977, Bol 2011, Chen 2014). Compression refers to the cross-sectional distribution of the ratings. Existing research shows that subjective ratings are often more compressed (i.e. exhibit less variance) than objective indicators of subordinate performance (e.g. Saal and Landy 1977, Bol 2011, Chen 2014). Leniency and compression can be costly for organizations, in particular when some managers in the organization are more lenient and more likely to compress ratings than others.

While several accounting studies have found evidence of managers' tendency to provide lenient and compressed subjective performance ratings (e.g. Moers 2005, Bol 2011, Bol et al. 2016), relatively little is known about the factors that influence this rating behavior (Bol 2008, Golman and Bhatia 2012). Specifically, no research that we are aware of has examined whether the level of – and variation in – managers' performance evaluations are affected by the layout of the performance reports on which these managers base their evaluations. However, there is a substantial literature that suggests that the way in which accounting information is presented to decision-makers can alter judgment and decision-making processes and outcomes (e.g. Lipe and Salterio 2002, Cardinaels 2008, Cardinaels and van Veen-Dirks 2010, Tang et al. 2014, Kramer et al. 2016). Relying on this literature, and on insights from judgment and decision-making theory, we argue that the subjective weighting of performance-relevant information is likely influenced by two aspects of the way in which this information is presented. First, we focus on information organization, contrasting performance reports in which performance measures are listed without explicit categorization (but in alphabetical order) with reports in which measures are grouped by category (as in a balanced scorecard) (e.g. Lipe and Salterio 2002, Kaplan et al. 2007, Cheng and Humphreys 2012). Next, we look at presentation format and compare reports in which target and actual values are presented in tables with reports in which these numbers are presented in graphs (bar charts) (e.g. Jarvenpaa and Dickson 1988, Cardinaels 2008).

Our theory development builds on the general idea proposed by Bol (2011) that one important reason why managers' subjective ratings are often inflated and compressed is that managers lack information about employees' actions. To make accurate subjective assessments, managers need to invest time and effort in gathering such information, and it will not always be in their personal interest to do so. Instead, these managers will deliberately limit their information search and will try to keep their employees happy by giving them all relatively high ratings and avoiding ratings that clearly differentiate between stronger and weaker performers (Bol 2011, Bol et al. 2016). We extend this reasoning to settings in which information about employees *is available*, in the form of performance reports, but extracting this information from the reports and analyzing and comparing performance levels requires costly cognitive effort. Next, we argue that the effort required to extract relevant information from performance reports depends on the layout of these performance reports, and we hypothesize that ratings will be less lenient and compressed if managers are provided with performance reports in which measures are explicitly categorized and reports in which

information is presented in graphs, because managers will find it easier to digest information from such reports.

We test our hypotheses using an experiment. Participants are provided with performance data for eight store managers and have to give overall performance ratings based on the provided information. The results indicate limited support for the hypotheses. While there is a significant effect of information organization on rater leniency, such that ratings are higher if measures are listed alphabetically rather than grouped in balanced scorecard categories, we find no effect of presentation format (tables versus graphs) on leniency. Also, we do not find any effects of information organization or presentation format on evaluation compression.

Our paper contributes to several streams of recent accounting literature. First, a few recent studies have explored why some managers provide more lenient and compressed ratings than others (Bol 2011, Chen 2014, Bol et al. 2016). Our study contributes to this emerging literature by providing insights on how the layout of performance reports can affect leniency and compression. Next, there is a much larger stream of literature that examines how managers subjectively weigh alternative performance measures in providing overall performance judgments. This literature has found that subjective weighting processes are affected by intentional and unintentional biases. For example, studies have found evidence of favoritism (Ittner et al. 2003), common measure bias (Lipe and Salterio 2000, Libby et al. 2004), financial measure bias (Cardinaels and van Veen-Dirks 2010, Johnson et al. 2014), outcome effects (Long et al. 2015) and mood bias (Ding and Beaulieu 2011). However, with the exception of Bol et al. (2016), no other study has looked at the leniency or compression of overall ratings that are the outcome of subjective weighting processes.[1] Our study also contributes to the subjective performance evaluation literature by examining a setting in which managers need to rate the performance of eight managers, whereas the existing experimental literature has almost exclusively focused on settings in which evaluators need to evaluate the performance of one or two managers. In organizations, most supervisors will have more than two subordinates and they will usually have a limited amount of time available to compare and evaluate these subordinates' performance. Our experimental design captures this important element of actual evaluation settings.

Finally, our study contributes to the small but growing literature on the presentation and visualization of management accounting information. While there is a considerable body of research examining the use of graphical information in annual reports (Beattie and Jones 2008, Davison 2015), there are almost no studies that have examined how the 'look and feel' of managerial accounting reports impact decision-making within organizations (Kramer et al. 2016). We extend this literature by investigating how the layout of performance reports affects evaluation judgments.

The remainder of the paper is organized as follows. In the next section, we review the relevant prior literature and we develop our hypotheses. We then explain our research method and present the results from the experiment. In the final section we discuss these results and their implications, also paying attention to the limitations of our study and possible avenues for future research.

2. Hypotheses development

2.1. *Performance evaluation leniency and compression*

Managerial discretion in performance evaluation procedures generally results in performance ratings that are higher and less dispersed than ratings that are exclusively based on objective measures of subordinate performance, and cannot be influenced by the evaluator (Bol 2008, 2011). To understand managers' subjective rating behavior, it is important to consider their incentives in the evaluation process (Bol et al. 2016). Maas et al. (2012) and Bol et al. (2016) argue that it is not necessarily in managers' interest to provide ratings that are as accurate as possible. Higher

levels of accuracy in evaluations require managers to spend more time and effort collecting and analyzing data that are informative about subordinates' action choices (Bol 2011, Maas et al. 2012). Because there are opportunity costs associated with spending time and effort on performance assessments, it will not always be in managers' personal interest to provide accurate ratings. Instead, their incentives may lead managers to provide ratings that require little extra work and are acceptable to all evaluated employees and relatively easy to explain to both subordinates and superiors. Ratings that are relatively high and not too dispersed often satisfy these criteria (Bol et al. 2016).

Consistent with this reasoning, Bol (2011) argues and finds that managers who have less information about their employees tend to give higher and more compressed ratings. However, even when relatively much data about employee behavior are available to managers, for example, in the form of performance reports, identifying, analyzing and comparing such data will require time and effort. Moreover, the cognitive resources required to process the data may depend on the how the data are presented, i.e. on the layout of the performance reports. The more cognitive effort managers need to put into the process of translating raw performance data into balanced evaluation judgments, the more likely it is that they instead will settle for an 'easy outcome' and provide all employees with a relatively favorable rating. Thus, we argue that by presenting subordinate performance data in ways that allow managers to quickly observe and compare the performance levels of different individuals, organizations can improve the accuracy of subjective performance evaluations.

Leniency and compression of performance ratings can be costly for organizations for a number of reasons. An obvious cost of leniency is that organizations may end up spending too much on employee compensation, for example, if ratings are tied to variable pay levels according to some pre-determined formula. Also, leniency in ratings results in weakened incentives for employees. Low performers' will feel less pressure to improve because performance increases are not clearly reflected in better ratings. Similarly, high performers will have less incentives to keep their effort at a high level because they will get a favorable rating even for lower levels of performance (Bol 2011, Bol et al. 2016).[2] Rating compression is costly because it makes it more difficult to identify exceptionally talented or motivated individuals and to differentiate between employees whose work barely meets the standard and those whose work is clearly above the minimum level of acceptability. This, in turn, likely leads to lower quality promotion decisions, replacement decisions and resource allocation decisions (Bol 2011). Also, compression of ratings and rewards may lead organizations toward a 'culture of mediocrity', as stronger performers will look for alternative employment options which better recognize their superior efforts or skill levels (Lazear 2000, Bol et al. 2016). Finally, if some managers in the organization are more likely to give lenient and compressed evaluations than others, then this is another source of costs for firms. Differences in leniency and compression levels make evaluation judgments of different managers incomparable. A failure to correct evaluation judgments for the rating style of individual managers, for example, using calibration meetings (Demere et al. 2016), can easily result in flawed business decisions and feelings of unfairness and frustration (Murphy and Cleveland 1995, Ittner et al. 2003).

