On the Ignorance of Group-Level Effects – The Tragedy of Personnel Evaluation?

Momme von Sydow

University of Munich, University of Heidelberg,

Niels Braus

University of Heidelberg

Ulrike Hahn

Birkbeck University of London, University of Munich

Author Note:

Momme von Sydow, Munich Center for Mathematical Philosophy (MCMP), University of Munich; and Department of Psychology, University of Heidelberg,

Niels Braus, Department of Psychology, University of Heidelberg

Ulrike Hahn, Birkbeck University of London, Department of Psychology and MCMP, University of Munich

Correspondence concerning this article should be addressed to Momme von Sydow, Ludwig-Maximilians-Universität München, MCMP, Geschwister-Scholl-Platz 1, D-80539 München, Email: momme.von-sydow@lrz.uni-muenchen.de

Abstract: In social-dilemma situations (e.g., public-good games), people may pursue their local self-interests, thereby lowering the overall payoff of their group and, paradoxically, even their individual payoffs as a result. Likewise, in inner-individual dilemmas, even without conflict of interest between persons, people may pursue local goals at the expense of overall utility. Our experiments investigate such dissociations of individual and group-level effects in the context of personnel evaluation and selection. Participants were given the role of human resource managers selecting workers to optimize the overall payoff for the company. We investigated contexts where the individually best/worst 'employees' systematically caused the worst/best group performance. When workers in a team could substantially increase or decrease co-workers' performance, most participants (albeit not all) tended to focus solely on individual performance without considering their overall contribution even when instructed to maximize group performance. This undue focus on individual information meant that employees who enhanced team performance the most often received the most negative evaluations. This may result in a 'tragedy of personnel evaluation' relevant to maladaptive incentive structures (personnel evaluation), job offers (personnel selection), and a substantially negative impact on organizational effectiveness. At the same time, the results suggest ways this problem may be overcome.

Public Significance Statement: We investigate whether participants in the role of personnel managers who obtained information on both individual and team earnings readily take into account both individual and overall influences on group performance. We found participants neglected even large group-level effects. This raises important research questions and serves as a warning to practitioners in human resource management against a potentially tragic underestimation of those who are best overall for organisations and companies.

Keywords: global vs. local optimization; personnel evaluation and selection; Simpson's paradox; inner-individual dilemmas; altruist and egoist detection; teams; causal induction

Intra-Organizational Dilemmas:

Conflicts between Individual-Level and Team-Level Optimization

Adam Smith (1776) argued that altruism is not needed to promote the common good: “By pursuing his own interest he [the individual] frequently promotes that of the society more effectually than when he really intends to promote it.” In the wider context of evolutionary biology, philosophy, economics and psychology, however, it has been increasingly noted that social dilemmas can arise whereby groups of organisms or human agents individually maximising their self-interest (and in this sense acting ‘optimally’) may not only lead to outcomes that are sub-optimal for the collective, but actually reduce the payoffs ultimately obtained by the individuals. For instance, it has been argued that the over-exploitation and destruction of finite public resources by ‘rationally’ acting, selfish individuals is inevitable. This so-called “tragedy of the commons” has been discussed with regard to environmental pollution and sustainable development (Hardin, 1968). Social dilemmas, such as this, have been widely studied, and ensuing debates in behavioural economics and psychology have often concerned possible solutions to social dilemma situations. Research in various experimental paradigms has also eroded the expressed or tacit strict assumption of egoism, while still addressing problems and limits of altruism (Fehr & Fischbacher, 2003, 2004; Fehr & Schmidt, 1999; Fehr & Gächter, 2002; Henrich, 2005; Ostrom et al., 1999; Everett, Pizarro, & Crockett, 2016). Likewise, evolutionary biology and philosophy of biology have shifted from emphasizing individual egoism (or gene-egoism) to acknowledging multi-level approaches, which generally suggest the presence of egoistic as well as altruistic behaviour tendencies in social groups (Nowak & Sigmund, 2005; Sober & Wilson, 1999; Wilson & Wilson, 2007; von Sydow, 2012).

In companies and organizations, structural tensions between egoistic and group-serving behaviours have direct relevance as well (e.g., Dalal, 2005; George & Bettenhausen, 1990; Li, Kirkman, & Porter, 2014). The issue of egoistic and group-serving behaviours does not only raise moral questions regarding the common good of a society, but even arises within companies, even when the clear aim is to optimize key economic operating figures (e.g., net sales or operating profit). This may pose intra-organizational dilemmas, with respect to building efficient organizations.

One possible problem may arise with respect to personnel selection. A personnel manager, in the role of a neutral third party, may select people based on their individual (local) performance alone, ignoring their indirect effects on others and thus their overall effect on group performance. This may lead them to select teams that are clearly suboptimal.

Likewise, an individual may pursue too many projects (*local goals*), each with a positive utility, but may thereby reduce overall utility (global goal) by ignoring these projects' negative (or positive) external effects on other projects – that is, by ignoring interactions between these projects.

These kinds of problems have been called "inner-individual" or "intra-individual" dilemmas (IID; von Sydow, 2015). Social dilemmas are situations in which agents who individually optimize their outcome decrease the group-level outcome and thereby also their own payoffs. An IID likewise involves local and global levels, but within a single individual's payoff structure.

From a game-theoretical perspective, however, social dilemmas and inner-individual dilemmas differ substantially. Whereas a self-interested 'optimal' strategy in social dilemmas involves conflict of interests with others and should, game-theoretically, not lead to the common good on a group level, the rational-choice solution to an IID clearly requires deciding for the globally optimal solution (without any resulting motivational tensions). One should clearly optimize one's goals globally, and a personnel manager, as neutral third party, should quite clearly try to optimize the *overall* success of the company and thus refrain from selecting or rewarding employees with good individual operating figures if overall they are bad for the group, company or organization. Even for inner-individual dilemmas where the optimal solution clearly seems to be overall optimization, it has been shown that people may sometimes optimize locally at the expense of the global level (von Sydow, 2015). The local-global distinction in inner-individual dilemmas perhaps resembles the tension between short-term (local) vs. long-term (global) utilities in delay of gratification or inter-temporal choice paradigms (Mischel, Shoda, & Peake, 1988, Frederick, Loewenstein, & O'Donoghue, 2002; Curry, Price, & Price, 2008). If the findings on inner-individual dilemmas – that is, that people may sometimes perhaps simply not realize positive externalities of local effects – are transferable to a human resource management context, people should ignore interactions (e.g., the positive interactions of an altruist or the negative interactions of an egoist) with other employees.

One related fundamental question of human factors research is whether people in the role of personnel managers can detect that high-performing teams are not necessarily those composed of the best performing individuals. Can they assemble the best performing group in circumstances where individual performances interact? We will address this question in an experimental setting exploring number-based performance evaluation in a task where we provide repeated information on both individual and group performance levels (in what we call a Two-Level Personnel Evaluation Tasks, T-PETs).

This question is important, since organisational psychology has increasingly stressed importance of the role of inter- or super-individual factors, both under the heading of employee interactions and team performance. Research has acknowledged the role of *teams* beyond mere individual contributions (Beersma, Hollenbeck, Humphrey, Moon, Conlon, & Ilgen, 2003; DeShon, Kozlowski, Schmidt, Milner, & Wiechmann, 2003; Mathieu, Maynard, Rapp, & Gilson, 2008; Mathieu, Tannenbaum, Donsbach, & Alliger, 2014; Chen, Farh, Campbell-Bush, Wu, & Wu, 2013; Memmert, Plessner, Hüttermann, Froese, Peterhänsel, & Unkelbach, 2015). Similarly there are also several notions linked to employees' *interactions* with others or their environment; for instance, van Scotter & Motowidlo (1996) introduced the term "interpersonal facilitation" as a subcategory of contextual performance. Co-workers may interact in a variety of ways, including considerate, co-operative or helpful acts; or a co-worker may catalyze higher performance, for instance by increasing the others' motivation or by contributing to an atmosphere that enhances performance. In particular, the role of pro-social behaviour and altruism is increasingly understood to be important for functioning societies or smaller groups, in both psychology and behavioural economics (Engel, 2011; Hendrich et al., 2005; Gollwitzer, Rothmund, Pfeiffer, & Ensenbach, 2009; Post, 2005). Thus people's ability to detect both anti- and pro-social behaviours seems important. This is especially the case since organisational psychology has also begun to stress the role of prosocial or altruistic behaviours in teams (George & Bettenhausen, 1990; Beersma et al., 2003; Li, Kirkman, & Porter, 2014) as essential to the success of organizations and companies. In particular, organizational citizenship behaviour (OCB), concerned mostly with prosocial or altruistic behaviours not prescribed by role descriptions (or task descriptions; Organ, 1997) has become a flourishing field of research, finding reliable effects of OCB on productivity, efficiency, sales, revenue, reduced costs, customer satisfaction, and performance quality (Grant & Patil, 2012; Nielsen, Hrivnak, & Shaw, 2009; Organ & Ryan, 1995; Pearce & Herbik, 2004; Podsakoff, Whiting, Podsakoff, & Blume, 2009; Podsakoff, Whiting, Podsakoff, & Mishra, 2010). Nonetheless, some potentially dysfunctional effects also need mentioning (cf. Brief & Motowidlo, 1986). Advantages of adding team-level goals, motivations and incentives have been acknowledged (e.g., Chen & Kanfer, 2006; Chen et al., 2013; DeShon et al., 2003), and the positive role of prosocial behaviour (or high "agreeableness"), at least in a 'co-operative' incentive structure, has been shown (Beersma et al., 2003). Brief & Motowidlo (1986) have identified thirteen specific kinds of prosocial organizational behaviours that can be assumed mostly to improve performance. Prosocial interactions between employees in organizational contexts thus clearly occurs in many forms, and varies in several dimensions – for instance whether targeted at specific individuals or at larger organizational levels, whether role- or extra-role-behaviour, whether intentional or unintentional, or whether or not mediated by actions, emotions or cognitions (Brief & Motowidlo, 1986; Li, Kirkman, & Porter, 2014).

It is known that individuals' performance can vary as a function of team membership (Stewart & Nandkeolyar, 2007; cf. Mathieu et al., 2014). Hence the issue of detecting interactions between employees and altruists or egoists in particular needs further attention in Human resource management (Becker & Gerhart, 1996; Mathieu et al., 2014). Research has just begun to show that the acknowledged importance of group interactions and prosocial behaviour is at least sometimes reflected in personnel evaluation or selection. It has been suggested that in-role or extra-role prosocial behaviour is sometimes directly or indirectly rewarded by organizations (Organ, 1997; Scotter, Cross, & Motowidlo, 2000; Grant & Patil, 2012, 562). This suggests that managers at least in principle could recognize such behaviours. This may however be based in part on knowledge about moral attitudes of employees (cf. Everett, Pizarro, & Crockett, 2016), without considering performance data.

Contempory performance data is often number-based. Such number-based evaluations may represent altruism (or prosocial behaviour) mathematically as a positive performance interaction with others and decreased individual performance (cf. Sober, 1998). So it is interesting how people deal with such contradicting performance data on the individual and group levels and how readily employers detect 'value' in contexts in which group and individual perspectives conflict. This is the gap that the present study seeks to address. We employ a basic number-based task and treat the varieties of interaction processes as "black box", as is actually often done in outcome-based evaluations. Specifically, we investigate how people perceive an individual's direct effect on team-earnings (individual level) relative to his or her overall effect, including all indirect effects on others. Specifically, we provide participants with cases in which these two levels are clearly dissociated. The task mirrors that of employee-analysis based on *quantitative* operating figures. Although we have doubts about whether an exclusively quantitative and performance-based personnel-evaluation is a viable approach to human resource management, units in large and middle-size organizations increasingly base decisions on personnel information systems (PIS) and quantitative information only (Brandl, 2002). Thus our studies may also be interpreted as exploring potential consequences of this lack of qualitative data when people deal with quantitative performance on the individual and group or team levels only and do not have access to richer information about interactions about employees.

People in charge of personnel-evaluation and incentive systems might tend, in number-based evaluations, to ignore distributed positive or negative effects of individuals on the work of other team members (externalities), concentrating only on a person’s direct operating figures. This would suggest a potential “tragedy of personnel selection”, since people with the best overall effects on group-performance (e.g., by direct help or interpersonal facilitation) might be evaluated the most negatively.

Two-Level Personnel Evaluation Tasks (T-PETs)

The experiments here employ the same personnel evaluation or selection scenarios, involving a positive interactor (facilitator) or a negative interactor (inhibitor). The interacting employee who is individually the lowest (or highest) earner of a team, but who contributes the most (or least) overall to its success, is referred to here, somewhat simply, as “the altruist”, and the negative interactor as “the egoist”. One should note, however, that although altruists may often display patterns of lowered individual performance with increased group performance (in line with standard behavioural definitions of altruism; Sober, 1998), here information is given neither on the underlying motivation nor on whether the low individual performance of the altruist is *caused* by sacrificing in favour of the group.[1] The same precautionary note is in place for the term “egoist”.

[1] Even outside evolutionary biology the definition of altruism is sometimes purely behavioral, not requiring altruistic motives (see Li, Kirkman, & Porter, 2014). We do not generally endorse this terminologically position, but mainly use the terminology here for mnemonic reasons. Moreover, altruism will at least often lead to this structure. Even if the term ‘facilitator’ may perhaps be more appropriate, we

Thus the evaluation of the interactor should differ at the individual and the overall-group levels. Participants assume the role of human resource managers, evaluating the utility of employees in a local snack bar of a chain. A manager obtains daily information on direct earnings of individuals and on overall earnings of the shifts (changing teams of four employees). The observed overall profit strongly depends positively (or negatively) on the presence of one specific individual employee in a shift ($r = .99$, $p < .001$), suggesting that, despite having the lowest (highest) individual contribution, his or her presence substantially affects the others' (and thus the team's) outcome.

The aim was to see how far the participants (as human resources managers) considered the partly indirect overall monetary contributions of the "altruist" (or "egoist") when evaluating their workers. In all experiments here we study simple examples in which an interactor has a positive (or negative) effect on *all* members of a group (or on the team as a whole). This is assumed by influential biological models of altruism (Sober & Wilson, 1999; Wilson & Wilson, 2007; but cf. Nowak & Sigmund, 1998, 2005; Kiyonari & Barcley, 2008; Rand, Dreber, Ellingsen, Fudenberg, & Nowak, 2009) and reflects the notion of overall group-level effects, for instance, of team motivation, team goals, team climate, cohesion, collective cognition, leadership, and group identity (DeShon et al., 2003; Mathieu et al., 2008, 2014; Haslam et al., 2017). Moreover, in the personnel management context, we focus on only *one* altruist (or egoist) who has a *large* impact on a relatively *small* group. Even though many real-life examples may involve larger groups, more interactors, or a more subtle impact of the interactor, it seemed interesting to explore first whether personnel managers had problems detecting strong group level effects in a relatively simple setting.

kept the more specific notion 'altruist' as metonym, because it is the common term linked to the presumably most important exemplification of designated class.