The existing literature points toward two possible strategies to make evaluation tasks easier and reduce the resources that managers need to spend on translating basic accounting data into accurate evaluation judgments. First, information items in performance reports can be categorized in meaningful ways, which facilitates pattern recognition and reduces cognitive processing requirements. Next, information can be presented in graphical formats, which makes it easier to recognize extreme values and to compare values across different items or, in the case of subordinate evaluation, across individuals.

2.2. *Information organization*

Some organizations use performance reports in which measures are grouped based on the specific type of performance about which they are informative. One example of such a report is the balanced scorecard, which groups performance measures into four categories: financial performance, customer-related performance, learning and growth performance and internal business processes performance (Kaplan and Norton 1996). The measures in these categories should be linked through cause-and-effect relationships such that higher performance on 'leading' measures in the current period leads to higher performance on 'lagging' measures in future periods (Kaplan and Norton 1996, Lowe et al. 2011).

Building on theory and research in psychology and marketing (e.g. Bettman and Kakkar 1977, Payne et al. 1993), Lipe and Salterio (2002) argued that using a balanced scorecard format rather than a performance report in which measures are not categorized reduces the cognitive load of the evaluation task, which in turn increases the accuracy of performance evaluation judgments. The four balanced scorecard categories 'suggest a way for managers to mentally organize the large number of performance measures' (Lipe and Salterio 2002, p. 531), making the evaluation task easier. Specifically, grouping the measures in meaningful categories allows evaluators to use a 'divide and conquer' strategy (Shanteau 1988). For example, if a report contains eight measures from four categories, evaluators can first assess the overall performance of each subordinate in each category and only then evaluate and compare performance levels across individuals. In support of this reasoning, Lipe and Salterio (2002) found that, dependent on the pattern of performance results, organizing measures in a balanced scorecard affected evaluation judgments. If performance differences between two subordinate managers were located within a single category, managers using a balanced scorecard adjusted their ratings more than managers using an unformatted scorecard (Lipe and Salterio 2002).

Similarly, Kaplan et al. (2007) argued that managers are likely to use heuristics in evaluating subordinates' performance. They hypothesized that managers' evaluation judgments will be less strongly affected by (irrelevant) subordinate likability if performance information is provided in a balanced scorecard format instead of an unstructured format. However, their experiment failed to support this hypothesis and showed that 'likability bias' was not mitigated by information organization. Finally, Cardinaels and van Veen-Dirks (2010) studied the impact of information organization on the weights attached to specific measures. Similar to Lipe and Salterio (2002) and Kaplan et al. (2007), they examined a setting in which managers had to evaluate two divisions with diverging levels of performance. The results of Cardinaels and van Veen-Dirks (2010) indicate that if the performance difference is located in the 'financial' category, managers using a balanced scorecard place more weight on financial measures than managers using an unstructured scorecard. However, when the performance difference is located in one of the non-financial leading measure categories, there is no difference between the evaluations of managers using balanced scorecards and managers using unformatted scorecards.

In summary, the existing literature generally suggests that the organization of information in performance reports can affect evaluation judgments through the demands that are placed on evaluators' cognitive processing. Following this reasoning, we hypothesize that information organization will also influence the extent to which evaluators who need to simultaneously evaluate several subordinates will exhibit a tendency to provide ratings that are relatively lenient and compressed. Grouping measures into meaningful categories, such as the four balanced scorecard categories, will allow evaluators to use a 'divide and conquer' strategy when processing the performance information (Shanteau 1988, Lipe and Salterio 2002). Such a strategy reduces the costly cognitive effort that is required to arrive at a set of evaluation judgments which are perceived to accurately reflect the differences in subordinates' performance levels. Thus, because

cognitive information acquisition costs are lower if information is structured in a meaningful way, leniency and compression will be lower if performance measures in performance reports are organized in a balanced scorecard format than in a format without explicit categories (e.g. in alphabetical order). This is summarized in the two following hypotheses:

H1a: Overall performance ratings are more lenient when performance measures are organized alphabetically than when performance measures are organized categorically.

H1b: Overall performance ratings are more compressed when performance measures are organized alphabetically than when performance measures are organized categorically.

2.3. *Presentation format*

Another important aspect of the layout of performance reports that potentially affects subjective weighting processes is the presentation format of the information contained in the report, specifically whether the information is presented in tables or in graphs. While research on the use of graphs to present financial information goes back to the 1970s (e.g. Moriarty 1979), there is very little research that has examined the effects of presentation format on judgment and decision-making in a management accounting context. One exception is Cardinaels (2008), who found that individuals with relatively low accounting knowledge performed better on a complex decision-making task when relevant information was provided in graphs instead of tables, whereas more knowledgeable individuals performed better with information provided in tables than with graphs.

The findings of Cardinaels (2008) suggest that the effects of differences in presentation format (tables versus graphs) are context dependent. This conclusion is in line with cognitive fit theory (Vessey 1991), which is the dominant theoretical perspective in the psychology and marketing literatures on information presentation formats. Cognitive fit theory poses that graphs provide spatial representations of information which '[…] facilitate viewing the information contained therein at a glance without addressing the elements separately or analytically' (Vessey 1991, p. 225). Tables on the other hand, emphasize symbolic information and primarily facilitate extracting specific data values (Larkin and Simon 1987, Vessey 1991). Cognitive fit theory not only distinguishes between spatial and symbolic information representations but also between spatial and symbolic tasks. Its basic tenet is that spatial tasks require spatial information and symbolic tasks require symbolic information and that a fit between the characteristics of the task and the characteristics of the presentation format increases task performance. Tasks that are better performed with data presented in tables are, for example, tasks that require decision-makers to extract one specific piece of information from a larger set (e.g. financial statements). Tasks that benefit from information being presented in graphs are, for example, tasks that require decision-makers to integrate information, to make comparisons or to identify relationships or trends over time (Vessey 1991, Cardinaels 2008, Kelton et al. 2010).

We argue that evaluating the performance of a number of subordinates based on information contained in performance reports is a task that will be easier to achieve when data are presented in graphs than when data are presented in tables. Evaluators need to compare target and actual levels of performance for different measures and different individuals. Graphs allow evaluators to use their dominant sense (sight) and their spatial intelligence to decode the information in the reports relatively quickly and to arrive at an overall conclusion about an individuals' absolute and relative performance with relatively little cognitive effort. If there is limited time available for the evaluation process, managers relying on data from graphs are therefore likely to develop a more accurate picture of the overall performance of their subordinates than managers who have to extract the information from tables. For this reason, managers provided with graphs will less frequently resort to inflating and compressing ratings. The results of the only empirical

study that we are aware of that has examined the use of tables versus graphs in a performance evaluation setting (Johnston and Shields 1983) are consistent with this reasoning. Johnston and Shields (1983) examined the influence of the presentation format of six months of sales data on the performance appraisals of salespersons. While they did not find a consistent difference between the appraisals of managers who were evaluated based on data in tables and graphs, they did find that if there was much variance across the six months, evaluators provided lower ratings if the information was presented in graphs. This result suggests that evaluators using graphs found it easier to recognize patterns in the time series data, and were more likely to base their evaluation on a comparison of data points, than evaluators who had no access to a graph (Johnston and Shields 1983, p. 198). This is consistent with our reasoning that providing evaluators with graphs will increase the probability that their evaluations will be based on cross-sectional comparisons, which will in turn reduce their tendency to stick to a narrow range of relatively favorable ratings.

The two hypotheses below summarize our predictions:

H2a: Overall performance ratings are more lenient when performance reports contain tables than when performance reports contain graphs.

H2b: Overall performance ratings are more compressed when performance reports contain tables than when performance reports contain graphs.