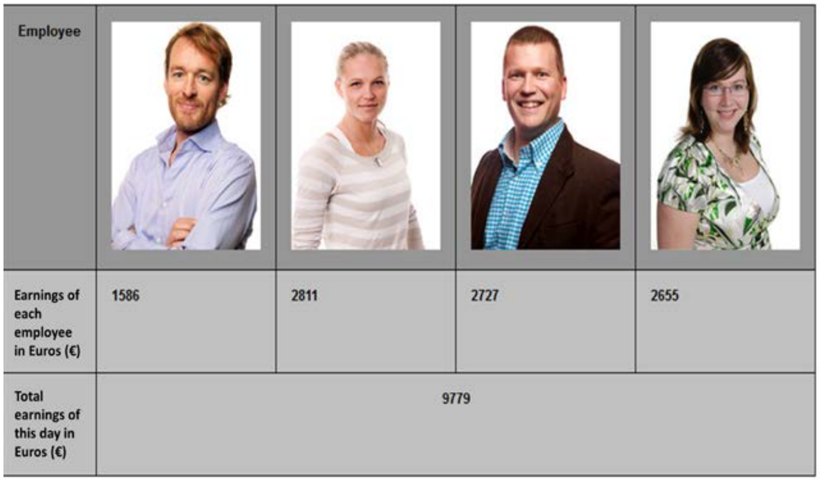
Exploratory T-PET

In a first exploratory T-PET study (reported in von Sydow & Braus, 2016), we looked at the general question of whether people detect more quickly the individual contribution or the overall group-level contribution of an ‘altuist’ in the above sense, and whether they realise the strength of the altruists’ indirect impact on the overall group earnings. The study had a 2 x 2 (rounds: 10 working days, versus 20 working days; and low versus high impact of the altruist on earnings of the normal workers; with positive correlations *r* = .96, *r* = .99, both *p* < .001 in 10 rounds; see Table 1), within-subjects design.

Table 1. *Mean earnings of normal workers (NW) and altruist (A); and mean overall earnings with or without altruist in a pretest study (N = 124, MTurk)*

	Condition 1 / 2	Condition 3 / 4
	Small Impact of *A*	Large Impact of *A*
NW without A	2200 €	2000 €
NW with A	2800 €	3000 €
Altruist	1600 €	1600 €
Overall without A	8800 €	8000 €
Overall with altruist	10000 €	10600 €

Participants in a computerized experiment were asked to imagine being a human resources manager evaluating the staff of a particular snack bar. In the computer scenario there are five staff members, but each day only four are working. Compared to many real personnel-evaluation situations, this five-persons scenario is relatively simple (von Sydow, 2015, for a plausible inner-individual dilemma with ten rather than five ‘variables’ in a different context). Participants in this personnel evaluation scenario were instructed to establish which workers contributed most to the company’s overall profit, based on data provided by the reporting unit of the larger company. After reading the instructions, participants received information regarding the per-day individual earnings of the four employees (along with their photographs), followed by total group earnings (Figure 1).



Employee				
Earnings of each employee in Euros (€)	1586	2811	2727	2655
Total earnings of this day in Euros (€)	9779			

Figure 1. Example of shown earnings at the individual and group levels on a particular day.

The earnings of both the normal workers and the altruist worker in each shift were based on the mean earnings shown in Table 1. This value was presented with some noise surrounding each value (a normal distribution with SD = 600 €). This nonetheless left the group effect of the altruist the dominant effect. The altruist role was randomly assigned to one of the pictures and randomly appeared in 6 of 10 rounds (days) with the remaining normal workers for that shift also selected at random (thus on average 7.5 times in 10 rounds). Participants could view the overview panels for each day as long as they wanted. After the 10 (or, in the extended practice condition, 20) rounds, the ‘human resources managers’ evaluated the employees of the snack bar in several tasks. The results showed that, in all conditions, a clear majority of participants judged the altruist as having not the highest but the lowest “overall utility for the company”, although clearly being the most highly associated with overall utility.

Additionally, there was a small but reliable increase in group-oriented judgments in the high-group-impact conditions. Rating tasks likewise showed that the altruist was generally rated lower than the normal workers. We also coded the comments of the participants, and interestingly only 10 of the 124 mentioned interaction between participants or different effects on the individual or group level; as, for instance: “It is always important to look beyond what is obvious. Like in this task wherein at first glance the girl with short hair and blue shirt seemed to be lagging behind; but after careful scrutiny, she is obviously leading the group. It may be affecting her individual performance, but the group's earnings is way high[er] when she is around.” Or: “The blonde lady with short hair seemed to encourage people to do a better job with sales. Even though her numbers weren't high she drastically increased the days’ sales.“ Such insightful comments correlated in the exploratory study, for instance, with judging the altruist to be of highest utility for the company (r = .347, p < .001).

Nonetheless, the altruist was evaluated most negatively by the vast majority of participants. This exploratory study suggests that most participants, at least after 10 or 20 rounds, clearly tend to evaluate workers in a T-PET, not based on the overall strong correlation between presence of a worker and overall team earnings, but mostly rather based on their individual earnings alone.

Overview of experiments

The experiments that we report on more fully here explore further this potential neglect of taking the overall utility of a worker into account, and more generally how people in the role of personnel manager deal with conflicting number-based evidence on the individual and group level.

Since the preliminary study showed that participants quickly base their judgments on individual and not on overall payoffs, Experiment 1a used a higher number of observations and introduces repeated test phases to explore whether this phenomenon is stable over time. Moreover, it extends the scenario to the wider field of personnel management by adding a two-level personnel *selection* task (here treated as a special kind of T-PET), requiring people to select a team based on available information on two levels. This task should at least normatively be even more clearly responsive to an employee's overall effect on his or her team, since it seems reasonable for a personnel manager to select those teams that performed best in the past (which should be easy to identify, cf. Experiment 2). In addition, we also increased the number of measurement blocks from one to four. This should further draw people's focus to the global level. Finally, we introduced conditions varying the performance levels of the normal non-altruistic workers to see whether participants can detect relatively small individual differences.

Experiment 1b considers whether the same phenomena occur for *egoist* detection as for *altruist* detection, while again checking the sensitivity to smaller individual effects.

Finally, Experiment 2 investigates whether and how quickly participants detect group-level effects of employees if the focus is on global information alone, with no conflicting information on individual payoffs. It checks whether people are cognitively able to discern quickly which worker has the highest overall group impact on group information alone. Additionally, we look at higher numbers of shifts.

Experiment 1a – Altruist Detection in Personnel Selection as well as Personal Evaluation Tasks?

Experiment 1 again used a personnel evaluation scenario, building on the large-impact-of-the-altruist conditions from the above-mentioned preliminary study, again involving a single interactor only, for instance the 'altruist,' who was the lowest individual earner but overall contributed most to the success of the group. Experiment 1 addressed the following main issues:

- *Personnel selection*. We used a personnel *selection* task, where the manager determined an optimal configuration of selected persons for a shift. This may focus participants more clearly on the overall earnings of a group. We aimed to explore our prediction whether or how far one can show the Tragedy-of-Personnel-Evaluation results in such a situation, or whether people may perhaps even begin to ignore the individual in favor of the group level. In the now shown

40 shifts we additionally measured the outcomes repeatedly (four times), thus also involving repeated selection tasks. Particularly if done repeatedly this may lead to an easier detection of the altruist's highest overall contribution.

- *Sensitivity to distinguishing individual earnings*. We varied the individual earnings of the 'normal' workers (whose presence did not interact with other workers' performances) to allow checking whether participants tend to neglect individual differences. We used individual effects that were smaller than the overall group-utility of the altruist. Additionally, we used three kinds of individual differences between normal workers (non-interactors) to explore whether this had an effect.
- *Altruist detection versus egoist detection.* Whereas Experiment 1a is concerned with positive interactors or altruist detection, Experiment 1b is concerned with egoist detection (see Experiment 1b for more details).
- *Selection of participants*. We aimed to rule out effects of unmotivated participants from MTURK. Therefore we selected participants by attentiveness from the beginning.

Design

Table 1 shows the average earnings of normal workers and of the altruist in the four conditions (C1 to C4). In C1 we used the same payoff structure as in the conditions in the preliminary study that were characterized by a large impact of the altruist on the group's performance. Here all normal (non-interacting) workers did not differ in their individual average earnings. In C2 now, one individual normal worker (N1) stands out with a higher individual earning. In C3, both normal workers N1 and N2 differ from both N3 and N4, and in C4 all four normal workers differ from each other in average earnings.

Table 2. *Mean earnings of normal workers (NW: N1 to N4), the altruist worker (A), the overall groups, with and without the altruist in the four conditions, and the resulting predictions if focused on workers' individual or overall (including indirect) impact on the group earnings in Experiment 1a.*

			C1	C2	C3	C4
Workers' earnings	NW without atruist	N1	2000€	2300€	2400€	2600€
		N2	2000€	1900€	2400€	2200€
		N3	2000€	1900€	1600€	1800€
		N4	2000€	1900€	1600€	1400€
	NW with altruist	N1	3000€	3300€	3400€	3600€
		N2	3000€	2900€	3400€	3200€
		N3	3000€	2900€	2600€	2800€
		N4	3000€	2900€	2600€	2400€
	altruist	A	1600€	1600€	1600€	1600€
Teams' earnings	Without altruist		8000€	8000€	8000€	8000€
	With altruist		10600€	10600€	10600€	10600€
Predictions	Individual		N1=N2=N3=N4>A	N1>N2=N3=N4>A	N1=N2>N3=N4>A	N1>N2>N3>N4>A
	Overall		A>N1=N2=N3=N4	A>N1>N2=N3=N4	A>N1=N2>N3=N4	A>N1>N2>N3>N4

In all conditions of Experiment 1a, we only have one positive interactor, the 'altruist.' The presence of the altruist modifies the outcome of the other workers' mean earnings; and although individually the lowest contributor, he or she clearly affects most positively the overall earnings of the groups. This results in a dissociation of individual and group level earnings. Individually the altruist has the lowest, whereas for the overall team the altruist has the most positive impact. Note also that the summed earnings of the four normal workers were kept constant over the conditions.

Method

Participants and selection criteria. Based on passing a participation-criterion (time spent on the first page > 20 sec. and < 6 min.), 156 people began the task.[2] Subsequently, 140 people (90%) passed a further selection criterion – correctly choosing (from four options) the task description: "You are instructed to evaluate the performance of the different employees for optimizing the companies' overall earnings by analyzing data provided by the reporting unit." Of these 140, 120 (86 %) finished the experiment. We worked solely with this selected pool of participants in order to ensure relatively motivated participants. Of the participants, 52% were male, 48% were female, the mean age was 33 years, and 59% mentioned having a Bachelor's or Master's degree and 38% a high school degree as their highest level of education. The participants obtained a reward of $1. Participants were randomly assigned to one of four experimental conditions (cf. Table 2).

Material and procedure. The materials were nearly identical and we used a similar procedure as the above-mentioned preliminary study (cf. Figure 1). We were again concerned with shifts of four workers out of potentially five. In each round, participants obtained information about the individual and overall payoffs of a shift. Additionally, we changed the payoff structure, to investigate how far participants distinguished relatively small individual differences and whether these latter might affect the detection of the larger impact of the altruist (cf. Table 2). Moreover, participants now obtained information on 40 working days, with each day displayed on a single page, and repeated test phases.

[2] We do not know the rate of people not fulfilling the first criterion because the recorded numbers also included people who after initial information decided not to participate.

We now required participants to remain on each page for at least four seconds. Only then could the 'continue' button be pressed. Again, participants could stay on these pages as long as they wished. In this experiment the altruist was working 50% of the days.

The workers now had to be evaluated by the participants (as managers) after Rounds 10, 20, 30 and 40 (four test phases, repeated T-PET). In the first three test phases, we presented one evaluation task (the rating task) and the additional personnel selection task. In the rating task, participants rated each worker's contribution on a scale of 1 to 10 ("Please rate contribution of the person shown above"). In the personnel selection task, participants were asked which four from the five employees they would select to work in a hypothetical further shift. Participants were told that all five employees would like to work and that their choice must optimize the profit for the company on that day. In the fourth test phase, we used all the evaluation measures from previous test phases (rating task and selection task) and some further tests in the order: the rating task, a maximal and minimal utility task, a ranking task, and the selection task. In the highest and lowest utility task[s? see above] the question read: "Which person has the greatest [or lowest] total utility for the company". The ranking task was formulated: "Please rank the order of employees with regard to their total utility for the company in the present situation".[3]

Finally, we used (a) one item of a Kimchi-Palmer task (Kimchi & Palmer, 1982, similar to a Navon task; Navon, 1977), assessing global versus local perception preferences by asking whether nine (3 x 3) squares look more similar to nine (3 x 3) triangles (global similarity) or to 4 x 4 squares (local similarity). In addition we used (b) a miniature attention test;[4] and (c) participants supplied

[3] We will omit analyzing the ranking task, here and in the other experiments, since it seems redundant with the other measures.

[4] Under the heading 'Height' participants could choose an appropriate height from a drop-down list, but the text under the headlines

demographic data and comments on the task. The comments on what they had learned in the experiment (provided by almost all participants) were coded as ‘insightful’ if and only if participants mentioned any kind of group-level effects in contrast to only individual effects, or interactions of one (or more) employee(s) with at least one other employee (i.e., having any effect on others). Thus ‘Insight’ was coded independently of whether the interactor mentioned was in fact doing the “interacting.”[5]

Results

Figure 2 presents the results of the rating task over the four test phases. The descriptive results are: (a) The average rating of the altruist in all conditions was lower than the ratings of the normal workers. Thus the average ratings even in the last phase seem quite clearly to reflect the individual contributions (individual impact) rather than the overall contributions of the workers (overall team impact). (b) The average ratings were sensitive to individual *differences* that were smaller than the overall impact of the altruist. (c) The ratings for the altruist in all conditions and phases were actually lowest *from the first test phase* onward. Thus the ratings appear quite stable over time.

instructed them to use the ‘not-specified’ option as a test. We do not consider it irrational to overlook such a text, given the context. Nonetheless, this easily assessed item could serve as a simple indicator of participants’ attentiveness or tendency to process information in a detailed, “bottom-up” way.

[5] First one experimenter coded each comment; then another one checked this. Differences were resolved accordingly.

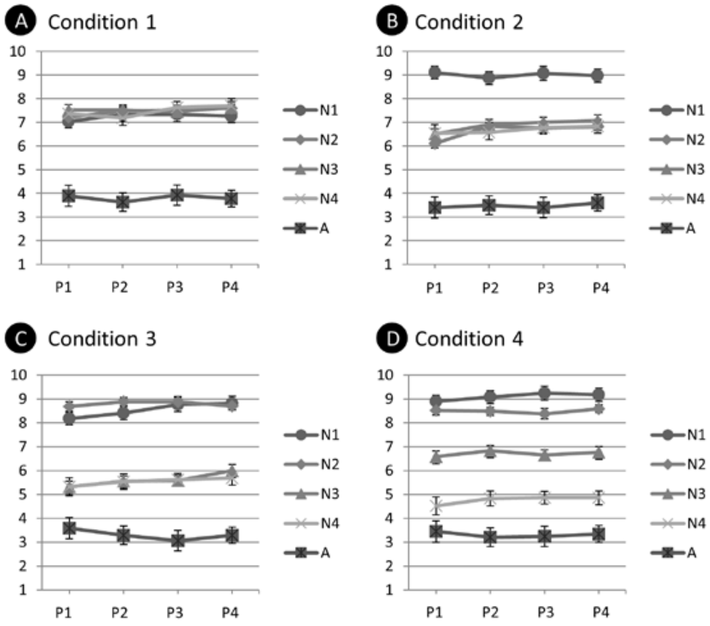


Figure 2. Average ratings (with SE) in Experiment 1a for the four normal workers (*N*) and altruist workers (*A*) in the test phases P1 to P4 of Conditions 1, 2, 3, and 4 (Panels A to D).

Correspondingly, inferential statistics, using a global analysis of variance (ANOVA), with Phases and Workers as within-subject factor and Conditions as between-subject factor, show reliable effects of Workers (PST: due to violations of sphericity we again used Pillai-Spur Tests, PST, *F*(4, 112) = 121.19, *p* < .001), Workers × Conditions (PST, *F*(12, 342) = 27.69, *p* < .001), but also of Phases (PST, *F*(3, 113) = 6.01, *p* = .009). Phases × Workers only approached significance (PST, *F*(12, 104) = 1.67, *p* = .08).

Even in the empirically most critical condition (C4), the mean ratings in Phase 4 are obviously at odds with the overall-level prediction based on the worker's *overall* contribution (A > N1 > N2 > N3 > N4; cf. Table 2). As speculated, contrasts show significant results for all five predicted mean differences, based on *individual* contributions (N1 > N2 > N3 > N4 > A): N1 > N2: *F*(1, 28) = 5.47; *p* < .027; N2 > N3: *F*(1, 28) = 39.8; *p* < .001; N3 > N4: *F*(1, 28) = 34.49; *p* < .001; N4 > A: *F*(1, 28) = 37.96; *p*

< 001.