3. Experimental design and method

3.1. *Experimental design*

We use an experiment with a 2×2 design to test our hypotheses. Information organization and presentation format were manipulated orthogonally between subjects. To manipulate information organization, we provide one group of participants with performance reports in which measures are organized by the four balanced scorecard categories whereas the other participants receive reports in which the same performance measures listed in alphabetical order. To manipulate presentation format, we present the data in the performance reports in either tables or graphs (red and blue bar charts). The observed dependent variables are the leniency and the compression of eight subjective overall performance ratings based on performance measures contained in performance reports.[3]

3.2. *Participants*

The participants are 183 undergraduate business and economics students from a large Western-European university. All students had taken several accounting courses including at least one course in management accounting. The participants' age varies between 18 and 38 years old, with a mean (standard deviation) of 21.13 (2.53). The majority of the participants (61.7%) is male. Only a small minority of 17 (9.3%) participants reports having experience in evaluating employees in real world settings.

3.3. *Task and manipulations*

In the experiment, participants assumed the role of a regional manager of a retail company. Their task was to evaluate the performance of eight store managers in their region.[4] For each store manager, they received a performance report with target and actual values for eight performance measures. The reports also contained the names of the store managers to increase experimental realism. The performance measures were the same for all eight store managers and were described as important predictors of the future results of the company. Participants were asked to carefully

read the performance reports and to rate the overall performance of each store manager on a scale of 0–100. They were asked to assume that their ratings would be used by the firm's 'evaluation committee' to make promotion decisions.

The performance reports of the eight store managers are based on the scorecards in Cardinaels and van Veen-Dirks (2010). Specifically, we selected 8 of the 16 measures from the scorecard in Cardinaels and van Veen-Dirks (2010), two from each balanced scorecard category.[5] Next, using the target values of Cardinaels and van Veen-Dirks (2010) as a starting point, we systematically varied the target and actual levels for each measure to create eight performance reports signaling different levels of overall performance. Target levels varied between −20% and +20% of the targets in Cardinaels and van Veen-Dirks (2010).[6] Actual performance levels varied between −15% and +15% of target. The figures in the performance reports were selected such that the average deviation from target varied across the eight stores, with the best performing store outperforming its targets by on average 7.5%, and the worst performing store's actual performance being on average 7.5% below target. All participants saw the stores in the same random order.[7] Table 1 presents the data from the performance reports of the eight stores.

The independent variables were manipulated by varying the design of the performance reports. First, we manipulated information organization by changing the order in which the eight performance measures are presented in the reports. In the alphabetical organization conditions, the measures were presented in alphabetical order with 'Customer satisfaction rating' on top and 'Sales Growth' at the bottom. In the categorical organization conditions, each measure was labeled as belonging to one of the four balanced scorecard categories (financial performance, customer performance, internal business processes performance, and learning and growth performance). They were presented in the 'traditional' balanced scorecard order of a for-profit firm, with financial performance on top, followed by customer, internal business processes and learning and growth performance (Kaplan and Norton 1996).

Presentation format was manipulated by either presenting the data in a table or in a graph. While the information provided was the same in the table and the graph conditions, participants in the graph conditions also saw a graphical representation of the target and actual levels for each performance measure in the form of a bar chart. Importantly, there were separate charts for each of the eight performance reports and no overall chart allowing direct comparison of the performance of the store managers. We present the charts in color, with target values in blue and actual values in red. Examples of performance reports in each of the four conditions are included in the appendix. (Please refer to the online version for color reference)

3.4. *Experimental procedures*

The students participated during class hours of an intermediate accounting course that contained both financial accounting and cost accounting elements. Neither of the authors was involved in the teaching of this course. The instrument was distributed by one of the authors in 11 tutorial groups of about 15–20 students. Students were asked whether they wanted to participate in a research study. It was emphasized that participation was voluntary and would take about 15 minutes. Except for one, all students agreed to participate. Participants were asked to work individually, not to talk with each other, and to go through the materials in the specified order. One of the authors was present from the beginning to the end of each of the 11 sessions to validate that the participants followed these basic rules. The participants did not receive any form of payment, but they could indicate that they were interested in receiving the results of the study.

Within each session, instruments were distributed randomly, with one important exception. Because we could not rule out that participants might be able to catch a view of each other's research instrument, we decided to always distribute either only instruments with performance

Table 1. Performance reports for the eight stores in the experimental task.

	Store 1			Store 2			Store 3			Store 4		
	Target	Actual	Δ	Target	Actual	Δ	Target	Actual	Δ	Target	Actual	Δ
Customer satisfaction rating (%)	86%	94.6%	+10%	86.0%	86.0%	0	75.0%	82.5%	+10%	94.0%	84.6%	−10%
Employee satisfaction rating (%)	79%	90.9%	+15%	79.0%	86.9%	+10%	82.0%	82.0%	0	88.0%	88.0%	0
Employee training (# hours)	14.0	14.6	+5%	15.5	16.3	+5%	14.4	15.8	+10%	14.0	14.0	0
Orders per week (# orders)	2910	2910	0	3060	3213	+5%	2970	2822	−5%	2430	2552	+5%
Repeat sales (%)	24.0%	25.2%	+5%	23.0%	26.5%	+15%	24.0%	25.2%	+5%	26.0%	24.7%	−5%
Return on sales (%)	65.0%	65.0%	0	55.0%	52.3%	−5%	66.0%	62.7%	−5%	59.0%	64.9%	+10%
Returned products (%)	6.0%	5.4%	+10%	7.0%	6.3%	+10%	7.0%	7.0%	0	5.0%	5.3%	−5%
Sales growth (%)	16.0%	18.4%	+15%	10.0%	10.0%	0	12.0%	12.6%	+5%	23.0%	24.2%	+5%
Average Δ			+7.5%			+5.0%			+2.5%			0
	Store 5			Store 6			Store 7			Store 8		
	Target	Actual	Δ	Target	Actual	Δ	Target	Actual	Δ	Target	Actual	Δ
Customer satisfaction rating (%)	81.0%	85.1%	+5%	86.0%	86.0%	0	81.0%	81.0%	0	95.0%	95.0%	0
Employee satisfaction rating (%)	89.0%	93.5%	+5%	85.0%	85.0%	0	71.0%	63.9%	−10%	82.0%	73.8%	−10%
Employee training (# hours)	15.8	15.8	0	14.0	13.3	−5%	15.6	14.0	−10%	14.3	14.3	0
Orders per week (# orders)	3240	3240	0	3570	3213	−10%	2700	2700	0	2940	2499	−15%
Repeat sales (%)	32.0%	35.2%	+10%	40.0%	42.0%	+5%	32.0%	30.4%	−5%	32.0%	27.2%	−15%
Return on sales (%)	67.0%	60.3%	−10%	59.0%	56.1%	−5%	70.0%	73.5%	+5%	67.0%	63.7%	−5%
Returned products (%)	7.0%	7.4%	−5%	7.0%	7.7%	−10%	5.0%	5.8%	−15%	7.0%	7.4%	−5%
Sales growth (%)	21.0%	20.0%	−5%	10.0%	10.5%	+5%	25.0%	23.8%	−5%	12.0%	10.8%	−10%
Average Δ			0			−2.5%			−5.0%			−7.5%

Notes: This table provides target values and actual values for the eight performance measures for the eight stores, as they were presented to the participants. The Δ columns indicate the difference between the actual and target values, with plus signs representing favorable differences and minus signs representing unfavorable differences. Note that experimental participants did not see these columns. The store numbers represent the stores' rank order based on the average Δ, not the order in which the participants encountered the stores in the experimental materials.

data presented in tables or only instruments with data presented in graphs in a session (whether a specific session was assigned to the table or the graph conditions was determined randomly). The reason is that we wanted to prevent any construct validity threats caused by carry over effects that might emerge if participants realized they were in an experiment with different conditions. We particularly wanted to mitigate the threat of confusion or resentful demoralization by participants in a table condition who accidentally observed that other participants received a full color graph. After the participants finished the experimental task, they filled out an exit questionnaire. This questionnaire contained manipulation checks, questions about demographics, and a number of items related to the participants' perception of the task and their evaluation judgments.