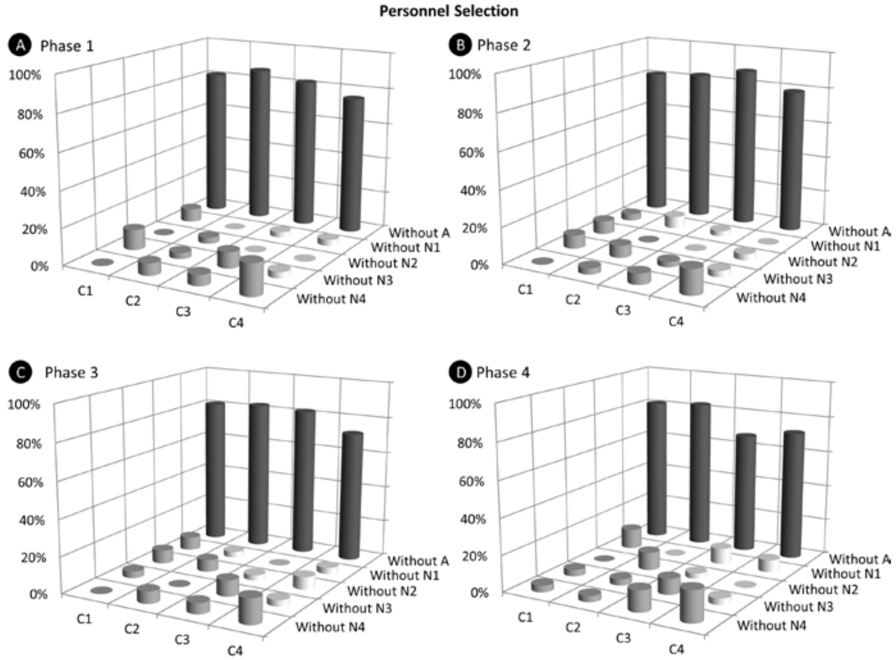


Figure 3. Results of the personnel selection task in the four test phases of Experiment 1a, depicting the proportion of ‘managers’ choosing a team of four from five candidates, thus excluding either worker N1, N2, N3, N4, or the altruist worker A. Optimal individual-related selections are black, optimal team-related selections are medium gray, and other selections are light gray.

The results of the personnel-selection task are shown in Figure 3. We used different levels of gray to mark individually optimal exclusions of the altruist (black), overall optimal team-related selections (medium gray), and other selections (light gray). The personnel-selection task could have yielded better results, since the selection task by definition seems to focus on the group level and explicitly emphasized *overall* earnings. Participants performing repeated selection tasks could thus realize the consistently and clearly lower outcome of teams without the altruist, relative to the only other four team-configurations. But Figure 3 shows that, even in this task, participants tended to exclude the altruist as the overall best worker of the team and to choose the candidate with the only individually optimal performance (black instead of medium gray selections). Again this was quite stable over time. Even in the final Round 4, the altruist was excluded clearly and highly reliably more often than *all* other workers put together, $\chi^2(1, N = 120) = 32.03, p < .001$). This was also the case for all single conditions, even the most critical C4 ($\chi^2(1, N = 29) = 5.83, p = .016$). Among those who did choose the altruist for the team, one can distinguish (at least in C2 to C4) those who additionally chose individually optimal normal workers (medium gray), or not (light gray); and in this subgroup the individually optimal normal workers were overall selected significantly more often, $\chi^2(1, N = 24) = 4.17, p = .041$).

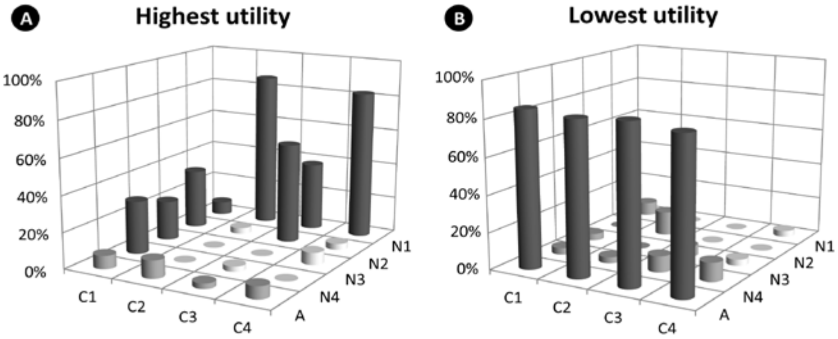


Figure 4. Percentage of 'managers' choosing either a normal worker (*N*) or the altruist worker (*A*) as being of highest (Panel A) or lowest (Panel B) utility for the company (Conditions C1, C2, C3, C4). The choices based directly on individual earnings are marked in black, those based on the overall earnings in medium gray, and the others in light gray.

Figure 4 presents the proportions of workers selected to be of highest or lowest '*total* utility' for the company in the final test phase (highest/lowest utility task). Here again, in all conditions a clear majority did not assign the *highest* utility based on the overall impact of the earnings of the group (medium gray), but instead favour the individual earnings (black) reliably over all others (C4: $\chi^2(1, N = 120) = 161.22, p < .001$). In all conditions, the majority of participants ascribed the *lowest* utility to the ‘altruist’ (Figure 4B). This was the case even in the empirically least clear condition (C4), where the black choices were still significantly more frequent than all other choices taken together, $\chi^2(1, N = 29) = 12.45$, $p < .001$.

With regard to the few selections that were not *individually* optimal, conditions C2 to C4 additionally allowed for testing whether the overall-optimal altruist selections in the highest utility task occurred at least with a higher relative frequency than the selections of other normal workers (exact binomial test, $N = 10$, $p = .011$). Additionally, the lowest utility task in C2 to C4 shows that among those who did not ascribe the lowest utility to the altruist, participants were sensitive in distinguishing individual earnings of normal workers. The lowest-utility assignments were more frequent for normal workers with lowest individual earnings (medium gray) than for normal workers with higher individual earnings (light gray) ($\chi^2(1, N = 15) = 8.97$, $p = .005$).

Participants’ comments quite clearly show that at least some grasped the high (or highest) utility of the altruist. In Experiment 2, ten comments by participants (8%) were coded as explicitly mentioning interactions between workers’ performances or differences between workers’ individual and overall performances. Several describe this discovery as a personal insight. Example 1: “What I learned is that I should look beyond my initial impression and study the data carefully. My first impression was that the female with the glasses was not contributing to the earnings of the company. However, after further analysis I realized that the highest earning days were when she was working. This led me to conclude that even though her sales might not be as much as the others, she was contributing in ways that increased the sales of the other employees.” Example 2: “At first, I was so focused on the employees' individual earnings that I viewed it as more of an indicator of who brought the most to the business... but then I started noticing a pattern: every time the man in the suit with checkered shirt was missing, total earnings plummeted. His individual earnings was the lowest, but the other workers always earned *more* when he was there than when he wasn't (causing total earnings to exceed other days). It was almost as if he added to their morale, or had a personality that bettered the working environment (i.e., encouraging, etc.). That is why I chose him as the highest utility, despite his low(er) individual earnings.” Example 3: “Well, I learned that it wasn't all about individual profits but sometimes something deeper. I realized that a certain individual in the group made the others work better when grouped with them. While this individual didn't earn much by himself, together with the others his group earned the most profit. So I think sometimes you have to look past the surface and delve deeper if you want to truly understand some things.” Overall, of the ten participants with insightful comments, all (100%) selected the altruist to be on the team in the personnel selection task in Phase 4 (cf. Figure 4), whereas of the participants without insightful comments, only 17% made this team selection ($r_\varphi = .53$; exact Fisher test, $p < .001$). Ninety percent of participants with insightful comments had already made this selection in Round 3 ($r_\varphi = .53$, $p < .001$), and 80 % in Round 2 ($r_\varphi = .51$, $p < .001$), but only 30% in Round 1 ($r_\varphi = .09$, $p = .388$). Additionally, insightful comments correlated with rating the altruist higher than *at least one* normal worker (Round 4, exact Fisher tests, $r_\varphi = .48$, $p < .001$; Round 3, $r_\varphi = .26$, $p < .001$; Round 2, $r_\varphi = .19$, $p = .077$; Round 1, $r_\varphi = .03$, $p = .658$), and with a higher rating than *all* normal workers (Round 4, $r_\varphi = .53$, $p < .001$; Round 3, $r_\varphi = .54$, $p < .001$; Round 2, $r_\varphi = .43$, $p = .006$; Round 1, $r_\varphi = .24$, $p = .054$).

With regard to the additional tests, in the Kimchi-Palmer task global answers (overall 30%) did not significantly correlate with insightful comments ($r_\varphi = –.09$, $p = .722$). As to the attention task (with 72% correct answers), all participants with ‘insightful’ comments correctly solved this (100%), whereas only 70% of the other participants did. If tested two-tailed, this correlation only approximated being significant ($r_\varphi = .19$, $p = .060$).

Discussion

Experiment 1a corroborates the idea that there may be a potential "tragedy of personnel evaluation." First, in Experiment 1a participants seem mostly to ignore the extremely positive effect of the altruist on a shift's overall performance even in the fourth test phase. The utility ratings seemed mostly to reflect the workers' individual but not overall performance. This holds for the numerical rating tasks as well as for the highest and lowest overall tasks. One might object that the question of the rating tasks was formulated relatively openly; it was not explicitly stated whether one should base judgments on the direct *individual* utility contribution rather than on the *overall* contribution. Nonetheless, it remains significant that participants' judgments tended to be based on individual utility alone. Moreover, we obtained similar results for the highest and lowest utility task as well, where participants were explicitly concerned with the workers' "*total* utility for your business." This strongly suggests that most people did not realise the great overall group effect of the altruist.

Second, in the repeated personnel selection tasks participants should assemble teams that are most profitable for their company. From their very nature, these team selection tasks focus on overall team performance. Note that the task is quite simple since there were only five configurations (cf. also Experiment 2). Nonetheless, most participants excluded the best team player (the altruist), even though the shifts without the altruist were clearly much less profitable.

Third, despite the lack of sensitivity regarding the overall utility of workers, the results show that people were actually quite sensitive to relatively small differences at the individual level.

Finally, there was a subgroup of participants whose comments revealed insight into the two dissociated levels of earnings (individual and group levels); and it was shown that this significantly correlated with selecting the altruist for the team and, to perhaps a slightly lesser degree, with evaluating the altruist more positively in the ratings tasks.

Overall, the results corroborate that participants in the role of personnel managers may systematically tend to evaluate the altruist with highest overall utility to be of lowest utility. Additionally, they mostly excluded this ‘altruist’ from the team.

Experiment 1b – Does ‘Egoist’ Detection in Personnel Evaluation Lead to Similar Results?

Whereas Experiment 1a investigated the evaluation of a *positively* interacting worker, Experiment 1b now concerns the analogous case of a *negatively* interacting worker, called the ‘egoist’. It is well known in organisational psychology that negative members can disproportionally affect the productivity of a whole team, for instance by spawning dysfunctional group processes (Felps, Mitchell, & Byington, 2006). Again the ‘egoist’ is only behaviourally defined as consistently having the *highest individual* earnings in the team while affecting the overall team performance most negatively.

Otherwise, Experiment 1b is almost identical to Experiment 1a. In this study, participants were again acting as personnel managers, repeatedly making personnel evaluations and selections concerning employees of a snack bar (working in teams of four out of five potential workers). Experiment 1b explores whether the tragedy of personnel evaluation is unique to altruist detection, or whether an analogous phenomenon exhibits itself for egoist detection as well.

From the viewpoint of a related debate on deontic rule testing, this might be doubted, since it has been claimed that people detect cheaters, but not altruists or co-operators (Cosmides, 1989; Gigerenzer & Hug, 1992). However, this strand of research, mainly concerned with Wason’s selection task (WST), has arguably mainly highlighted the differences between the domains of deontic rules and descriptive hypotheses (Oaksford & Chater, 1994; Fodor, 2000; Beller, 2010; von Sydow, 2006), with deontic reasoning reflecting a fairly systematic deontic faculty (Beller, 2010; Bucciarelli & Johnson-Laird, 2005; von Sydow, 2006), despite potential sub-classes within the deontic domain (Fiddick, Cosmides, & Tooby, 2000). Moreover, the putative faculty for cheater-detection may depend upon specific goals (von Sydow, 2006; von Sydow & Hagmayer, 2006; cf. Rand, Dreber, Ellingsen, Fudenberg, & Nowak, 2009; Sperber & Girotto, 2002), and perhaps the WST paradigm is too restrictive to investigate these phenomena (Sperber & Girotto, 2002). The WST question does not settle the more general question of how readily people deal with dissociations of individual-level and group-level performance or how they manage to detect pro-social or altruistic behaviour in general. Research in repeated public-good games suggests that the option of mutual reward may be as effective as (or even more so than) mutual punishment to increase payoffs (Kiyonari & Barclay, 2008; Rand, Dreber, Ellingsen, Fudenberg, & Nowak, 2009). However, the T-PETs investigated here address a different issue, investigating whether a third-party player can detect the dissociated individual and team effects of an ‘egoistic’ player. Do people ignore the group effects of the egoistic interactor as they did for the altruistic interactor? Or, inversely do they perhaps ignore the individual-level information? Despite differences to the WST-debate, such findings may shed light on more general issues raised in that debate.

Design

Table 3 shows the average earnings of the negative interactor, the egoist (*E*), and, depending on the latter’s presence or absence, the average earnings of the normal workers in the four conditions. The conditions, as in Experiment 1a, vary the homogeneous and heterogeneous earnings for the normal workers (C1 most homogeneous; C4 most heterogeneous) to investigate participants’ sensitivity to small differences on the individual level – much smaller than the egoist’s overall effect on the team’s performance. In all conditions the presence of the individually most successful egoist leads to a substantial decline in the team’s overall performance.

Table 3

Mean earnings of normal workers (NW: N1 to N4) and of the negatively interacting ‘egoist’ worker (E), as well as the overall earnings with or without the egoist in the four conditions; and resulting predictions if based on a worker’s individual or the overall performance in Experiment 1b.

			C1	C2	C3	C4
Workers’ earnings	NW without egoist	N1	3000€	3300€	3400€	3600€
		N2	3000€	2900€	3400€	3200€
		N3	3000€	2900€	2600€	2800€
		N4	3000€	2900€	2600€	2400€
	NW with egoist	N1	2000€	2300€	2400€	2600€
		N2	2000€	1900€	2400€	2200€
		N3	2000€	1900€	1600€	1800€
		N4	2000€	1900€	1600€	1400€
	Egoist	E	3400€	3400€	3400€	3400€
Teams’ earnings	Without egoist		12000€	12000€	12000€	12000€
	With egoist		9400€	9400€	9400€	9400€
Predictions	Individual		E>N1=N2 =N3=N4	E>N1>N2 =N3=N4	E>N1=N2 >N3=N4	E>N1>N2 >N3>N4
	Overall		N1=N2=N3=N4>E	N1>N2=N3=N4>E	N1=N2>N3=N4>E	N1>N2>N3>N4>E

To keep most of the payoff structure in Table 3 similar to Experiment 1a (cf. Table 2), we simply exchanged the earnings of the normal workers in the with-interactor context with those in the without-interactor context. Thus again, the overall earnings of all workers together remained constant over conditions. Moreover, the egoist's individual earnings in C1 increased (relative to an average normal worker without egoist) as much as the altruist's individual earnings had decreased (relative to a normal worker without altruist). Thus the predictions of Experiment 1b are reversed (relative to Experiment 1a): If the judgments are based on individual performance alone, it is now expected that the interactor – here the egoist – should be rated highest (not lowest, as with the altruist); and if they are based on overall impact, the egoist should now obtain the lowest ratings (rather than the highest ratings, as with the altruist).

Method

Participants. The first participation-criterion (again the time spent on the first page > 20 sec. and < 6 min.) yielded 161 participants from MTURK. Without advanced notice, participants were then required to rephrase the described task; 151 people passed this second test. Of these, 128 (85 %) finished the experiment. Additionally we excluded the data of eight participants who had previously taken part in a similar experiment. Thus, as in Experiment 1a, we analyzed data of 120 participants (52% male, 48% female; mean age 33), most of them with a high school or even an academic degree (59% Bachelor's or Master's; 38% high school). The participants obtained rewards of $1 for participation. They were randomly assigned to one of four conditions (cf. Table 2).

Material and procedure. Apart from the payoff structure (Tables 2 and 3), the scenario, procedure, and dependent variables used in Experiment 1b were identical to Experiment 1a (excluding minor corrections of typographical errors). In each round, participants obtained overview information in tables about workers' individual earnings, together with their photographs and information about overall earnings of the team (Figure 1). Again there were only five workers, with four workers per shift and five possible team configurations. The first three test phases included only rating tasks and a team selection task; the final round additionally involved the maximal and minimal utility task and further tests mentioned in the methods section of Experiment 1a.

Results

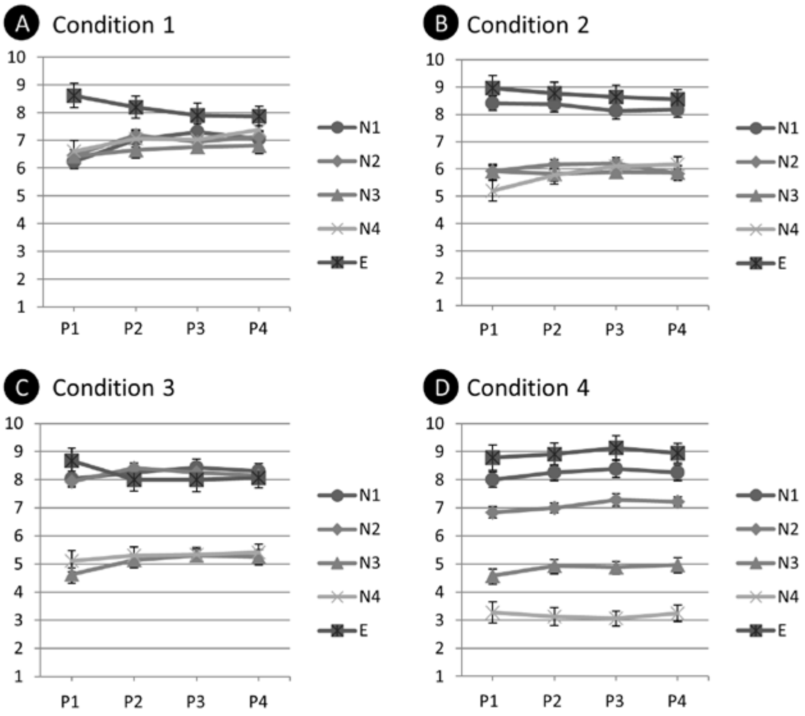


Figure 5. Average ratings (with SE) in Experiment 1b for the four normal workers (*N*) and egoist worker (*E*) in the test phases P1 to P4 of Conditions 1, 2, 3, and 4 (Panels A to D)

Figure 5 shows the average utility ratings for all snack bar workers in the four conditions. In all conditions the average ratings resemble more closely the predictions based on individual rather than overall team-contributions (cf. Table 3). A global mixed ANOVA with the between-subjects factor Conditions and the within-subjects factors Workers and Phases (in a multivariate Pillai-Spur Test, PST) showed significant effects of Workers, *F*(4, 110)=95.86, *p* < .001, Workers × Conditions, *F*(12, 336)=21.94, *p* < .001, and Phases, *F*(3, 111) = 3.41, *p* = .020. The effects of Phase × Person, *F*(12, 102) =1.78, *p* = .061, did not reach significance.

To obtain more specific insights we analyzed each condition. In Condition 1, a repeated-measurement ANOVA, with Workers and Phases as factors, shows effects of Workers, PST, *F*(4, 26) = 4.19, *p* = .001, with the egoist on average having higher ratings than the other four workers throughout all phases. Did the repeated measurement factor Phase (i.e., change over time) play a role? Despite a descriptive decline in the difference between the egoist and the normal workers throughout the test phases, the factor Phase and the interaction Workers × Phase only approached significance (*p* = .08, *p* = .12). In Condition 2, there was again only a reliable effect of Workers, PST, *F*(4, 24) = 25.20, *p* < .001, with no other statistically fully significant effect of Phase or Workers × Phase (*p* = .73, *p* = .07). Although the average rating of the egoist was descriptively higher than the average rating of all other workers, the egoist had similar ratings to the normal worker predicted to be highest. Bonferroni-corrected post hoc comparisons showed no differences between the egoist and normal worker 1 (E, N1) (*p* = 1.00); but N1, as predicted, also had a higher rating than the other normal workers, which also was the case for the egoist (all comparisons, *p* < .001). In Condition 3, the factor Workers was also significant (PST, *F*(4, 22) = 24.10, *p* < .001), but no other factors were (*p* = .79, *p* = .59). Bonferroni-corrected post hoc comparisons showed that the egoist was not rated higher than the predicted individually good normal workers (N1, N2) (both *p* = 1.00), but rather that the worker in this group, as well as the egoist, clearly and reliably differed from those in the second group of normal workers (N3, N4; all *p* < .001). Condition 4 likewise shows only a significant effect of Worker (PST, *F*(4, 28) = 58.61, *p* < .001), but no other significant effect (*p* = .171, *p* = .845). Bonferroni comparisons here showed significant effects of all five workers in the order predicted by individual earnings only (all *p* < .01).

Overall, the main pattern of ratings-task results largely corresponds with the order predicted by individual earnings. It is clearly at odds with the order that would follow from the overall impact of workers on the group level (cf. Table 3). Although the ratings for the egoist were often higher than those of the normal workers (and never significantly lower than any of them), the difference at least regarding the most efficient normal worker (or group of normal workers) was not always large or statistically significant. This seems to suggest a somewhat increased (but not high) influence of group-level predictions.

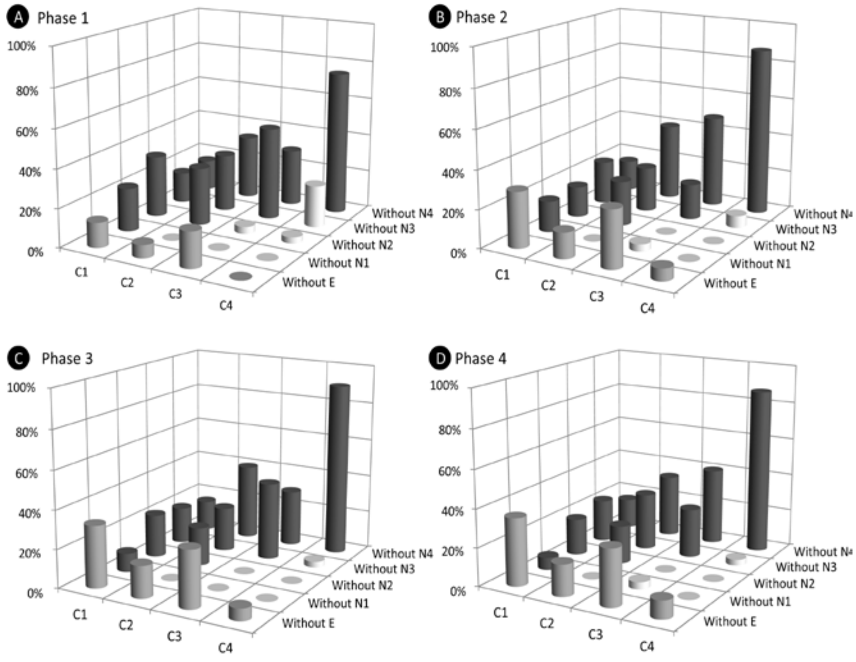


Figure 6. Results of the personnel selection task in the four test phases of Experiment 1b, showing the proportion of 'managers' choosing a team of four out of five, thus excluding worker *N1*, *N2*, *N3*, *N4*, or the egoist worker *E*. Individual-related optimal selections are marked in black, with team-related optimal selections in medium gray and other selections in light gray.

In the repeated personnel selection tasks the main question is whether the ‘managers’ included the egoist on the team (who consistently had the lowest team earnings). The results, shown in Figure 6, show that the majority selected teams with optimal earnings on the individual level (black; individual-related selections). Only a few selected the team configuration *without* the egoist (from five possible team configurations), although this team had clearly and consistently the best overall performance (medium gray; team-related selections). The remaining selections (light gray) selected the egoist for the team, along with other, individually non-optimal workers.

Even in the final test phase, Phase 4, the individual-related selections (black) in all conditions together occurred more frequently than the team-related ones (medium gray), $\chi^2(1, N =118) = 34.71$, $p < .001$. In contrast to Experiment 1a, in C1 to C3 this might be due to the higher number of classes of individual-related rather than team-related classes. However, the black selections still occur more frequently than *all* other selections (both with an equal number of classes), $\chi^2(1, N = 120) = 32.03$, $p < .001$. Even in Condition 4, where the black and medium-gray selections are represented by one class only, the individual-related selections (black) have a higher relative frequency than all other selections together (Round 4, $\chi^2(1, N = 32) = 18.00$, $p < .001$). With regard to temporal changes, there seems to be an increase in the proportion of team-related selections medium gray) from Phase 1 (9%) to Phase 4 (23%); $\chi^2(1, N = 240) = 8.00$, $p = .005$. Additionally, in Phase 4 (in C2 to C4) the optimal team-related selections (medium gray) without egoist occur more often (relative to the number of categories) than the non-optimal selections with egoist (light gray), $\chi^2(1, N = 18) = 25.00$, $p < .001$, corroborating that these ‘deviations’ from the individual-related selections are not merely chance. Finally, clear differences between conditions likewise show participants’ sensitivity to even relatively small individual differences.

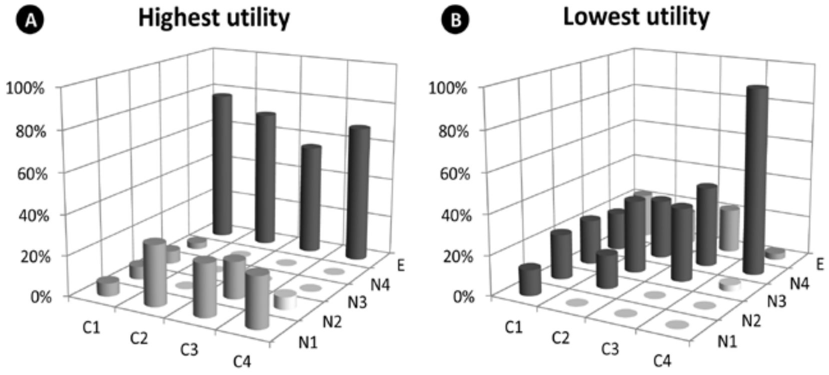


Figure 7. Percentage of ‘managers’ choosing either a normal worker (*N*) or the egoist worker (*E*) as of the highest (Panel A) or lowest (Panel B) utility for the company (Conditions C1, C2, C3, C4) in Experiment 1b. The individual-related choices are marked in black, the team-related choices in medium gray and the neither-individual-nor-team-related choice in light gray.

Figure 7 shows which workers in the final phase (Phase 4) were judged to have the highest or lowest ‘total utility’ for the company. The highest-utility task (Panel A) reveals relatively frequent ‘egoist’-judgments (black), even though the egoist’s presence was consistently correlated with lowest team earnings. In all conditions these judgments were clearly above chance level ($\chi^2(1, N = 31) = 63.88, p < .001$; $\chi^2(1, N = 30) =46.8, p < .001$; $\chi^2(1, N = 27) = 21.33, p < .001$; $\chi^2(1, N = 32) = 47.53, p < .001$). Looking at the *team*-related judgments (medium gray), they are also above chance level relative to the remaining judgments (neither optimal on group or individual level; light gray), $\chi^2(1, N = 31) =50.58, p < .001$ (C2 to C4).

In the lowest-utility task (Figure 7, Panel B), the individual-related selections (black) were chosen more often than chance-level, $\chi^2(1, N = 120) =175.21, p < .001$. In the three conditions (C2 to C4), where one can contrast the team-related judgments (egoist has overall the lowest utility; medium gray) with judgments that were neither individually nor on group-level optimal (light gray), the team-related judgments occurred reliably more often (exact binomial test, $N = 13, p < .001$).

With regard to the comments, now 21% of the participants mentioned explicitly that the individual- and group-level contributions of a worker differ, or that there are interactions between participants. Example 1: “I learned that while the woman in the striped shirt outperformed everyone individually, on the days she worked the overall profit went down. Because of this, she proved to actually be the least valuable member of the team, despite being the strongest individual. The best group to work in the shop would be the four employees other than the woman in the striped shirt.” Example 2: “It seemed to me that the brown haired guy was bringing down the productivity of the two female workers, the one with glasses and the one with the striped shirt.” Example 3: “Interesting to me that one person can dramatically outperform her peers but at the same time seemingly be the reason why the store's earnings are 75% of the earnings in which she isn't working.”

Eighty percent of the participants with such insightful comments in Phase 4 excluded the egoist in the personnel selection task from the team, whereas only 7% of the other participants did so (r_φ = .76, p < .001). This strongly suggests that the detection of interaction has a substantial effect on the selection task. Additionally, in earlier phases there were substantial and statistically reliable correlations of insightful comments with team-related selections: Phase 3, r_φ = .70, p < .001; Phase 2, r_φ = .64, p < .001; Phase 1, r_φ = .26, p < .003 (where some selections may have been due to chance). This suggests a high stability from Phase 2 onward. Furthermore, insightful comments correlated positively but again a bit weaker with ratings, coding positively for *at least one* normal worker rated above the egoist (Phase 4, r_φ = .45, p < .001; Phase 3, r_φ = .30, p = .002; Phase 2, r_φ = .25, p = .007; Phase 1, r_φ = –.08, p = .429) or for *all* normal workers rated higher than the egoist (Phase 4, r_φ = .58, p < .001; Phase 3, r_φ = .55, p < .001; Phase 2, r_φ = .31, p = .003; Phase 1, r_φ = .09, p = .377).

The additional tests did not reliably correlate with insightful comments. In the Kimshi-Palmer test, global answers (overall 31%) did not correlate with insightful comments (r_φ = –.08, p = .471); and in the attention task (with 52% correct answers) the positive correlation with insightful comments did not approach significance (r_φ = -.10, p = .377).

Discussion

The results of Experiment 1b show that egoist detection seems to be affected by similar problems as altruist detection. When the negative interactor (or ‘egoist’ for short) individually contributed the highest earnings but overall led to the lowest group earnings, the majority of participants nonetheless rated the egoist as most valuable. Moreover, in the personnel selection task they even systematically chose him for the team, even though this team consistently performed the worst. This was the case even though participants were quite aware of small differences at the individual level. Thus the results suggest a corresponding tragedy of personnel selection with regard to ‘egoists’, or negative interactors, despite the fact that participants only had to identify one global interactor. The phenomenon seems to affect both positive and negative interactions.

In comparison to Experiment 1a, Experiment 1b, despite its negative results, suggests a moderate advantage of egoist detection over altruist detection. In particular, the results of the rating tasks seem to be less dominated by individual-level judgments for egoist detection than for altruist detection.[6] Furthermore, there were significantly more insightful comments in Experiment 1b (egoist detection) than in Experiment 1a (altruist detection), $\chi^2(1, N = 240) = 7.53$, p = .006. This suggests that the content of egoist versus altruist detection leads to different interaction detection rates.