4. Results

4.1. *Preliminary analysis and measures*

To check whether the participants in the categorical information organization conditions were more aware of the different categories of performance measures than the participants in the alphabetical information organization conditions, we asked participants to indicate on a five-point Likert scale with anchors *Strongly disagree* (1) and *Strongly agree* (5) to what extent they agreed with the statement: 'I recognized four different categories of performance indicators.' The results confirm that participants in the categorical information organization conditions agreed stronger with this statement than participants in the alphabetical information organization conditions (mean = 3.92, std. dev. = 1.035 versus mean = 3.26, std. dev. = 1.157; $t = 4.082$, $p < .001$).[8]

Our analysis focuses on the leniency and the compression in the performance ratings of the eight store managers. As indicators of leniency, we focus on the average performance rating (*AVERAGE RATING*) and the median performance rating (*MEDIAN RATING*) provided to the eight store managers. Regarding compression, we look at three alternative measures. First, we measure the standard deviation of the eight ratings (*SD RATINGS*). While the standard deviation is a good indicator of the overall variance in an evaluator's ratings, it does not directly allow us to assess the extent to which raters are willing to clearly differentiate between top performers and weak performers. For this reason, we also look at the absolute difference between a rater's highest and lowest rating (*ABSOLUTE DIFFERENCE*) and the relative difference between a rater's highest and lowest rating, focusing on the lowest rating expressed as a percentage of the highest rating (*RELATIVE DIFFERENCE*).[9] Table 2 contains descriptive statistics for each of the dependent variables on a cell-by-cell basis.

4.2. *Hypothesis tests*

To test our hypotheses, we use factorial ANCOVA with the two manipulations as factors, and gender and evaluation experience as a covariates. We include gender as a covariate in our analyses because research suggests that women tend to rate differently than men (Furnham and Stringfield 2001). Evaluation experience is included as a covariate to control for possible effects of participants' experience in subjectively evaluating employees on their ratings. The first two ANCOVA models test our hypotheses related to performance evaluation leniency, and focus on the variance in the variables *AVERAGE RATING* and *MEDIAN RATING*. The results of the analyses are presented in Table 3, panels A and B. These panels show that in both models the interaction effect of information organization and presentation format is insignificant ($F_{1,177} = 0.070$, $p = .792$ in the *AVERAGE RATING* model and $F_{1,177} = 0.124$, $p = .725$ in the *MEDIAN RATING* model). This comes as no surprise given that our theory did not predict an interaction effect. Next, in both models the main effect of information organization is significant ($F_{1,177} = 13.613$, $p < .001$ in

Table 2. Descriptive statistics by treatment.

	Presentation format								
	Tables			Graphs			Overall		
	N	Mean	Std. Dev.	*N*	Mean	Std. Dev.	*N*	Mean	Std. Dev.
Information organization: Alphabetical									
AVERAGE RATING	49	75.55	6.71	43	76.74	8.56	92	76.11	7.61
MEDIAN RATING	49	75.91	7.47	43	76.91	9.07	92	76.38	8.22
SD RATINGS	49	9.96	4.42	43	9.57	4.39	92	9.78	4.39
ABSOLUTE DIFFERENCE	49	28.86	12.65	43	28.28	12.90	92	28.59	12.70
RELATIVE DIFFERENCE	49	68.25	12.71	43	68.49	14.62	92	68.36	13.56
Information Organization: Categorical									
AVERAGE RATING	45	71.49	8.78	46	71.89	8.92	91	71.70	8.80
MEDIAN RATING	45	72.13	9.00	46	72.06	9.86	91	72.09	9.39
SD RATINGS	45	10.80	5.58	46	9.22	4.61	91	10.01	5.15
ABSOLUTE DIFFERENCE	45	32.67	16.13	46	27.83	13.73	91	30.22	15.08
RELATIVE DIFFERENCE	45	62.77	17.04	46	67.35	15.47	91	65.09	16.34
Overall									
AVERAGE RATING	94	73.61	7.99	89	74.24	9.03	183	73.91	8.50
MEDIAN RATING	94	74.40	9.74	89	74.10	8.41	183	74.25	9.06
SD RATINGS	94	10.37	5.00	89	9.39	4.49	183	9.89	4.77
ABSOLUTE DIFFERENCE	94	30.68	14.47	89	28.04	13.26	183	29.40	13.92
RELATIVE DIFFERENCE	94	65.63	15.11	89	67.90	14.99	183	66.73	15.05

Notes: Participants rate the performance of eight store managers on a 0–100 scale. Presentation format was manipulated between-subjects. In the Tables conditions participants received the performance reports in a table format. In the Graphs conditions participants received the performance reports in the form of bar charts. Information Organization was also manipulated between-subjects. In the Alphabetical Organization conditions participants received performance reports in which the measures were presented in alphabetical order. In the Categorical Organization conditions participants received performance reports in which the measures were presented in the four balanced scorecard categories (first financial performance measures, next customer-related performance measures, next measures related to internal business processes and finally learning and growth-related measures). AVERAGE RATING is the average of the eight performance ratings provided by the participant. MEDIAN RATING is the median of the eight performance ratings provided by the participant. SD RATINGS is the standard deviation of the eight performance ratings provided by the participant. ABSOLUTE DIFFERENCE is the difference between the highest and the lowest of the eight ratings provided by the participant. RELATIVE DIFFERENCE is the relative difference between the highest and the lowest of the eight ratings provided by the participant and is calculated as (lowest rating/highest rating) × 100.

the *AVERAGE RATING* model and $F_{1,177} = 10.985$, $p = .001$ in the *MEDIAN RATING* model), whereas the main effect of presentation format is not significant ($F_{1,177} = 0.488$, $p = .486$ in the *AVERAGE RATING* model and $F_{1,177} = 0.148$, $p = .701$ in the *MEDIAN RATING* model). The significant main effects of information organization indicate that participants in the alphabetical organization conditions provided significantly higher average ratings and median ratings than participants in the categorical organization conditions. The mean (standard deviation) of *AVERAGE RATING* in the alphabetical organization conditions is 76.11 (7.61), whereas it is 71.70 (8.80) in the categorical organization conditions. The mean (standard deviation) of *MEDIAN RATING* in the alphabetical organization conditions is 76.38 (8.22), whereas it is 72.09 (9.39) in the categorical organization conditions. Together, these results provide support for H1a but not for H2a. Notably, the ANCOVA results also show significant effects of gender in both models ($F_{1,177} = 7.290$, $p = .008$ in the AVERAGE RATING model and $F_{1,177} = 6.067$, $p = .015$ in the *MEDIAN RATING* model), indicating that female participants' average and median ratings were more lenient than male participants' ratings, which is consistent with existing research (Furnham and Stringfield 2001).

Table 3. Analysis of covariance results.

Source	Sum of squares	df	Mean square	*F*	Sig.
Panel A: Dependent variable is AVERAGE RATING					
Model	1446.938	5	289.388	4.382	.001
Gender	481.446	1	481.446	7.290	.008
Evaluation experience	30.542	1	30.542	0.462	.497
Presentation format	32.247	1	32.247	0.488	.486
Information organization	898.974	1	898.974	13.613	<.001
Presentation format × Information organization	4.606	1	4.606	0.070	.792
Error	11,689.019	177	66.040		
Panel B: Dependent variable is MEDIAN RATING					
Model	1370.393	5	274.079	3.577	.004
Gender	464.871	1	464.871	6.067	.015
Evaluation experience	35.943	1	35.943	0.469	.494
Presentation format	11.360	1	11.360	0.148	.701
Information organization	841.669	1	841.669	10.985	.001
Presentation format × Information organization	9.527	1	9.527	0.124	.725
Error	13,562.041	177	76.622		
Panel C: Dependent variable is SD RATINGS					
Model	98.067	5	19.613	0.859	.510
Gender	26.587	1	26.587	1.164	.282
Evaluation experience	9.629	1	9.629	0.422	.517
Presentation format	47.893	1	47.893	2.097	.149
Information organization	3.327	1	3.327	0.146	.703
Presentation format × Information organization	14.626	1	14.626	0.640	.425
Error	4042.035	177	22.836		
Panel D: Dependent variable is ABSOLUTE DIFFERENCE					
Model	1161.504	5	232.301	1.206	.308
Gender	466.402	1	466.402	2.421	.121
Evaluation experience	39.981	1	39.981	0.208	.649
Presentation format	363.016	1	363.016	1.885	.172
Information organization	136.239	1	136.239	0.707	.401
Presentation format × Information organization	199.635	1	199.635	1.036	.310
Error	34,092.375	177	192.612		
Panel E: Dependent variable is RELATIVE DIFFERENCE					
Model	1903.291	5	380.658	1.712	.134
Gender	913.644	1	913.644	4.110	.044
Evaluation experience	28.436	1	28.436	0.128	.721
Presentation format	295.660	1	295.660	1.330	.250
Information organization	513.966	1	513.966	2.312	.130
Presentation format × Information organization	211.935	1	211.935	0.953	.330
Error	39,346.974	177	222.299		

Notes: This table presents results of ANCOVAs for the five dependent variables described in Table 2. Presentation format and Information organization are between-subjects manipulations that are also described in Table 2. Gender is a dummy that takes on the value 1 if the participant is female and zero otherwise. Evaluation experience is participants' self-reported experience in evaluating subordinates (in years).