Although the direction of this effect is indeed in line with the idea of an advantage of egoist or

[6] Also the results of the highest utility task suggest that more people in Experiment 1b did not select the negative interactor (the egoist) to be of highest utility than selected the positive interactor (the altruist) in Experiment 1a; thus there seems an apparent increase of team-related judgments in the highest utility task. However, the basis of this comparison may be unfair, since for the highest utility task there were more group-related categories in Experiment 1b than in Experiment 1a.

cheater detection over altruist detection (cf. Cosmides, 1989; Gigerenzer & Hug, 1992), the small size of the effect supports the idea that there is in principle no (substantial) difference between these faculties (Oaksford & Chater, 1994; von Sydow, 2006; Beller, 2010; Bucciarelli & Johnson-Laird, 2005; Sperber & Girotto, 2002). In any case, the results prompt the question of why people in our setting have problems detecting *both* kinds of interactors despite the relatively strong group effects..

Nonetheless, the insightful comments made clear that at least a relevant minority was able to detect the difference between a worker's individual versus overall contributions to a group. This was associated with group-related answers in the personal evaluation and the personnel selection tasks, to some extent even in early test phases.

Experiment 2 – Optimal Evaluations and Selections Without Conflicting Information?

Experiment 2 aims to test a possible explanation for the tragedy of personnel evaluation observed in the preliminary study and Experiments 1a and 1b. One possibility is that (most) people are simply unable to detect group-level correlations given the shown number of days (the correlation between a worker's presence and the overall outcome at a group level). Alternatively, they may at least in principle be able to do so if forced to concentrate on this correlation only, without being distracted by (inconsistent) individual-level information. To test this we ran a version without individual-level information. Perhaps memorising and integrating information over time is too difficult even if the correlation is very high (r = .99, cf. preliminary study and Experiment 1) and though we found in Experiments 1a and 1b that participants were very well able to detect much smaller differences on the individual level. The resulting finding would in any case remain tragic, particularly since we worked with a relatively simple setting (only five workers). The other possibility would be that people are in principle able to see the correlation between an individual's presence and team performance over time, while for some reason fail to pay attention to the group level, focusing instead (almost) solely on the (conflicting) individual impacts. The evidence that up to one-fifth of the participants understood the two-level task is inconclusive on this point; it may either show that only these participants had an *ability* to see group-level effects in the given time (several group-level findings were above chance) or that they alone considered the group-level question.

After having considered the possibility of content effects of altruist vs. egoist detection, Experiment 2 elaborates on the issue of focusing people on the group level. We kept the initial story and again used personnel evaluation and selection tasks, here with *altruist* detection, in a sequential learning setting (T-PET, Experiment 1a). However, to test the the explanatory hypotheses, we investigated whether people were able to detect the group-level correlation if forced to focus on this level. Additionally, we again increased the number of rounds in conditions focusing on the group level as well as in other conditions.

Experiment 2 thus differs from Experiment 1 in the following two important aspects:

- *Global-information-only condition*: In these new conditions, participants obtained overall group-earnings, without the direct individual contributions of each worker. If people were simply incapable of learning the correlations between individual presence and group-level earnings, the effects should completely persist, whereas they should completely disappear if people are able to detect these correlations and in the local-and-global-condition do not focus on these conditions. For the global-only condition we at least expected that people would somewhat more often pick the altruist, since they are forced to concentrate on an individual's overall (i.e., group-level) effects, with no supplementary individual analysis. We aimed to explore whether this was the case, and whether most participants could discern the overall best worker under such conditions. If so, we wanted to see how many rounds were required.
- *Prolonged learning*: We increased the number of rounds (now 80 rounds with overview information, each round with no time limit). Even though the results of the previous

experiments suggest a slow increase of altruist detection, we looked for a point after which altruist detection increased dramatically. That is, in the only-global-information conditions more rounds may be needed to recognize which worker has the overall best outcome.

Design

The experiment had a mixed 2 (information: global-only vs. local-and-global) × 2 (earnings of normal workers: homogeneous versus heterogeneous) between-subjects design, with an additional within-subjects factor of four test phases. In each test phase both evaluation and personnel-selection tasks were applied. Additionally, in the last round, highest and lowest utility tasks and other tests were completed (as in Experiments 1a and 1b). We added a need-for-cognition (NFC) test, to explore the prediction that participants who detect the two levels of analysis (individual and overall impact) engage in cognitive activities with higher elaboration.

Table 4

Overview of the four conditions, also showing the overall vs. direct impact of a worker on the earnings of a group in Experiment 2 (for the detailed earnings cf. Experiment 1a, Table 2, C1 and C4).

Condition	C1	C2	C3	C4
Information	Local and global	Global only	Local and global	Global only
Earnings of NW	Homogeneous, cf. Table 2, C1		Heterogeneous, cf. Table 2, C4	
Overall impact	A >> NW1 = NW2 = NW3 = NW4		A >> NW1 > NW2 > NW3 > NW4	
Direct impact	NW1 = NW2 = NW3 = NW4 > A		NW1 > NW2 > NW3 > NW4 > A	

Note: NW = normal worker; A = altruist.

Table 4 illustrates four conditions. The local-and-global conditions use the same overview information as in each round of the previous experiments (Figure 1). Each round presents information on *both* the direct earnings of the four workers on a shift and the overall earnings of a given team. The overall earnings of course involve not only the direct earnings of individuals but also the overall earnings. In the global-only conditions, only the *overall* earnings of a group (shift) are presented, without showing individual contributions.

The homogeneous versus the heterogeneous conditions correspond to either identical or different individual impact of all workers (using the numbers of C1 and C4 of Experiment 1a; see Table 2). As before, the 'altruist' (*A*) in both conditions has the most positive impact on the overall earnings although individually the lowest direct impact. The impact of the normal workers (NW) does not differ in the homogeneous conditions but did differ in the heterogeneous condition (while keeping their mean contribution identical; Table 2). Participants were randomly assigned to one of the four resultant conditions.

Method

Participants. The same relatively strict selection criteria for participants were used as in Experiments 1a and 1b to ensure high data quality. The first participation criterion (time spent on the first page) was passed by 150 people. For the second criterion, 7 people failed (correct rephrasing of the task from four options). Of the remaining 143, 122 finished the experiment, and we again worked with those who finished the main task. The participants were recruited from MTURK: 57% male, 42% female; mean age 33, and 68% with a Bachelor's or Master's degree (32% had a high school degree). Participants obtained a reward of $2 for participating.

Material and procedure. We used almost identical materials to and a similar procedure as in Experiment 1a. The experiment had 80 rounds rather than 40, with four test phases administered after Rounds 20, 40, 60, and 80. In all four test phases again, both a personnel evaluation task and a personnel selection task were assessed. In the final test phase, we again administered an additional highest- and lowest-utility task, a ranking task, and kept the Kimchi-Palmer test-item, and the attention test. Finally, we added an 18-item version of the Need for Cognition Test (Cacioppo, Petty, & Kao, 1984).[7]

In the global-and-local conditions, the overview information tables presented in each round correspond to Figure 1. In the global-only conditions, the second line of this panel, presenting the individual earnings of each employee, was omitted in order to assess the potential of people to see overall differences if forced to concentrate on the overall (global) information. Thus participants in these conditions for each of the 80 days saw only pictures of the four workers (out of five) on the shift as well as their overall team-output. Again, only five team combinations were possible.

Results

Figure 8 shows mean ratings for the workers' contributions to the company earnings. An overall ANOVA with Workers (five workers) and Phases (four phases) as within-subjects factors, and Conditions as between-subject factor, yielded a highly significant effect of Conditions × Workers (Pillai-Spur Test, PST, $F(12, 306) = 22.52$, $p < .001$). A main reason for this expected interaction effect seems to be the change of rank in the altruist's ratings in the global-only versus global-and-local conditions. Furthermore, the Workers factor became significant (PST, $F(4, 100) = 17.34$, $p < .001$).[8]

[7] Each item was assessed on a symmetric five-level scale of agreement, with "extremely" (–2, 2) or "somewhat (–1, 1) uncharacteristic /characteristic of me," and "uncertain" (0) as labels. The scores of the items (–2, –1, 0, 1, 2) were summed up to calculate the overall test score.

[8] This may reflect that for all conditions (despite their considerable variation) the altruist was at least on average rated lower than some other workers (e.g., N1).

Phase × Condition as well as Phase × Workers approached significance-level (PST, $F(9, 309) = 1.69$, $p = .09$; PST, $F(12, 92) = 1.51$, $p = .133$), and potential changes over the Phases were not substantial.

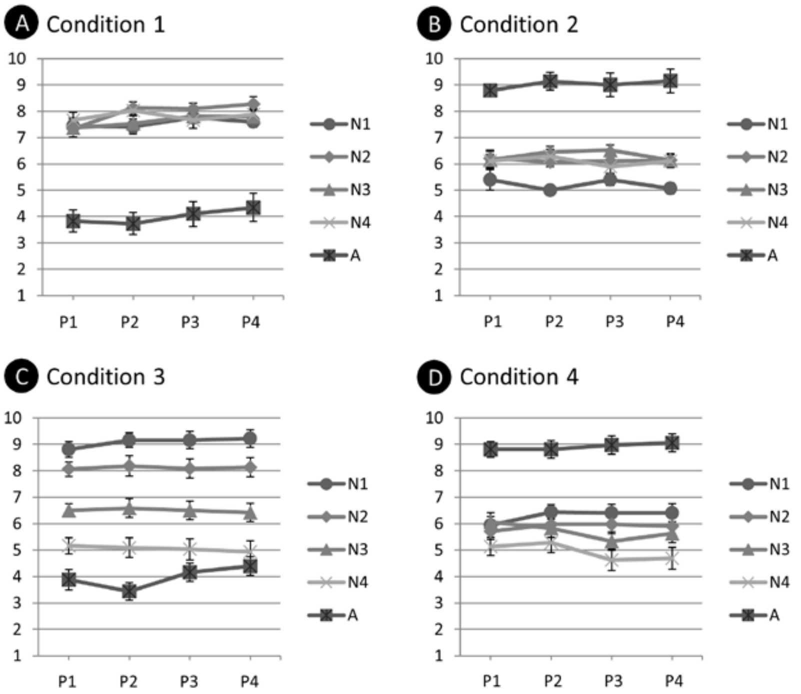


Figure 8. Average ratings (with SE) in Experiment 2 for the four normal (*N*) and altruist (*A*) workers in test phases P1 to P4 of Conditions 1, 2, 3, and 4 (Panels A to D).

The local-and-global conditions, both in the homogeneous Condition 1 and the heterogeneous Condition 3 (Panels A and C), broadly seem to replicate findings of Experiment 2 (Conditions 1 and 4), even with double the number of learning rounds. The altruist is still evaluated as lowest, even though most strongly correlated with high overall earnings. In the homogeneous Condition 1, the four normal workers were evaluated similarly, with each higher than the altruist. In an ANOVA for test phase 4, the within-subject factor Workers was clearly significant (PST, $F(4, 23) = 10.19$, $p < .001$), descriptively mainly referring to the lower rating of the altruist. Furthermore, individual contrasts with normal workers confirmed that all normal workers were rated higher than the altruist (all $p < .05$). Another

ANOVA showed that no significant differences between the normal workers remained when the altruist was removed (PST = .12, *F*(3, 25) = 1.16, *p* = .344). In the heterogeneous Condition 3, an ANOVA likewise showed a general effect of Workers (PST, *F*(4, 24) = 38.87, *p* < .001), but also, as expected, the four normal workers were rated differently, with significant contrasts in the predicted order N1 > N2 > N3 > N4 > A (each contrast, *p* < .001). Again, the ratings did not significantly change over time.

In the new global-only conditions (homogeneous Condition 2 and heterogeneous Condition 4), in which people were to base their ratings of a worker's utility on the teams' overall earnings alone, they clearly detected that, of all workers, the altruist correlated most demonstrably with high overall earnings. Although we had speculated that participants might grasp this later on, they in fact did so surprisingly early (even before the first test phase after round 20, corresponding to the insignificant results of the factor Phase in the omnibus ANOVA). The results, however, also suggest a negative effect of the exclusive focus on the overall contributions, since in the heterogeneous global-only condition (C4) participants were less able (than in C3) to distinguish (smaller) individual differences. In more detail, comparing the homogeneous conditions C2 to C1, the pattern of ratings is reversed and the results in Condition 2 are now at least roughly in line with the hypothesis: People induce that the altruist contributes more to the good of the company than the normal workers: A > NW1 = NW2 = NW3 = NW4. An ANOVA for Condition 2 (test phase 4) shows significant results for the factor Workers (PST, *F*(4, 22) = 23.04, *p* < .001); and pair-wise contrasts show that here the altruist is rated higher than all normal workers (always with *p* <.001). Although these were the largest effects (Figure 10), an ANOVA *without* the altruist also showed remaining differences between the normal workers (PST, *F*(3, 24) = 5.37, *p* = .006) (which clearly seems due to chance, since all normal workers had roughly the same earnings). In the heterogeneous Condition 4, the order of the average ratings of the altruist and the normal workers was likewise reversed (relative to the also heterogeneous Condition 3).

In an ANOVA a significant effect of the Workers factor was found in C4 (PST, *F*(4, 26) = 15.47, *p* < .001); and contrasts show that the altruist was rated significantly higher than even the normal worker, who was rated highest (*p* < .001). Although the mean ratings for the other normal workers, at least in Phase 4, are in the order that is optimal (NW1 > NW2 > NW3 > NW4), and an ANOVA without the altruist reaches significance (PST, *F*(4, 27) = 3.62, *p* = .02), only one Bonferroni-corrected post hoc comparison between normal workers (the one that is expected to differ most: NW1-NW4) led to significant results (*p* < .05). In sum, despite clearly detecting that the altruist has a greater effect on the overall output in the global-only conditions, participants show a reduced ability to distinguish between the normal workers.

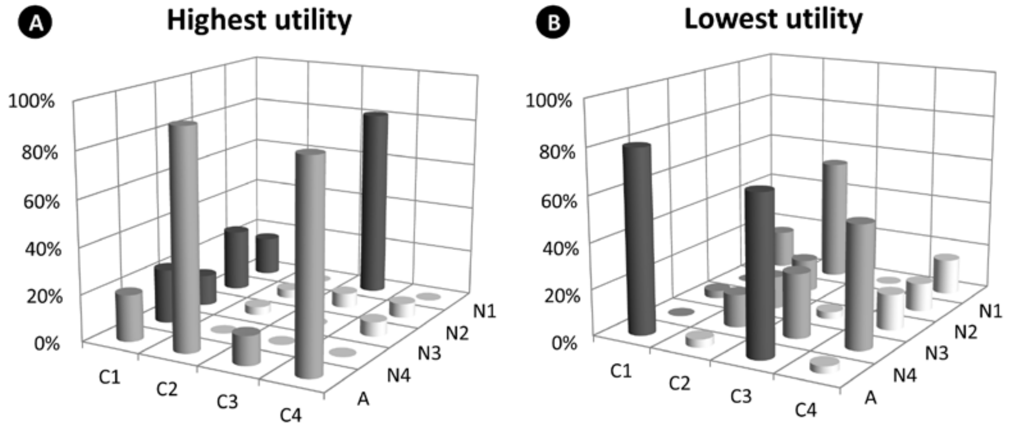


Figure 9. Percentage of participants choosing a normal worker (*N*) or the altruist worker (*A*) as overall of highest (Panel A) or lowest (Panel B) utility for the company in the final phase (after 80 rounds) in Conditions C1, C2, C3, C4 of Experiment 2. The choices that correspond to individual earnings are marked in black; those that correspond to overall earnings in medium gray.