Next, we analyze the effects of the manipulations on rating compression, looking, in turn, at three alternative measures of this dependent variable. We again use a factorial ANCOVA with gender as a covariate. First, we assess the effects of information organization and presentation format on the standard deviation of the eight performance ratings provided by the participants (*SD RATINGS*). The results are in Panel B of Table 3. There are no significant main effects and neither is there a significant interaction effect (all $p > .05$). Also, gender has no significant effect on *SD RATINGS* ($p = .282$). If we run the same analysis with *ABSOLUTE DIFFERENCE*

(the difference between the highest rating and the lowest rating) as the dependent variable, we find similar results. As is clear from Panel C of Table 3, also with this dependent variable, there are no significant main effects, nor is there a significant interaction (all $p > .05$). Finally, looking at *RELATIVE DIFFERENCE,* we find a significant effect of gender ($F_{1,\ 177} = 4.110$, $p = .044$), but we do not find main effects or an interaction effect of our manipulations (all $p > .05$). Together, these results suggest that information organization and presentation format do not affect rating compression. Thus, there is no support for H1b and H2b.

4.3. *Supplemental analyses*

The exit questionnaire contained several items that allow us to gain a better understanding of the participants' thought processes and evaluation judgments. First, the questionnaire contained two items that are informative about participants' perception of the quality of their ratings. These items ask participants to indicate on a five-point Likert scale with anchors *Strongly disagree* (1) and *Strongly agree* (5) to what extent they agree that they are confident about their ratings and that they believe their ratings are accurate. The scores on these items do not vary significantly across conditions ($F_{3,179} = 0.200$, $p = .896$, respectively; $F_{3,\ 179} = 0.720$, $p = .541$). However, we do find significant gender effects. Men tend to agree more that they are confident about their ratings than women (mean = 3.42, std. dev. = 0.811 versus mean = 2.87, std. dev. = 0.947; $t = 4.205$, $p < .001$) and men also agree more that they believe their ratings are accurate than women (mean = 3.27, std. dev. = 0.835 versus mean = 2.96, std. dev. = 0.859; $t = 2.402$, $p = .017$). Neither participants' confidence about their ratings nor the extent to which they believe their ratings are accurate is correlated with leniency. However, belief that ratings are accurate is significantly correlated with *SD RATINGS* ($r = 0.173$, $p = .019$), *ABSOLUTE DIFFERENCE* ($r = 0.163$, $p = .027$) and *RELATIVE DIFFERENCE* ($r = -0.165$, $p = .025$). Participants who compressed ratings to a lesser extent, had a stronger belief in the accuracy of their ratings.

Finally, the questionnaire contains two items that give some insight into the mediating role of perceptions of performance report layout. First, we asked participants to indicate on a five-point Likert scale with anchors *Strongly disagree* (1) and *Strongly agree* (5) to what extent they agreed that the information on which they had to rely in performing the evaluation task was presented clearly. Our reasoning suggests that information in graphs might be perceived as clearer than information presented in tables. While the mean score on this item is higher in the graph conditions (mean = 3.81, std. dev. = 0.954) than in the table conditions (mean = 3.61, std. dev. = 0.944), this difference is not significant ($t = 1.197$, one-tailed $p = .117$). Our reasoning also suggests that managers' tendency to inflate and compress performance ratings increases with the cognitive effort that is required to extract relevant information from performance reports. Stronger agreement with the statement that information was clearly presented might therefore be associated with less leniency and compression. However, evaluation of the correlation coefficients shows that the performance ratings of participants who report stronger agreement with this statement were neither higher nor more compressed (all $p > .05$).

Next, we analyze participants' responses to an item that asked them to indicate (also on a five-point Likert scale with anchors *Strongly disagree* (1) and *Strongly agree* (5)) to what extent they agreed that the order of the performance indicators made the task easier. The data show that there is no significant difference between the average levels of agreement of participants in the categorical organization conditions (mean = 3.13, std. dev. = 0.991) and the alphabetical organization conditions (mean = 3.29, std. dev. = 1.073; $t = 0.702$, one-tailed $p = .242$). At first glance, this seems to contradict our reasoning that it requires less cognitive effort to extract relevant information from performance reports that are organized categorically than from reports in which measures are ordered alphabetically. However, we believe this is not necessarily the case.

While we can only speculate about the reasons for this result, one possibility is that participants in the alphabetical organization conditions *made the task easier for themselves* by giving relatively lenient and compressed ratings. This is consistent with our finding that there are significant correlations between participants' level of agreement with the statement that the order of the performance indicators made the task easier and the leniency and compression of their evaluation judgments. Participants who agree more strongly, score higher on *AVERAGE RATING* ($r = 0.179$, $p = .015$), *MEDIAN RATING* ($r = 0.156$, $p = .034$) and *RELATIVE DIFFERENCE* ($r = 0.183$, $p = .013$) and lower on *SD RATINGS* ($r = -0.172$, $p = .020$) and *ABSOLUTE DIFFERENCE* ($r = -0.168$, $p = .023$). We also note, however, that other explanations are possible and that findings from post-experimental questionnaires should be interpreted with much care as psychology research consistently shows that that it is difficult for decision-makers to critically evaluate the differential role of various aspects of information organization and presentation in their (partly unconscious) judgment and decision-making processes (DeNisi et al. 1984, Spencer et al. 2005).

5. Discussion and conclusion

In this paper, we set out to investigate how – if at all – the layout of performance reports affects the leniency and compression of overall performance judgments based on the subjective weighting of several performance measures. Building on theory that suggests that managers give relative high and compressed ratings because providing accurate ratings requires them to exert costly cognitive effort, we predicted that two aspects of performance report layout would affect the average level of – and the variance in – overall performance ratings. First, we examined information organization, contrasting performance reports in which measures are listed alphabetically and reports in which measures are organized by the four balanced scorecard categories. Next, we examined presentation format, focusing on the different effects of reports in which target and actual values of performance measures are presented in tables and reports in which these are presented in graphs. An experiment with 183 participants showed that while raters who were provided with reports in a balanced scorecard format were less lenient, as predicted, there were no effects of information organization on rating compression. Neither were there any effects of presentation format on the level or compression of the ratings.

The finding that evaluators provided higher ratings if performance measures were not grouped in meaningful categories is important for both management accounting theory and practice. It is important for theory because it adds credibility to the basic idea that contemporary performance measurement systems such as the balanced scorecard are effective because they are particularly well adapted to the inherent limitations of human information processing (Lipe and Salterio 2002, Salterio 2012). As Salterio (2012) notes, the organizational features of the scorecard fit very well with recommendations about information provision based on classic psychology research on information chunking (the 'divide and conquer strategy') (Chase and Simon 1973, Shanteau 1988) and working memory limitations (Miller 1956, Craik and Lockhart 1972). The finding is also important for management accountants, controllers and consultants who are involved in the design of performance measurement systems in practice. It is easy to overlook the consequences of issues that might appear trivial at first sight (Thaler and Sunstein 2008), such as the order in which performance measures are presented. Our study shows that the organization of measures in a balanced scorecard format has direct observable effects on evaluators' overall performance ratings. One specific implication of our study for practice is that organizations might be able to reduce the costs of lenient performance ratings by redesigning their performance reports. Another implication is that the accuracy and fairness of performance ratings can be improved by using a uniform design of performance reports throughout the organization. If some managers receive performance data in a balanced scorecard format whereas other receive the same type

of data in a less structured format, it will be difficult to compare managers' ratings and to draw correct inferences about, for example, which subordinates should be considered for promotion or for a pay raise. Also, if subordinates in business units that have implemented a balanced scorecard score consistently lower in the annual performance review process than subordinates in other units, this may well result in feelings of unfairness, frustration and anger (Ittner et al. 2003).

While the observed effect of information organization on leniency is consistent with our theory, it is important to acknowledge that we cannot rule out that mechanisms different from the one we predicted, contributed to this observed effect. For example, information organization may affect evaluation judgments by directing evaluators' attention in certain directions (Birnberg and Shields 1984, Kramer and Maas 2016), or by changing evaluators' moods or motivation (Ding and Beaulieu 2011, Schwarz 2011). To illustrate, research on processing fluency suggests that individuals like information that feels easy to cognitively process, and tend to respond to processing fluency with more favorable judgments (e.g. Alter and Oppenheimer 2009, Rennekamp 2012). Verifying that the effect that we observed in our experiment is due to the cause–effect mechanism that we predicted, is not straightforward because this mechanism consists of a largely unconscious cognitive process. Consequently, it is difficult, if not impossible, to accurately capture this process using items in a post-experimental questionnaire. We therefore leave it to future research to identify and clearly disentangle all possible mechanisms through which information organization in performance reports can affect performance evaluation judgments.

The data from our experiment support H1a, but not the other three hypotheses, H1b, H2a and H2b. Future research is also needed to establish whether these findings should lead to refutation of our theory, or whether they are caused by insufficient power or particular features of the design of the experiment. Three aspects of our experiment require some specific attention because they point toward important limitations of our study. First, our participants were undergraduate students with relatively little – if any – experience in evaluating subordinates. Next, we used a scenario-based experiment in which we asked participants to assume the role of a manager. While scenario-based experiments are widely used in judgment and decision-making research in accounting and other fields, we acknowledge that we cannot rule out that our results are affected by the fact that participants only had to imagine what it would feel like to provide performance ratings and never really experienced the costs associated with making decisions that might not be well received by subordinates. Finally, one reason that we did not find any differences between the table and the graph conditions might be that our graphs manipulation was not as strong as it could have been. Specifically, rather than using 'pure' graphical constructs (Vessey 1991, p. 235) we chose to provide participants in the graph conditions with horizontal bar charts with numerical figures placed at the end of the bars. The reason is that it is unlikely that firms will ever provide pure graphs in performance reports and that our primary aim was to investigate how performance report design affects evaluation judgments. We agree with Vessey (1991) that more research is needed to establish how salient the 'spatial', as opposed to 'symbolic', aspects of a (performance) report need to be, before graphs and tables lead decision-makers toward different conclusions. Also, given the large variety of graphs that management accountants can choose to include in performance reports (Jarvenpaa and Dickson 1988), we are looking forward to seeing future studies that not only compare graphs and tables but also studies that compare different types of graphs.

Our finding that information organization affects leniency but not compression also raises questions about the similarity of the processes behind these two phenomena. Following existing literature (e.g. Moers 2005, Bol 2011, Chen 2014), we theorized that leniency and compression would be two different outcomes of the same process. In other words, we assumed that managers who would rather give their subordinates 'the benefit of the doubt' than search for information that might indicate that a lower rating would be more appropriate, would also be more likely

to underestimate than overestimate performance differences between their subordinates. The results require us to reassess the validity of this assumption. Similarly, we reasoned that information organization and presentation format would have similar effects on managers' performance ratings, but found that that was not the case. While, as indicated, it is possible that this is due to some experimental design choices, for example, the specific types of graphs that we used, we cannot rule out that the reason is more fundamental. For example, it is possible that information organization and presentation format affect different aspects or subroutines of cognitive information acquisition processes. We are looking forward to future research that sheds more light on these important issues, and that develops and tests theories about how managers search, compare and weight items in performance reports, such as recent accounting studies using eye-tracking technology to increase our understanding of how decision-makers extract information from balanced scorecards (e.g. Chen et al. 2016, Dalla Via et al. 2016, Kramer and Maas 2016).

In summary, our paper makes some important contributions, but also raises a number of questions which point toward interesting avenues for future research. While both evaluators and evaluees generally prefer performance appraisal and feedback to take place through rich media channels (Mintzberg 1973, Brutus 2010), such as face-to-face meetings, evaluation and bonus determination of unit managers often takes the form of evaluators subjectively weighting a number of objective performance signals in a report to arrive at one overall rating. A better understanding of how evaluators deal with the inherent limitations of such performance evaluation systems will enable organizations to more optimally use the opportunities provided by modern technologies.

Acknowledgements

We thank the editors and two anonymous reviewers for their guidance and suggestions. We are also grateful to Sophie Hoozée, Kun Huo, Stephan Kramer, Andreas Ostermaier, Marcel van Rinsum, the class of 2015 of the MSc Accounting, Auditing and Control at Erasmus School of Economics, and workshop participants at the 2014 AAA Annual Meeting in Atlanta and the 2015 EAA Annual Meeting in Glasgow for useful comments on earlier drafts of the paper.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes

1. Bol et al. (2016) examine how information accuracy (i.e. the extent to which performance reports are informative about employees' actions) and outcome transparency (i.e. the extent to which performance evaluation outcomes are public knowledge) affect managers' tendency to compress overall performance ratings. They find that compression is significantly lower in settings in which information is relatively accurate and outcome transparency is relative high, compared to all other settings.
2. There is some debate in the literature about whether leniency in evaluation can also have positive consequences for organizations. For example, Bol (2011) found that lenient ratings were associated with higher increases in employee effort.
3. Our hypotheses focus on two possible interventions to reduce the leniency and compression of performance ratings; however, our theory is silent on the combined effect of these interventions. We use a full factorial design to test our hypotheses because we could not rule out *ex ante* that the effect of one intervention is conditional upon the presence of the other.
4. The decision to have the participants rate eight (as opposed to some other number of) employees followed a tradeoff of theoretical and practical considerations. From a theory perspective, it is important that the number is high enough to make the evaluation process cognitively difficult. Much research shows that cognitive limitations manifest themselves when individuals need to process seven or

more information cues in a judgment task ('the magical number seven'; e.g. Miller 1956, Baddeley 1994). Having many more than seven employees, however, would have caused practical problems, given the limited timeframe within which the evaluation decisions had to be made.

5. In each of the four categories, we picked two measures that seemed relatively easy to understand. In a few cases, we slightly altered the exact wording of the measure (e.g. we used 'Return on sales' instead of 'Sales margins'). Using a pilot study, Cardinaels and van Veen-Dirks (2010) explicitly established that all measures in their scorecard were considered typical for their category and that there were no differences in the extent to which measures were considered typical for their category across categories. This provides us with some confidence that also in our experiment the measures were typical for their category.
6. We varied the target levels because varying target levels across units is common practice in multi-outlet businesses (e.g. Ghosh and Lusch 2000, Ittner et al. 2003, Bouwens and Kroos 2011) and contributes to the cognitive complexity of the evaluation task.
7. The order was Store 7, Store 6, Store 4, Store 3, Store 5, Store 1, Store 2 and Store 8. The store numbers here indicate the rank order in terms of average deviation from target performance, from best (Store 1: +7.5%) to worst (Store 8: −7.5%).
8. Unless indicated otherwise, all reported *p*-values are two-tailed. While this check provides evidence of the overall effectiveness of the manipulation, the mean scores in both treatment groups are both between 3 and 4, indicating that, on average, participants in the alphabetical order conditions also signaled agreement with the item. A closer look at the data reveal that 45 participants (48.9%) in the alphabetical information organization treatment scored higher than the midpoint of 3 on this item. Also, 10 participants (11%) in the categorized information organization treatment scored lower than the midpoint of 3, indicating that they disagreed with that statement. If we run the hypotheses tests in a reduced sample consisting of only the 128 participants who scored on the 'correct side' of the answer scale (including the neutral 3 for both groups), the results are qualitatively similar as in the main analysis. We find marginally significant main effects of information organization on *AVERAGE RATING* ($F_{1,122} = 3.73$; $p = .056$) and on *MEDIAN RATING* ($F_{1,122} = 3.69$; $p = .057$), supporting H1a, but we find no support for the other hypotheses.
9. Note that higher values of *ABSOLUTE DIFFERENCE* indicate less compression, whereas higher values of *RELATIVE DIFFERENCE* indicate more compression.