Figure 9 shows the proportion of ‘managers’ choosing a particular worker to have the “highest” (Panel A) or “lowest” (Panel B) “total utility for the company” in the final test phase. The utility attributions differ systematically, particularly between the global-and-local conditions (C1 and C3) and the new global-only conditions (C2 and C4). In the homogeneous global-and-local Condition 1, even after 80 rounds a majority (83 %) selected a normal worker rather than the altruist as being of the highest utility (Panel A). This was not statistically significant (testing against the Null hypothesis of random selection), but this test cannot differentiate between random choice and systematic 4 to 1 preference of regular workers over the altruist. However, in the same condition a majority of participants (80 %) judged the altruist to be of the *lowest* utility for the company and this choice highly reliably differed from random selection, $\chi^2(1, N = 30) = 67.50$, $p < .001$ (Panel B). In homogeneous global-only Condition 2, by contrast, a majority of the participants (93%) selected the altruist rather than a normal worker to be of the highest utility ($\chi^2(1, N = 28) = 92.89$, $p < .001$). In the heterogeneous global-and-local Condition 3, a majority (81 %) judged the individually optimal normal worker N1 to be of the highest utility, with N1 being selected reliably most often ($\chi^2(1, N = 32) = 72.03$, $p < .001$). In contrast, the altruist, the overall most useful worker, is most frequently judged as of the lowest total utility for the company (69 %), $\chi^2(1, N = 32) = 47.53$, $p < .001$. In the heterogeneous global-only Condition 4, in contrast to Condition 3, a majority of participants (87 %) selected the altruist to be the most useful ($\chi^2(1, N = 32) = 91.13$, $p < .001$), and a majority (53 %) even selected the specific normal worker with the lowest individual utility (NW4) to be of the lowest overall utility to the company ($\chi^2(1, N = 32) = 21.95$, $p < .001$), suggesting at least some remaining sensitivity on the individual level.

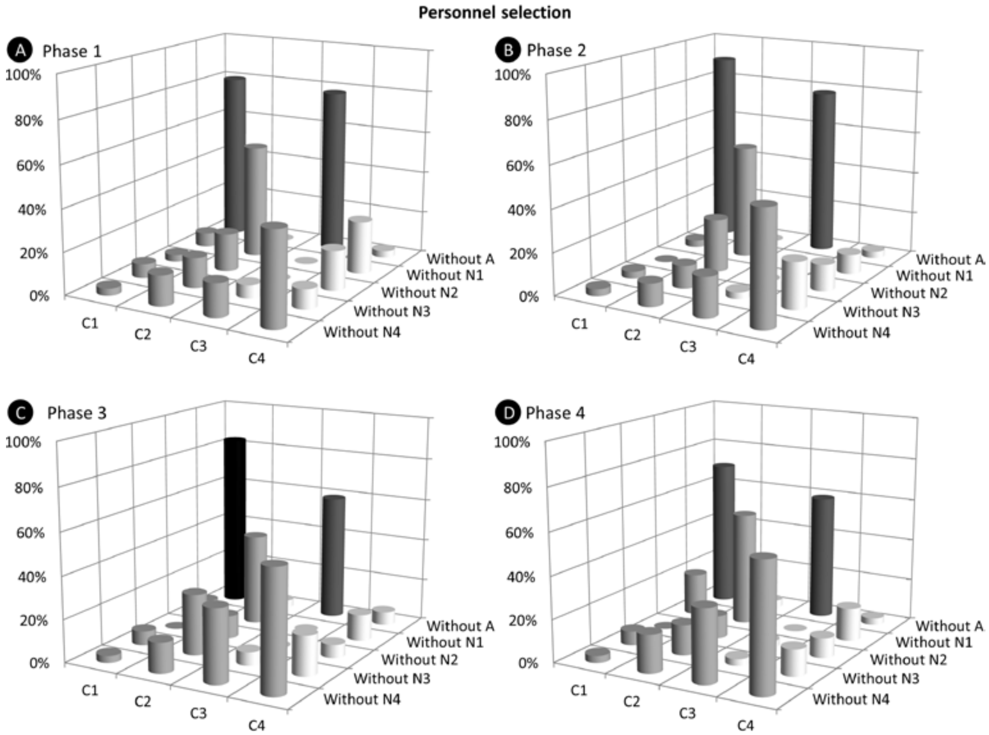


Figure 10. The results of the personnel selection task in the four test phases of Experiment 2 show the proportion of ‘managers’ choosing a particular team (four out of five), thus excluding either worker N1, N2, N3, N4, or altruist worker *A*. In the global-and-local conditions C1 and C3, the black columns correspond to the predicted selections based on *individual* performance only. In the global-only conditions, C2 and C4, no individual-level information was available. In all conditions, the medium gray columns represent the optimal selection(s) based on *overall* past performance of teams.

Finally, Figure 10 presents the team selections in the personnel selection task, a kind of task that most clearly emphasizes the importance of the team as a whole. The team earnings observed by the participants in all conditions were consistently the highest for teams with the altruist. In the homogeneous, local-and-global Condition 1, we nonetheless replicated a strong tendency to select the specific team *without* the altruist (from five possible configurations). Even in Phase 4, after 80 rounds, 70% of the participants selected this team, which is clearly above chance, $\chi^2(1, N = 30) = 46.88$, $p <$

.001. This selection was reduced from 80% in Phase 1 to 70% in Phase 4, but the change was not reliable, $\chi^2(1, N = 60) = .80$, $p = .371$. In contrast, the homogeneous global-only Condition 2 shows that participants provided with global-only information were highly capable of quickly detecting the altruist should be part of the team. As early as Phase 1, participants reliably selected one of the four with-altruist teams rather than the no-altruist team, $\chi^2(1, N = 28) = 7.00$, $p < .001$. Throughout all phases, here fewer than 5% selected the without-*A* team. In addition, statistically the contrast between Conditions 1 and 2 was highly significant (tested in Phase 4), $\chi^2(1, N = 58) = 27.15$, $p < .001$.

In the heterogeneous global-and-local Condition 3, as predicted, selections began with a high proportion (78%) of no-altruist team choices in Phase 1 (Figure 10), $\chi^2(1, N = 32) = 67.57$, $p < .001$. In Phase 4, these individual-related selections, which exclude *A*, are likewise found to be above chance (59%), $\chi^2(1, N = 32) = 31.01$, $p < .001$; but now the group-related selections are above chance as well ("without N4", with 34%); $\chi^2(1, N = 32) = 4.13$, $p = .042$. In contrast, in the global-only Condition 4 (also heterogeneous), even in Phase 1 the optimal *team*-related selection (the with-altruist team excluding N4) is the most frequently selected (43 %), $\chi^2(1, N = 32) = 11.28$, $p < .001$; and the without-altruist team, by contrast, is selected even significantly below chance (3 %), $\chi^2(1, N = 32) = 5.70$, $p = .017$. In Phase 4, the selection of the individually optimal no-altruist team ("without A") is still selected with low relative frequency (3 %), and the overall optimal team (N4) by 59 % of participants. If one compares the two heterogeneous conditions, Condition 4 and Condition 3 (Phase 4), the proportion of the "without-altruist" team selections, is significantly reduced in Condition 4, $\chi^2(1, N = 64) = 23.56$, $p < .001$; and the selection of the team with the altruist that also involves the optimal normal worker (despite the absence of information on the individual level), is selected significantly more frequently, $\chi^2(1, N = 64) = 4.02$, $p = .045$.

Additionally, participants' comments were again coded if showing insight into differences between an individual's direct and overall earnings (individual versus group level) or into interactions between workers. This measure is only reasonably calculated for local-and-global conditions, where conflicting information on both levels was provided. After 80 learning rounds, at least 38% of the participants in these conditions were classified as providing such insightful comments (33% in Condition 1 and 43% in Condition 3). Of these participants, 87% selected the altruist in the personnel selection task (in the same Phase 4), whereas of the participants not demonstrating insight only 3% made this selection. Correspondingly, insight correlated highly and reliably with putting the altruist on the team in Phase 4 (r_φ = .86, p < .001), but also in Phase 3 (r_φ = .72, p < .001), Phase 2 ((r_φ = .55, p < .001) and, to some extent, perhaps, in Phase 1 (r_φ = .24, p = .06). Additionally, the results suggest that once participants recognized the value of *A* they mostly tended to continue to include *A* in the team. With respect to the rating tasks, we also found that insight correlated reliably, although somewhat less strongly, with whether altruists were *rated* higher than *all* other workers (Phase 4, r_φ = .38, p = .006; Phase 3, r_φ = .38, p < .006; Phase 2, r_φ = .23, p = .146; Phase 1, r_φ = .04, p = 1.000), and with whether the altruist was rated higher than *at least one* other worker (Phase 4, r_φ = .58, p < .001; Phase 3, r_φ = .44, p = .001; Phase 2, r_φ = .33, p = .019; Phase 1, r_φ = .11, p = .513). Whereas the altruist is detected by a large proportion of participants in the global-only condition early on, we see that in the local-and-global condition, optimal answers on the global level are less frequent but correct group-level ratings were correlated relatively early on with later stated insight.

Finally, with regard to the three additional tests, the correlation between insightful comments and a global-perception preference in the Kimchi-Palmer test did not reach significance, r_φ = .16, p = .26; and insight did not reliably correlate with the correct solution (61%) in the attention test, r_φ = –.05, p = .791. Nevertheless, participants with insightful comments showed higher average values in the Need for Cognition Scale (NFC, 2.6 vs. 10.5), t(60) = 1.93, p = .03 (one-tailed), and (b) relatively more altruist-selections in the selection task (t(60) = 2.64, p = .011).

Discussion

The results of Experiment 2 show, on the one hand, that the suggested ‘tragedy of personnel selection’ is quite stable even over 80 rounds although the level of insight seems somewhat higher than in Experiment 1a. This stability revealed itself despite investigating again strong and consistent effects of the altruist, relatively low number of involved workers, and although we now used 80 rounds. Although the results in the conditions with individual-and-group information seem tragic (the majority judges the presumably best worker as the worst), it also must be pointed out that about a third of the participants even in these conditions showed insight into interactions or individual-versus-group-level differences. Likewise, a substantial proportion of participants selected the ‘altruist’ as part of the team. The rating task seems somewhat less affected – perhaps because it is formulated more neutrally. Nevertheless, insight substantially correlated not only with selecting the altruist in the personnel selection task (*r* =.72), but also with analogous rating patterns (*r* = .53). Although insightful comments did not correlate with a global perception style (mini Kimchi-Palmer test) or with the attention test item, the results suggest that a higher Need-for-Cognition score is a positive predictor of insight.

Second, and importantly, in conditions in which participants were forced to focus on the group level, they could quickly detect that the presence of the altruist was correlated to the best outcomes on the group level. Thus it can be excluded that people are unable to detect such correlations. In the new global-only conditions (C2 and C4), participants were shown in principle to be even highly capable of seeing that the altruist had the strongest impact on the overall earnings of the group when they concentrated on the group-level information. Here correct group-level answers in both tasks were dominant even in the first test phase. The results are reminiscent of those concerning people’s performance after either group feedback or group-and-individual feedback, where it emerged that people who obtained individual *and* group feedback were unable to capitalize on this multiple-goal feedback (DeShon et al., 2004). However, we were here concerned with the different task that participants were not cast as individual workers, but rather as neutral human resource managers, excluding motivational conflicts. Participants in this role obtained overview data on both the individual *and* the group level or only on the group level and had to evaluate the workers.

The results strongly suggests that the detected lack of sensitivity to group-level information in this and previous experiments could not be due to a general inability to see such correlations, but rather – at least for the majority of participants – to a tendency to focus on this information if information on individual earnings was available. The pace at which people were able to learn the overall correlation even within 20 rounds suggests that spending only a quarter of the time on this level of analysis would have been enough to determine this correlation.

Nonetheless, the results show that most participants in the global-and-individual condition did not detect this correlation and presumably did not take into account the possibility that the global and individual impact of a worker may differ, even if would have been able to detect the correlation.

General Discussion

Findings

The experiments presented here were two-level personnel evaluation tasks (T-PETs) involving information provided at individual and group levels. Specifically, they examined the evaluation or selection of single individuals who interacted positively (negatively) with all other workers giving rise to strong positive (negative) overall effects on a team's performance at the group level, that were at odds with their individual level contributions. Faced with opposing effects on the individual and group levels, participants' focus seemed to be on direct individual contributions rather than on overall contribution to a team.

Experiments 1 and 2 showed that a clear majority of T-PET participants evaluated workers based on their individual earnings alone, without considering the higher overall contribution to the earnings of the group made by the ‘interactors’, and did so in all conditions examined. Experiment 1a on ‘altruist’ detection (or ‘facilitator’ detection), in keeping with the pilot study, showed that judgments failed to reflect the strong group-level effects or even a combination of both the group-level and individual-level effects, but instead remained mostly focused on the individual level. Despite frequently completely ignoring the large group-level effects of the ‘altruist’, participants were mostly able to distinguish small performance differences on the *individual* level. At the same time, the ‘altruist’-detection rate, did not seem substantially dependent on the homogeneity or heterogeneity of the normal workers. Moreover, the tendency to evaluate the group-serving ‘altruist’ as least well performing, was established by participants early on and remained stable throughout the task even though they were asked to repeatedly rate and select employees four times. Finally, even the personnel-selection tasks, which explicitly asked participants to assemble the best teams and thus might have focused participants on the overall team-earnings, mostly saw the best team player excluded from those teams.

In Experiment 1b the ‘interactor’ had a negative impact on group performance (strongly lowering their performance), while individually displaying the best performance, thus behaviourally displaying a kind of ‘egotism’. ‘Egoist’ detection was likewise affected by a tragic inability to appreciate the group level, with participants ignoring the large negative correlation between the egoist’s presence and the team’s overall performance. Once again, this deficit was accompanied by participant ability to detect much smaller performance differences at the individual level, indicating that it is not a lack of sensitivity to variation *per se* that drives these results. The detection of ‘egoists’ seems reminiscent of ‘cheater detection’, so that a past literature on the latter might have led one to expect greater sensitivity to ‘egoists’ than to ‘altruists’. However, we found only a slight advantage for egoist detection over altruist detection, suggesting that there is no strong preference of the one over the other which could have explained the phenomenon (Cosmides, 1989; Sperber & Girotto, 2002; von Sydow, 2006).

One possible explanation of our results could have been that participants are simply unable to detect group level correlations. However, Experiment 2 explicitly investigate whether and how quickly participants were able to detect the altruist's overall effect on the group's performance when forced to focus on group-level information only. In the absence of individual information, participants were clearly able to detect altruists. In fact, they mostly evaluated the altruist as the best employee, often selecting him or her for the team. Surprisingly, participants showed this as early as in the first test phase (after 20 trials). This shows that participants across our studies could have easily detected the correlation in a fraction of the available time and using this information is, in principle, well within their grasp. Hence the majority's failure to use this correlation must have other reasons – for instance because they simply do not consider the possibility that this team-level information might be important and differ from the individual-level information.

Likewise, a number of methodological concerns or limitations of the paradigm, such as lack of concentration or motivation can be ruled out: poor performance obtained, even though participants had repeated test phases and 80 learning trials, and were exposed to only five possible group configurations (teams of four out of five employees). In addition, our strict selection criteria for participants, the high sensitivity of participants to small individual differences (documented both in Experiment 1a/b and Experiment 2) that were considerably smaller than the overlooked group effects, and the result that our short final attention test did not reliably correlate with judgments related to overall (group-level) effects, seem to exclude lack of concentration as an explanation. This points to a substantive deficit that should be of genuine theoretical and practical concern. We consider a number of possible explanations for this inability to factor in group-level information appropriately below.

Before doing so, however, it must be stressed that a relevant minority in all experiments did provide evaluations, selections, and comments revealing that they clearly detected the crucial importance of the individually weak 'altruist' for overall group performance. Several of these participants commented that it was a surprising insight for them that both levels may differ, and most took this insight into account in their judgments and selections.