ORCID

Victor S. Maas http://orcid.org/0000-0002-6132-707X

References

Alter, A. and Oppenheimer, D., 2009. Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review*, 13 (3), 219–235.

Baddeley, A., 1994. The magical number seven: still magic after all these years? *Psychological Review*, 101 (2), 353–356.

Beattie, V. and Jones, M.N., 2008. Corporate reporting using graphs: a review and synthesis. *Journal of Accounting Literature*, 27, 71–110.

Bettman, J.R. and Kakkar, P., 1977. Effects of information presentation format on consumer information acquisition strategies. *Journal of Consumer Research*, 3 (4), 233–240.

Birnberg, J.G. and Shields, M.D., 1984. The role of attention and memory in accounting decisions. *Accounting, Organizations and Society*, 9 (3–4), 365–382.

Bol, J.C., 2008. Subjectivity in compensation contracting. *Journal of Accounting Literature*, 27, 1–32.

Bol, J.C., 2011. The determinants and performance effects of managers' performance evaluation biases. *Accounting Review*, 86 (5), 1549–1575.

Bol, J.C., Kramer, S., and Maas, V.S., 2016. How control system design affects performance evaluation compression: the role of information accuracy and outcome transparency. *Accounting, Organizations and Society*, 51, 64–73.

Bouwens, J. and Kroos, P., 2011. Target ratcheting and effort reduction. *Journal of Accounting and Economics*, 51 (1), 171–185.

Brutus, S., 2010. Words versus numbers: a theoretical exploration of giving and receiving narrative comments in performance appraisal. *Human Resource Management Review*, 20 (2), 144–157.

Cardinaels, E., 2008. The interplay between cost accounting knowledge and presentation formats in cost-based decision-making. *Accounting, Organizations and Society*, 33 (6), 582–602.
Cardinaels, E. and van Veen-Dirks, P.M.G., 2010. Financial versus non-financial information: the impact of information organization and presentation in a balanced scorecard. *Accounting, Organizations and Society*, 35 (6), 565–578.
Chase, W.G. and Simon, H.A., 1973. The mind's eye in chess. In: W.G. Chase, ed. *Visual Information Processing*. Oxford, UK: Academic Press, 215–281.
Chen, Y.-L., 2014. Determinants of biased subjective performance evaluations: evidence from a Taiwanese public sector organization. *Accounting and Business Research*, 44 (6), 656–675.
Chen, Y., Jermias, J., and Panggabean, T., 2016. The role of visual attention in the managerial judgment of balanced scorecard performance evaluation: insights from using an eye-tracking device. *Journal of Accounting Research*, 54 (1), 113–146.
Cheng, M.M. and Humphreys, K.A., 2012. The differential improvement effects of the strategy map and scorecard perspectives on managers' strategic judgments. *Accounting Review*, 87 (3), 899–924.
Craik, F.I.M. and Lockhart, R.S., 1972. Levels of processing: a framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, 11 (6), 671–684.
Daft, R.L. and Lengel, R.H., 1986. Organizational information requirements, media richness and structural design. *Management Science*, 32 (5), 554–571.
Dalla Via, N., van Rinsum, M., and Perego, P. 2016. How accountability influences investment decision quality in a balanced scorecard setting: an eye-tracking experiment. Unpublished Working Paper.
Davison, J., 2015. Visualising accounting: an interdisciplinary review and synthesis. *Accounting and Business Research*, 45 (2), 121–165.
Demere, W., Sedatole, K.L., and Woods, A. 2016. The role of calibration committees in subjective performance evaluation systems. Unpublished Working Paper. Available from SSRN: https://ssrn.com/abstract=2360076.
DeNisi, A.S., Cafferty, T.P., and Meglino, B.M., 1984. A cognitive view of the performance appraisal process: a model and research propositions. *Organizational Behavior and Human Performance*, 33 (3), 360–396.
Ding, S. and Beaulieu, P., 2011. The role of financial incentives in balanced scorecard-based performance evaluations: correcting mood congruency biases. *Journal of Accounting Research*, 49 (5), 1223–1247.
Furnham, A. and Stringfield, P., 2001. Gender differences in rating reports: female managers are harsher raters, particularly of males. *Journal of Managerial Psychology*, 16 (4), 281–288.
Ghosh, D. and Lusch, R. F., 2000. Outcome effect, controllability and performance evaluation of managers: some field evidence from multi-outlet businesses. *Accounting, Organizations and Society*, 25 (4), 411–425.
Golman, R. and Bhatia, S., 2012. Performance evaluation inflation and compression. *Accounting, Organizations and Society*, 37 (8), 534–543.
Höppe, F. and Moers, F., 2011. The choice of different types of subjectivity in CEO annual bonus contracts. *Accounting Review*, 86 (6), 2023–2046.
Ittner, C.D., Larcker, D.F., and Meyer, M.W., 2003. Subjectivity and the weighting of performance measures: evidence from a balanced scorecard. *Accounting Review*, 78 (3), 725–758.
Jarvenpaa, S.L. and Dickson, G.W., 1988. Graphics and managerial decision making: research-based guidelines. *Communications of the ACM*, 31 (6), 764–774.
Johnston, W.J. and Shields, M.D., 1983. Evaluating the newer salesperson. *Industrial Marketing Management*, 12 (3), 193–199.
Johnson, E.N., Reckers, P.M.J., and Bartlett, G.D., 2014. Influences of timeline and perceived strategy effectiveness on balanced scorecard performance evaluation judgments. *Journal of Management Accounting Research*, 26 (1), 165–184.
Kaplan, R.S. and Norton, D.P., 1996. *The Balanced Scorecard: Translating Strategy into Action*. Boston, MA: Harvard Business School Press.
Kaplan, S.E., Petersen, M.J., and Samuels, J.A., 2007. Effects of subordinate likeability and balanced scorecard format on performance-related judgments. *Advances in Accounting*, 23, 85–111.
Kelton, A.S., Pennington, R.R., and Tuttle, B.M., 2010. The effects of information presentation format on judgment and decision making: a review of the information systems research. *Journal of Information Systems*, 24 (2), 79–105.
Kramer, S. and Maas, V.S., 2016. Selective attention to performance measures and bias in subjective performance evaluations: an eye-tracking study. Working Paper. Available from SSRN: http://ssrn.com/abstract = 2457941.