Implications

Our results suggest that we may be facing a ‘tragedy of personnel evaluation’ (and of personnel selection), in that evaluations often exclusively based on individual performance may not reflect who is best for a team’s performance. This may lead to both suboptimal evaluations and inefficient selection decisions. In our experiments this is demonstrated by the fact that employees who contributed the most to a team’s performance overall were systematically evaluated most negatively and those who contributed the least overall were evaluated most positively. Even though in the real world differences between group-level and individual-level performance may often be less pronounced, and the individually best employee may even be the one who is best for the group, there is no guarantee that individual and group factors will always work in parallel. Moreover, smaller differences will be even harder to detect, yet may still easily be large enough to have measurable impact on organisations. Although the boundary conditions of the observed phenomenon need to be explored in future, the tragic outcomes, observed here in the laboratory, warn of the danger of potentially tragic personnel selection in the real world as well.

One supporting reason is that, as reviewed in the Introduction, altruistic (or prosocial) as well as egoistic (or antisocial) interactions with others are central aspects of human behaviour in groups (Engel, 2011; Fiddick, Cosmides, & Tooby, 2000; Gollwitzer et al., 2009; Henrich et al., 2005; Nowak & Sigmund, 1998, 2005; Post, 2005; Rand et al., 2009; Sober & Wilson, 1999; von Sydow, 2006; Wilson & Wilson, 2007); and, more specifically, that various positive and negative interactions between individuals or of individuals within whole groups are crucial to the performance of teams, organizations and companies (Brief & Motowidlo, 1986; Li, Kirkman, & Potter, 2014; Mathieu et al., 2008, 2014; Memmert et al., 2015; Nielsen et al., 2009; Organ & Ryan, 1995; Pearce & Herbik, 2004; Podsakoff et al., 2009, 2010; van Scotter & Motowidlo, 1996). Although we found small differences between altruist and egoist detection and at least some detection of ‘interactors’, and others found some weak indications of positive evaluation of positive interactors or ‘facilitators’ (see discussion below), our results seem to indicate substantial problems in detecting and appropriately evaluating interactors. Given the ubiquity and importance of interactions between members of teams the observed tragic effects may well have high significance.

Although the current findings only show ‘tragic results’ in a laboratory setting, they warn that an unreflective use of number-based employee evaluation by managers may lead to sub-optimal outcomes. Thinking of group dynamics may require more subtle attentiveness on the part of personnel managers, but the overall yield may well warrant the effort.

To sum up, our results point out the need for managers to be aware of the difference between individual-level and group-level effects when evaluating employees working in teams. The paradoxical two-level nature of team performance studied here was only understood by a minority of participants who showed insight into the difference between individual-level and group-level performance and tended to select teams who had the best overall outcome. They in fact often described experiencing an “Aha” (or Eureka) moment. The possibility of such insight provides hope of further understanding of human state- or trait-variables and task or content factors facilitating such insight. Here we have just begun our investigation; and in fact we found that insightful comments made by the participants did not significantly correlate with the Cognitive Reflection Test (CRT, Frederick, 2005, in the pre-test), nor with a Kimchi-Palmer test item (Kimchi & Palmer, 1982), or a simple attention task. Indeed, our results suggested that, as hypothesized, people with high “Need-for-Cognition” values (Cacioppo, Petty, & Kao, 1984) may provide more insightful comments.

Limitations and Future Avenues of Research

The main goal of the present article is to point out that, given our demonstration of potential tragedy in personnel evaluation there is need for future avenues of research to help determine (1) the domains of application and boundary conditions, (2) potential mediating mechanisms, and (3) ways to mitigate the disastrous results.

(1) Future work needs to further clarify the boundary conditions of the reported phenomenon, particularly with respect to real world contexts. Since interactions between employees are almost ubiquitous in real-world scenarios, and individual-level and team-level effects need not have the same strength or direction, the present evidence suggests that a tragedy of personnel selection may well take place in real-life scenarios as well. However, whether, how far and under which conditions this is really the case needs to be explored. It would be interesting to explore, for instance, the following issues.

(a) Does the tragic group-level neglect occur also for other data patterns as well? As discussed above, it seems plausible that similar phenomena occur with less clear dissociations of individual and group effects or in more complex situations, with both being presumably much more frequent. However, we here modelled the interaction of general altruists' (or egoists') interacting unconditionally with all others. This is in line with central literature on altruism (e.g., Sober & Wilson, 1999; Wilson & Wilson, 2007) and with the idea of the importance of general team-level or unit-level influences (cf. Mathieu, Maynard, Rapp, & Gilson, 2008; Chen & Kanfer, 2006; Chen et al., 2013); but it needs to be investigated how far these findings fully or partly generalize to other kinds of more specific interactions or the evaluation of even smaller teams.

(b) Since we did not find effects based on level of formal education, the results seem to be quite general (apart from the Need for Cognition test). Although it therefore seems implausible that personnel managers are totally immune to the strong effects documented here, it would be interesting to explore whether samples of real personnel managers would demonstrate different effects.

(c) Our present findings used *number-based* personnel evaluations and personnel selections only. Perhaps the explored group-level neglect or interaction-neglect may occur less frequently or even disappear when people have not quantitative but qualitative information at hand. The research on effects of (extra-role) altruistic behavior on evaluations at least suggests that the neglect in such domains is not complete (Organ, 1997; Scotter, Cross, & Motowidlo, 2000; Grant & Patil, 2012, 562). This may partly be due to not investigating situations with clearly conflicting individual and group-level performance. The recognition of group effects may also still be underestimated in these more positive reports (but number-based T-PETs may particularly well be suited to investigate the exact weighting of such effects). In real life, managers may also use cues instead of performance information, since for instance people are often known to choose their social partners based on qualitative cues indicating moral aspects of their character (Everett, Pizarro, & Crockett, 2016). Interestingly, a large meta-analysis shows that prosocial behaviour (Organizational Citicenship Behavior, OCB) is positively correlated with employee’s performance ratings and also weaker with actual reward allocation (Podskoff et al., 2009). However, for our purpose it would not only be important to know which aspect of OCB elicited these effects, but also whether the effects are due to group-performance detection at all, or rather to taking OCB behaviour as a proxy for good performance on the individual or group level. The apparent correlation between individual performance and facilitation of the team's performance (Podskoff et al., 2009) may ,mitigate the tragic results found. Finally, even if it remains well possible that underweighting of group-level effects occurs even in less abstract settings, it also seems plausible to us, prima facie, that managers who work in direct contact with their teams and the involved work processes may less be prone to fall prey to such neglect. These questions require closer investigation; but the increasing role of number- and outcome-based evaluations conducted from far, instead of by senior-level managers in close contact to the team, suggest an increasing neglect (or underweighting) of group-level effects in real-world personnel evaluation as well.

(d) The evaluation-context used in our experiments may have elicited a context of competition and self-interest (Grant & Patil, 2012). This would be in accordance with the finding that providing feedback about the earnings of peers reduces acts of altruistic punishment in public-good games (Nikiforakis, 2010). Experiment 1b, however, shows that substantial *negative* interaction with other workers (i.e., ‘egoist detection’) was almost as difficult for the participants to assess. Nonetheless, a variant of this idea, which one may call ‘functional binding,’ may apply. Classic personnel selection rested almost exclusively on individual-level assessments; thus the way personnel evaluation and selection is done traditionally may serve as an individualist standard even if information on both the individual and team levels was provided. An *individualist* context may have prevented participants from considering both altruistic and egoistic interactions in the team; thus future research should investigate whether and how contexts may change the experiment-outcomes.

(e) Even if the tragic findings arise similarly for situations described here, where the composition of the optimal team is known based on full information about the performance of each possible constellation, future research needs to investigate situations where teams have to be build anew. Here the issue arises about the generality and transferability of altruistic or egoistic behavior or, more generally, of facilitating/inhibiting effects. Perhaps in such situations it would be easier to train teams to work cooperatively than selecting an optimal team. We think that it would be worthwhile to investigate these issues in the future.

Although the degree to which our findings can be generalized may depend on a wealth of factors, our results point to a pressing need. Just as evolutionary biologists should never confuse individual with inclusive fitness, managers must be aware of the dangers of confusing individual performance with an individual’s overall (or inclusive) effect on a team.

(2) We are now going to discuss further potential explanatory mechanisms. The discovered potential

tragedy of personnel evaluation may be based on quite general cognitive mechanisms. Here we have shown that altruist and egoist selections lead to what is at least a similar neglect of the interactor, and that we cannot attribute it to specific problems of altruist detection alone (cf. Cosmides, 1989; Sperber & Girotto, 2002; von Sydow, 2006; cf. Discussion of Experiment 1). Moreover, we have shown that this group-level neglect could not be attributed to a complete inability to detect correlations between the presence of certain employees and the overall earnings (Experiment 2). However, there are several factors and processes that may lead people not to assess this correlation. The tragedy may, for instance, be related to (a) problems of detecting interaction effects (Novick & Cheng, 2004), even though people may well to be able to detect trivariate causal relationships (Waldmann & Hagmayer, 2001; Meder, Hagmayer, & Waldmann, 2008), bivariate correlations (McKenzie & Mikkelsen, 2007) or probabilistic-logical patterns (von Sydow, 2011, 2016, 2017); (b) a difficulty realizing that many small externalities can add up to large payoffs (Dörner, 1989/1993; von Sydow, 2015); or (c) problems dealing with multilevel representations and with the Simpson's paradox (Fiedler, Walther, Freytag, & Nickel, 2003; Waldmann & Hagmayer, 2001; von Sydow, Hagmayer, & Meder, 2016). These general explanations may relate to and be of importance for several fields of basic research, such as causal induction, induction of correlations, complex problem solving or dynamic decision theory (Funke, 2001, 2014; Meder, Hagmayer, & Waldmann, 2008; Osman, 2010; Waldmann & Hagmayer, 2001). A disentangling of these potential and perhaps supplementary causes of our tragic results would certainly be a reasonable focus of future research.

In connection with an elaboration of such explanations, one might question, given our results, the assumption of the optimality of selecting teams that in previous rounds showed the highest overall performance. Indeed, in several central domains of cognitive science a number of known deviations from normative standards have led not only to suggestions that people are fundamentally biased, but also to alternative normative standards (for instance, in hypothesis testing, argumentation, or probability judgment; Fiedler & von Sydow, 2015, Gigerenzer, 1996; Oaksford & Chater, 1994; Hahn & Oaksford, 2007; Hertwig & Volz, 2013; von Sydow, 2011, 2016). Nonetheless, it seems difficult to question and almost a truism that seeking the best team requires examining the best configuration of employees in the past. It seems self-evident to choose the option with the best expected outcome. We pointed out that a neglect of positive or negative group effects of interactors (for instance altruistic cooperators or egoistic defectors) in evaluation contexts may have disastrous effects for companies and organisations. Moreover, our relatively simple task as well as the results of Experiment 2 strongly suggests that people are in principle cognitively able to detect interactors quickly based on the strong grouplevel correlations shown.

Nonetheless, it may in some sense be reasonable that participants did not *use* this information but perhaps aimed at a more *detailed* understanding of the problem than merely choosing the overall best of five teams. First, in many situations the sole analysis of individual contributions is a reasonable goal, and this may have prevented people from assessing overall team performance (including effects of interactions between individuals). Second, this may be linked to the perspective of causal decision making, emphasizing not only correlations but also causal pathways (Waldmann & Hagmayer, 2001, Lagnado, Waldmann, Hagmayer, & Sloman, 2006; Sloman & Hagmayer, 2006; Hagmayer & Meder, 2012; von Sydow et al., 2016). With this perspective the tragedy may be a clearly negative – or disastrous – side effect of an otherwise reasonable strategy to construct local pathways (von Sydow et al., 2016; cf. also Hebbelmann & von Sydow, 2017). Participants in the studied T-PETs may perhaps have tried to obtain process knowledge by exploring positive or negative effects of employees on other individual employees. Although in these tasks even the effects of the altruist on each single individual employee were greater than the generally detected smallest individual difference between normal employees, they were smaller than the greatest individual difference across employees. If people have tried to construct a causal (or logical) network of local relationships rather than detect the massive overall correlation of altruist and overall outcome, this task becomes more difficult and may substantially contribute to neglecting the positive effects of the “altruist”. This does not change the fact that a resultant neglect of an overall impact is false or even disastrous in terms of non-causal as well as causal decision making. However, the latter perspective may perhaps shed light on rational forces underlying these disastrous effects for personnel evaluation.

(3) Finally, from an applied perspective it would be interesting to explore conditions that might help to mitigate the neglect of group-level effects which seems disastrous in the reported experiments. The insightful solutions, at least by a substantial minority, show that the task is solvable; and the results of Experiment 2 show that people can quickly find the best group solution if focusing on group-level information alone. For instance, our results suggest that, *ceteris paribus*, an increased focus on the group-level may mitigate the current disastrous findings, perhaps at the price of lowering the high resolution at the level of individual performance.[9] To summarise, stressing the role of interaction in the instructions; stressing the role of teams in the culture of companies; using team-selection tasks without any evaluation task; and comparing several groups, perhaps together with stressing the importance of the teams, may help improve the evaluations and team selections.

Conclusion

Independent of the significance of further clarifications and investigations of the details of the processes involved and of exact boundary conditions, the practical importance of pointing out the danger of overlooking group-serving or group-harming employees and building inefficient teams

[9] In some more recent studies, however, we show that using two groups may indeed strengthen the group-level focus, but that this is not sufficient to truly overcome the tendency to focus on individuals (preliminary report in von Sydow & Braus, 2017, and von Sydow, Braus, & Hahn, 2017).

(based only on individual excellence) can hardly be overestimated, with regard to companies and other organizations. Overall, our findings underline a necessary awareness of the danger of a tragedy of personnel evaluation in the real world, with implications for incentive-structures (personnel evaluation), employee-advancement (personnel promotion) and job offers (personnel selection). Pointing out this problem may help find ways to overcome and even prevent it.

Acknowledgments

This article elaborates and extends previous presentations at the Cognitive Science Conference in Pittsburgh (von Sydow & Braus, 2016) and London (von Sydow & Braus, 2017; von Sydow, Braus, & Hahn, 2017). Parts of the work have also been presented at the 58th TeaP at the University of Heidelberg, the MuST 9 at the Munich Center for Mathematical Philosophy, at an invited talk at Tilburg University (Prof. Dr. Gideon Keren), at the annual German national psychology conference of the DGPs 2016 and the DGPs meeting on organizational and economic psychology in Dresden (AOW, 2017). This work was supported by the Sy111/2-1 Grant from the Deutsche Forschungsgemeinschaft (DFG) as part of the priority program *New Frameworks of Rationality* (SPP 1516). We are grateful to Maria Waltmann and Linda McCaughy for support, to Klaus Fielder, Joachim Funke, Daniel Holt, Sandra v. Sydow, and Michael Waldmann for discussing related issues, and to Martha Cunningham for corrections. The pictures in Figure 1 are from S. ter Burg, A. Hjartat, B. van Kleef, and T. Arbeider-parti (creative commons license: https://creativecommons.org/licenses/by/2.0/; www.flickr.com/photos/ and ter-burg/8172735997, apotekhjartat/13406119904, mvon-ederland/12364694653, tromsarbeiderparti/4577244545). We thank previous listeners at conferences and anonymous reviewers for their helpful comments and suggestions. We are grateful for the positive and helpful feedback of three anonymous reviewers.

References

Becker, B., & Gerhard, B. (1996). The Impact of Human Resource Management on Organizational Performance: Progress and Prospects. *Academy of Management Journal, 39*(4), 779–801.

Beersma, B., Hollenbeck, J. R., Humphrey, St., Moon, H., Conlon, D. E. & Ilgen, D. R. (2003). Cooperation, competition, and team performance - towards a contingency approach. *Academy of Mannagement Journal, 46*(5), 592-590.

Beller, S. (2010). Deontic reasoning reviewed: psychological questions, empirical findings, and current theories. *Cognitive Processing, 11*, 123-132.

Brief, A. P. & Motowidlo, St. J. (1986), Prosocial Organizational Behavior. *Academy of Management Review, 11*(4), 710–725.

Bucciarelli, M., & Johnson-Laird, P. N. (2005). Naïve deontics: A theory of meaning, representation, and reasoning. *Cognitive Psychology, 50*, 159–193.

Brandl, J. (2002): Die Problematik der Kennzahlen in Personalinformationssystemen. *Personalführung, 9*, 42-47.

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition, 31*(3), 187–276. doi:10.1016/0010-0277(89)90023-1

Cacioppo, J. T., Petty, R. E., & Kao, C. F., (1984). The Efficient Assessment of Need for Cognition. *Journal of Personality* Assessment, *48,* 306–307. doi:10.1207/s15327752jpa4803_13

Chen, G., Farh, J., Campbell-Bush, E. M., Wu, Z., & Wu, X. (2013). Teams as innovative systems: Multilevel motivational antecedents of innovation in R&D teams. *Journal of Applied Psychology, 98*(6), 1018-1027. doi:10.1037/a0032663

Chen, G., & Kanfer, R. (2006). Toward a systems theory of motivated behavior in work teams.

Research in Organizational Behavior, 27, 223–267. doi:10.1016/S0191-3085(06)27006-0

Curry, O., Price, M., & Price, J. (2008). Patience is a virtue: Cooperative people have lower discount rates. *Personality and Individual Differences, 44*(3), 780–785.

DeShon, R. P. Kozlowski, S. W. J., Schmidt, A. M., Milner, K. R., & Wiechmann, D. (2004). A Multiple-Goal, Multilevel Model of Feedback Effects on the Regulation of Individual and Team Performance. *Journal of Applied Psychology, 89*(6), 1035-1056.

Dalal, R. S. (2005). A Meta-Analysis of the Relationship Between Organizational Citizenship Behavior and Counterproductive Work Behavior. *Journal of Applied Psychology, 90*(6), 1241–1255.

Dörner, D. (1989). *The logic of failure*. New York, NY: Holt (cf. German edition, *Die Logik des Mißlingens,* 1993).

Felps, W., Mitchell, T. R., & Byington, E. 2006. How, when, and why bad apples spoil the barrel: Negative group members and dysfunctional groups. *Research in Organizational Behavior, 27*: 175-222.

Fiedler, K., Walther, E., Freytag, P., & Nickel, S. (2003). Inductive reasoning and judgment interference: Experiments on Simpson's paradox. *Personality and Social Psychology Bulletin, 29*, 14–27.doi:10.1177/0146167202238368

Fodor J. (2000). Why we are so good at catching cheaters. *Cognition, 75*(1): 29-32.

Grant, A. M. & Patil, S. V. (2012). Challenging the norm of self-interest. Minority influence and transitions to helping norms in work units. *Academy of Management Review, 37(4)*, 547–588.

Engel, C. (2011). Dictator Games: A Meta Study. *Experimental Economics, 14*, 583–610.

Everett, J. A. C., Pizarro, D.. A., & Crockett, M. J. (2016). Inference of Trustworthiness from Intuitive Moral Judgments, *Journal of Experimental Psychology: General*, 145(6):772–787. doi:10.1037/xge0000165

Fehr, E. & Schmidt, K. (1999). A Theory of Fairness, Competition, and Cooperation. *Quarterly*

Journal of Economics, 114, 817–868. doi:10.1162/003355399556151

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature, 415*, 137–140. doi:10.1038/415137a

Fehr, E., & Fischbacher, U. (2004). Social norms and human cooperation. *Trends in Cognitive Science, 8,* 185–190. doi:10.1016/j.tics.2004.02.007

Fiddick, L., Cosmides, L., & Tooby, J. (2000). No interpretation without representation: The role of domain-specific representations and inferences in the Wason selection task. *Cognition, 77*, 1–79.

Fiedler, K., & von Sydow, M. (2015). Heuristics and Biases: Beyond Tversky and Kahneman's (1974) Judgment under Uncertainty (pp. 146-161). In: M. W. Eysenck & D. Groome. *Cognitive Psychology: Revisiting the Classical Studies*. Los Angeles, London: Sage.

Frederick, S., Loewenstein, G., & O'Donoghue, T. (2002). Time discounting and time preference: A critical review. *Journal of Economic Literature, 40*, 351–401.

Funke, J. (2001). Dynamic systems as tools for analyzing human judgment. *Thinking and Reasoning, 7*, 69–89.

Funke, J. (2014). Analysis of minimal complex systems and complex problem solving require different forms of causal cognition. *Frontiers in Psychology, 5*(739), 1-3. doi:10.3389/fpsyg.2014.00739.

George, J. M. & Bettenhausen, K. (1990). Understanding Prosocial Behavior, Sales Performance, and Turnover: A Group-Level Analysis in a Service Context. *Journal of Applied Psychology, 75*(6), 698–709.

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky (1996). *Psychological Review, 103*, 592–596.

Gigerenzer, G., & Hug, K. (1992). Domain-specific reasoning: Social contracts, cheating, and perspective change. *Cognition, 43*, 127–171.

Gollwitzer, M., Rothmund, T., Pfeiffer, A., & Ensenbach, C. (2009). Why and when justice sensitivity leads to pro- and antisocial behavior. *Journal of Research in Personality, 43(6)*, 999–1005.

Hagmayer, Y., & Meder, B. (2013). Repeated causal decision making. *Journal Of Experimental Psychology: Learning, Memory, And Cognition, 39*, 33-50. doi:10.1037/a0028643.

Hahn, U. & Oaksford, M. (2007). The rationality of informal argumentation: A Bayesian approach to reasoning fallacies. *Psychological Review, 114,* 704-732.

Hardin, G. (1968). The tragedy of the commons. *Science, 162* (3859), 1243–1248. doi:10.1126/science.162.3859.1243

Haslam, S. A., Steffens, N. K, Peters, K, Boyce, R. A., Mallett C. J., & Fransen, K. (2017). A social identity approach to leadership levelopment. *Journal of Personnel Psychology, 16*(3), 113–124. doi:10.1027/1866-5888/a000176

Hebbelmann, D. & von Sydow, M. (2017). Betting on Transitivity in an Economic Setting. *Cognitive Processing, 18*, 505-518. doi:10.1007/s10339-017-0821-x.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., et al. (2005). 'Economic man' in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences, 28*(6), 795–855.

Hertwig, R., & Volz, K. G. (2013). Abnormality, rationality, and sanity. *Trends in Cognitive Sciences, 17*, 547-549.

Kimchi, R., & Palmer, St. E. (1982). Form and texture in hierarchically constructed patterns. *Journal of Experimental Psychology: Human Perception and Performance, 8*(4), 521–535. doi:10.1037/0096-1523.8.4.521

Kiyonari, T., & Barclay, P. (2008). Cooperation in social dilemmas: Free riding may be thwarted by second-order reward rather than by punishment. *Journal of Personality and Social Psychology, 95*(4), 826–842.

Lagnado, D. A., Waldmann, M. R., Hagmayer, Y., & Sloman, S. A. (2006). Beyond covariation: Cues to causal structure. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation.* Oxford: Oxford University Press.

Li, N., Kirkman, B. L., & Porter, C. O. L. H. (2014). Toward a model of work team altruism. *Academy of Management Review, 39(4),*541–565. doi:10.5465/amr.2011.0160

Mathieu, J., Maynard, M. T., Rapp, T., Gilson, L. (2008). Team effectiveness 1997-2007: A review of recent advancements and a glimpse into the future. *Journal of Management, 34*(3), 410-476. doi:10.1177/0149206308316061

Mathieu, J. E., Tannenbaum, S. I., Donsbach, J. S., & Alliger, G. M. (2014). A review and integration of team composition models moving toward a dynamic and temporal framework. *Journal of Managment, 40*(1), 130-160. doi:10.1177/0149206313503014

McKenzie, C. R. M., & Mikkelsen, L. A. (2007). A Bayesian view of covariation assessment. *Cognitive Psychology, 54,* 33-61.

Meder, B., Hagmayer, Y., & Waldmann, M. R. (2008). Inferring interventional predictions from observational learning data. *Psychonomic Bulletin & Review,* 15 (1), 75-80.

Melis, A. P., Hare, B., Tomasello, M. (2006). Chimpanzees recruit the best collaborators. *Science, 311,* 1297–1300.

Memmert, D., Plessner, H., Hüttermann, S., Froese, G., Peterhänsel, C., & Unkelbach, C. (2015). Collective fit increases team performances: Extending regulatory fit from individuals to dyadic teams. *Journal of Applied Social Psychology, 45*, 274–281. doi:10.1111/jasp.12294

Mischel, W., Shoda, Y. & Peake, P. K. (1988). The nature of adolescent competencies predicted by preschool delay of gratification. *Journal of Personality and Social Psychology, 54,* 687-696.

Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology, 9*, 353–383.

Nielsen, T. M., Hrivnak, G. A., & Shaw, M. (2009). Organizational Citizenship Behaviour and

Performance. A Meta-Analysis of Group-Level Research. *Small Group Research, 40*(5), 555-577. doi:10.1177/1046496409339630

Nikiforakis, N. (2010). Feedback, punishment and cooperation in public good experiments. *Games and Economic Behavior, 68*(2), 689–702.

Novick, L. R., & Cheng, P. W. (2004). Assessing Interactive Causal Influence. *Psychological Review, 111*, 455–485. doi:10.1037/0033-295X.111.2.455

Nowak, M. A., & Sigmund, K. (1998). The dynamics of indirect reciprocity. *Journal of Theoretical Biolology, 194*, 561–574.

Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature, 437/27*, 1291–1296. doi:10.1038/ nature04131.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review, 101*, 608–631.

Organ, D. W., & Ryan, K. (1995). A meta-analytic review of attitudinal and dispositional predictors of organizational citicenship behavior. *Personnel Psychology, 48*, 757–802.

Organ, D. W. (1997). Organizational Citizenship Behaviour: It's Construct Clean-Up Time. *Human Performance, 10*(2), 85–97.

Osman, M. (2010). Controlling Uncertainty: A Review of Human Behavior in Complex Dynamic Environments. *Psychological Bulletin, 136*(1), 65-86.

Ostrom, E., Burger, J., Field, C. B., Norgaad, R. B., & Policansky, D. (1999). Revisiting the Commons: Local Lessons, Global Challenges. *Science, 284* (5412), 278–282. doi:10.1126/science.284.5412.278.

Pearce, C. L. & Herbik, P. A. (2004). Citizenship Behavior at the Team Level of Analysis: The Effects of Team Leadership Team Commitment, Perceived Team Support, and Team Size. *The Journal of Social Psychology, 144*(3), 293–310.

Podsakoff, N. P., Whiting, S. W., Podsakoff, P. M., & Blume, B. D. (2009). Individual- and

Organizational-Level Consequences of Organizational Citizenship Behaviours: A Meta-Analysis. *Journal of Applied Psychology, 94(1),* 122–141. doi:10.1037/a0013079.

Podsakoff, N. P., Whiting, S. W., Podsakoff, P. M., & Mishra, P. (2010). Effects of organizational citizenship behaviors on selection decisions in employment interviews. *Journal of Applied Psychology, 96 (2),* 310–326.

Post, St. G. (2005). Altruism, Happyness & Health: It's Good to be Good. *International Journal of Behavioural Medicine, 12(2),* 66–77.

Rand, D., Dreber, A., Ellingsen, T., Fudenberg, D., & Nowak, M. (2009). Positive interactions promote public cooperation. *Science, 325*(5945), 1272–1275.

Stewart, G. L., & Nandkeolyar, A. K. 2007. Exploring how constraints created by other people influence intraindividual variation in objective performance measures. *Journal of Applied Psychology, 92*: 1149-1158.

Sober, Elliott (1998). What is Evolutionary Altruism? In: Hull, David Lee; Ruse, Michael (Eds.): *The Philosophy of Biology*. Oxford, New York: Oxford University Press, 1998, pp. 459-475.

Sober, E., & Wilson, D. (1999). *Unto Others: The Evolution of Unselfish Behavior*. Harvard University Press.

Sloman, S. A., & Hagmayer, Y. (2006). The causal psycho-logic of choice. *Trends in Cognitive Sciences, 10*, 407– 412. doi:10.1016/j.tics.2006.07.00

Sperber, D. & Girotto, V. (2002). Use or misuse of the selection task? Rejoinder to Fiddick, Cosmides, and Tooby. *Cognition, 85*, 277–290.

Van Scotter, J. R., & Motowidlo, S. J. (1996). Interpersonal Facilitation and Job Dediction as Separate Facets of Contextual Performance. Journal of *Applied Psychology, 81*(5), 525–531.

Van Scotter, J. R., Motowidlo, S. J., & Cross, T. C. (2000). Effects of task performance and contextual performance on systemic rewards. *Journal of Applied Psychology, 85*(4), 526–535.

von Sydow, M. (2006). Towards a Flexible Bayesian and Deontic Logic of Testing Descriptive and Prescriptive Rules. Dissertation, Georg-August-Universität Göttingen.

von Sydow, M. (2012). From Darwinian Metaphysics towards Understanding the Evolution of Evolutionary Mechanisms. Göttingen: Universitätsverlag Göttingen.

von Sydow, M. (2015). The Tragedy of Inner-Individual Dilemmas. In D. Noelle, R. Dale, A. Warlaumont, J. Yoshimi, T. Matlock, C. Jennings, & P. Maglio (Eds.), *Proceedings of the 37th An. Conf. of the Cognitive Science Society* (pp. 2517-2522). Austin, TX: Cognitive Science Society.

von Sydow, M. (2011). The Bayesian Logic of Frequency-Based Conjunction Fallacies. *Journal of Mathematical Psychology, 55*(2), 119-139. doi:10.1016/j.jmp.2010.12.001

von Sydow, M. (2016). Towards a Pattern-Based Logic of Probability Judgements and Logical Inclusion “Fallacies”. *Thinking & Reasoning, 22(3),* 297-335. doi:10.1080/13546783.

von Sydow, M. (2017). Rational Explanations of the Conjunction Fallacies – A Polycausal Proposal. *Proceedings of the Thirty-Ninth Annual Conference of the Cognitive Science Society* (pp. 3472-3477). Austin, TX: Cognitive Science Society.

von Sydow, M., & Braus, N. (2016). On the Tragedy of Personnel Evaluation. In A. Papafragou, D. Grodner, D. Mirman, & J.C. Trueswell (Eds.), *Proceedings of the Thirty-Eighth Annual Conference of the Cognitive Science Society* (pp. 105-110). Austin, TX: Cognitive Science Society.

von Sydow, M., & Braus, N. (2017). Altruist vs. Egoist Detection and Individual vs. Group Selection in Personnel Management. *Proceedings of the Thirty-Ninth Annual Conference of the Cognitive Science Society* (pp. 3466-3471). Austin, TX: Cognitive Science Society.

von Sydow, M., Braus, N., & Hahn, U. (2017). Overcoming the Tragedy of Personnel Selection? *Proceedings of the Thirty-Ninth Annual Conference of the Cognitive Science Society* (pp. 3460-3465). Austin, TX: Cognitive Science Society.

von Sydow, M., & Hagmayer, Y. (2006). *Deontic Logic and Deontic Goals in the Wason Selection*

Task. In R. Sun & N. Miyake (Eds.). Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society (pp. 864-869). Mahwah, NJ: Erlbaum.

von Sydow, M., Hagmayer, Y., & Meder, B. (2016). Transitive reasoning distorts induction in causal chains. *Memory & Cognition*, *44*(3), 469–487. doi:10.3758/s13421-015-0568-5

Waldmann, M. R., & Hagmayer, Y. (2001). Estimating causal strength: The role of structural knowledge and processing effort. *Cognition*, *82*, 27–58. doi:10.1016/S0010-0277(01)00141-X

Wilson, D. S., & Wilson, E. O. (2007). Rethinking the theoretical foundation of sociobiology. *Quarterly Review of Biology*, *82*(4), 2007, 327–348. doi:10.1086/522809