Kramer, S., Maas, V.S., and van Rinsum, M., 2016. Relative performance information, rank ordering and employee performance: a research note. *Management Accounting Research*, 33, 16–24.
Larkin, J.H. and Simon, H.A., 1987. Why a diagram is sometimes worth ten thousand words. *Cognitive Science*, 11 (1), 65–100.
Lazear, E.P., 2000. Performance pay and productivity. *American Economic Review*, 90 (5), 1346–1361.
Libby, T., Salterio, S.E., and Webb, A., 2004. The balanced scorecard: the effects of assurance and process accountability on managerial judgment. *Accounting Review*, 79 (4), 1075–1094.
Liedtka, S.L., Church, B.K., and Ray, M.R., 2008. Performance variability, ambiguity intolerance, and balanced scorecard-based performance assessments. *Behavioral Research in Accounting*, 20 (2), 73–88.
Lipe, M.G. and Salterio, S.E., 2000. The balanced scorecard: judgmental effects of common and unique performance measures. *Accounting Review*, 75 (3), 283–298.
Lipe, M.G. and Salterio, S.E., 2002. A note on the judgmental effects of the balanced scorecard's information organization. *Accounting, Organizations and Society*, 27 (6), 531–540.
Long, J.H., Mertins, L., and Vansant, B., 2015. The effect of firm-provided measure weightings on evaluators' incorporation of non-contractible information. *Journal of Management Accounting Research*, 27 (1), 47–62.
Lowe, D.J., Carmona-Moreno, S., and Reckers, P.M.J., 2011. The influence of strategy map communications and individual differences on multidimensional performance evaluations. *Accounting and Business Research*, 41 (4), 375–391.
Maas, V.S., van Rinsum, M., and Towry, K.L., 2012. In search of informed discretion: an experimental investigation of fairness and trust reciprocity. *Accounting Review*, 87 (2), 617–644.
Marginson, D., 2006. Information processing and management control: a note exploring the role played by information media in reducing role ambiguity. *Management Accounting Research*, 17 (2), 187–197.
Miller, G.A., 1956. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, 63 (2), 81–97.
Mintzberg, H., 1973. *The Nature of Managerial Work*. New York, NY: Harper & Row.
Moers, F., 2005. Discretion and bias in performance evaluation: the impact of diversity and subjectivity. *Accounting, Organizations and Society*, 30 (1), 67–80.
Moriarty, S., 1979. Communicating financial information through multidimensional graphics. *Journal of Accounting Research*, 17 (1), 205–224.
Murphy, K.R. and Cleveland, J.N., 1995. *Understanding Performance Appraisal. Social, Organizational, and Goal-Based Perspectives*. Thousand Oaks, CA: Sage Publications.
Payne, J.W., Bettman, J.R, and Johnson, E.J., 1993. *The Adaptive Decision Maker*. Cambridge, UK: Cambridge University Press.
Rennekamp, K., 2012. Processing fluency and investors' reactions to disclosure readability. *Journal of Accounting Research*, 50 (5), 1319–1354.
Russ, G.S., Daft, R.L., and Lengel, R.H., 1990. Media selection and managerial characteristics in organizational communications. *Management Communication Quarterly*, 4 (2), 151–175.
Saal, F.E. and Landy, F.J., 1977. The mixed standard rating scale: an evaluation. *Organizational Behavior and Human Performance*, 18 (1), 19–35.
Salterio, S.E., 2012. Balancing the scorecard through academic accounting research: opportunity lost? *Journal of Accounting and Organizational Change*, 8 (4), 458–474.
Schwarz, N. 2011. Feelings-as-information theory. In: P.A.M. van Lange, A.W. Kruglanski, and E.T. Higgins, eds. *Handbook of Theories of Social Psychology: Volume One*. London, UK: Sage Publications, 289–308.
Shanteau, J., 1988. Psychological characteristics and strategies of expert decision makers. *Acta Psychologica*, 68 (1), 203–215.
Spencer, S.J., Zanna, M.P., and Fong, G.T., 2005. Establishing a causal chain: why experiments are often more effective than mediational analyses in examining psychological processes. *Journal of Personality and Social Psychology*, 89 (6), 845–851.
Tang, F., Hess, T.J., Valacich, J.S., and Sweeney, J.T., 2014. The effects of visualization and interactivity on calibration in financial decision-making. *Behavioral Research in Accounting*, 26 (1), 25–58.
Thaler, R.H. and Sunstein, C.R., 2008. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. New Haven, CT: Yale University Press.
Van Rinsum, M. and Verbeeten, F.H.M., 2012. The impact of subjectivity in performance evaluation practices on public sector managers' motivation. *Accounting and Business Research*, 42 (4), 377–396.
Vessey, I., 1991. Cognitive fit: a theory-based analysis of the graphs versus tables literature. *Decision Sciences*, 22 (2), 219–240.

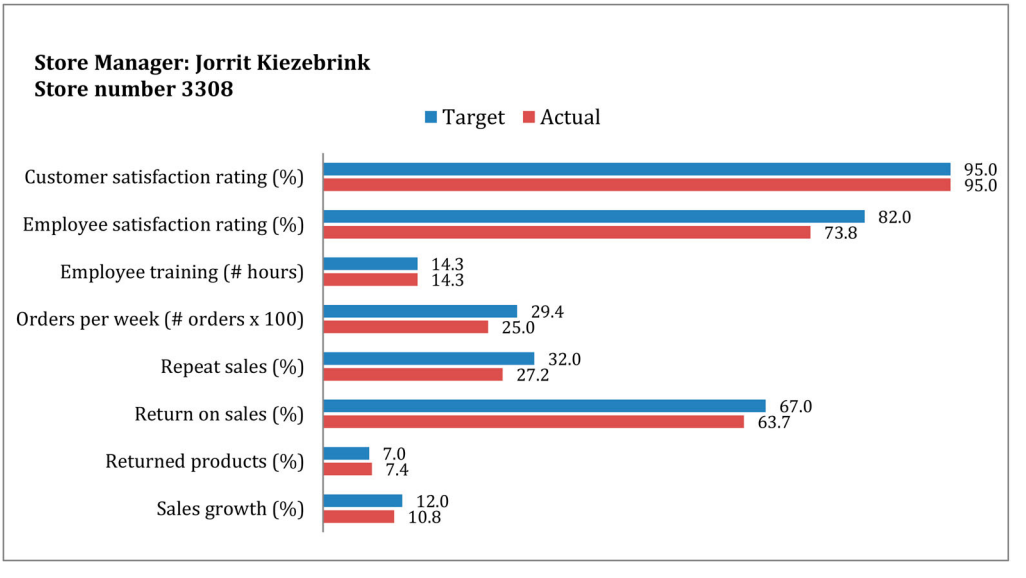
Appendix

The same performance report in each of the four experimental conditions

Alphabetical organization – Table condition

Store Manager Store number 3308	Jorrit Kiezebrink	
	Target	Actual
Customer satisfaction rating (%)	95.0	95.0
Employee satisfaction rating (%)	82.0	73.8
Employee training (# hours)	14.3	14.3
Orders per week (# orders × 100)	29.4	25.0
Repeat sales (%)	32.0	27.2
Return on sales (%)	67.0	63.7
Returned products (%)	7.0	7.4
Sales growth (%)	12.0	10.8

Alphabetical organization – Graph condition

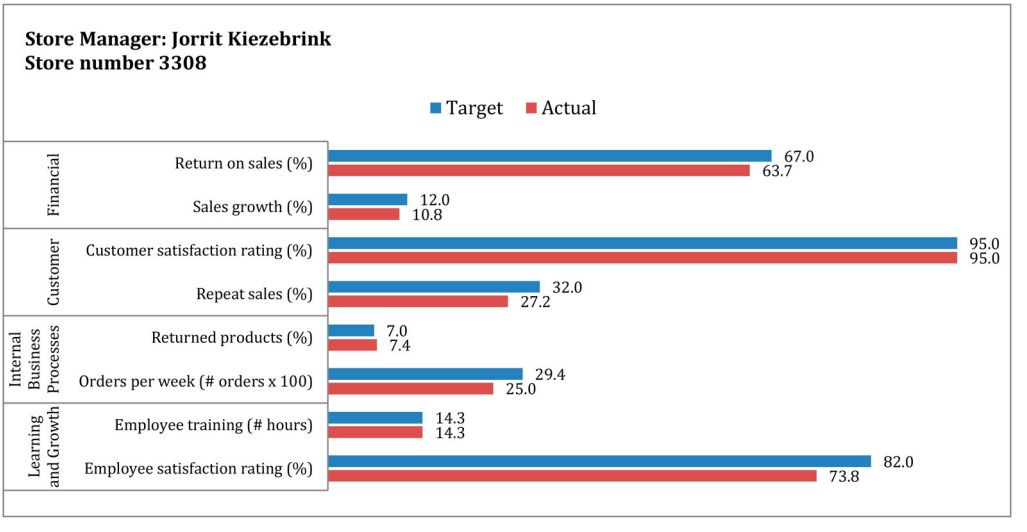


Note: Black and white in print and colored in online

Categorical (balanced scorecard) organization – Table condition

Store Manager Store number 3308		Jorrit Kiezebrink	
		Target	Actual
Financial	Return on sales (%)	67.0	63.7
	Sales growth (%)	12.0	10.8
Customer	Customer satisfaction rating (%)	95.0	95.0
	Repeat sales (%)	32.0	27.2
Internal Business	Returned products (%)	7.0	7.4
Processes	Orders per week (# orders × 100)	29.4	25.0
Learning and Growth	Employee training (# hours)	14.3	14.3
	Employee satisfaction rating (%)	82.0	73.8

Categorical (balanced scorecard) organization – Graph condition



Note: Black and white in print and colored in online