Tuyttens, F. A. M., Stadig, L., Heerkens, J., Van laer, E., Buijs, S., & Ampe, B. (2016). Opinion of applied ethologists on expectation bias, blinding observers and other debiasing techniques. Applied Animal Behaviour Science, 181, 27-33. DOI: 10.1016/j.applanim.2016.04.019

Opinion of applied ethologists on expectation bias, blinding observers and other debiasing techniques

Frank A. M. Tuyttens[a,b,*], Lisanne Stadig[a], Jasper L. T. Heerkens[a], Eva Van laer[a], Stephanie Buijs[a] and Bart Ampe[a]

[a] Animal Sciences Unit, Institute for Agricultural and Fisheries Research (ILVO), Melle, Belgium

[b] Faculty of Veterinary Medicine, Ghent University, Merelbeke, Belgium

[*] corresponding author: frank.tuyttens@ilvo.vlaanderen.be

Prof. Dr. Frank Tuyttens

Animal Sciences Unit

Institute for Agricultural and Fisheries Research (ILVO)

Scheldeweg 68, 9090 Melle, BELGIUM

Tel: +32 9 2272605, Fax: +32 9 2272601

E-mail: frank.tuyttens@ilvo.vlaanderen.be

ABSTRACT

There is increasing evidence that the field of applied ethology is prone to expectation biases invalidating research outcomes. Nevertheless, outcome assessors are rarely blinded. We surveyed delegates of the International Society for Applied Ethology (ISAE) 2014 congress shortly before (n=39 respondents) and after (n=51 respondents) a combined congress plenary and workshop on expectation bias in applied ethology. The aims were to evaluate the effect of the plenary and workshop on the opinion of applied ethologists in order to better comprehend why blinding outcome assessors seems so rarely practiced as a debiasing technique in this field of research. The results suggest that a moderate awareness about expectancy effects among ethologists and the logistic constraints of blinded observations rather than a perceived low susceptibility of the research field is the larger part of the explanation. Although awareness about expectancy effects and debiasing techniques was higher immediately after than before the congress plenary and workshop, a more sustained and concerted effort is needed throughout all stages of the research process to avoid expectation bias invalidating research finding and to improve the scientific credibility of the field of applied ethology.

Keywords: animal behaviour, blinding, cognitive bias, confirmation bias, context effects, observer bias,

1. Introduction

Although scientific research is meant to yield correct information about reality, there is increasing concern that many claimed research findings are false (Ioannidis, 2005). Attempts to reproduce research findings – even if published in top-tier journals – often fail (Begley and Ellis, 2012). False findings do not only lead to confusion and disappointment, they also cause a waste of resources, may pose a threat to the health of humans, animals or ecosystems, and hamper scientific progress and credibility such that its impact on society is sub-optimal. The publication of false or exaggerated positive findings appears to be more common in research fields where replication is difficult, theories are less clear, and methods are less standardized, because researchers have more "degrees of freedom" to produce the results they expect (Fanelli and Ioannidis, 2013; Ionnidis 2005, 2008). Behavioural studies are more likely to report exagerated effects (Fanelli 2010; Fanelli and Ioannidis, 2013) and have long been considered to run a particularly high risk of bias (Burghardt et al., 2012; Rosenthal, 1966; Tuyttens et al., 2014; van Wilgenburg and Elgar 2013).

Bias can be defined as the combination of various design, data collection, data analysis, interpretation and presentation factors that tend to produce false research findings (Ioannidis, 2005). These biases can be conscious (i.e. fraud and

scientific misconduct) or unconscious. The present paper focusses on a particular type of unconscious bias, namely expectation bias (also termed confirmation bias, cognitive bias, observer bias or context effects). Expectation bias is a common source of inaccuracy where outcome assessors observe or interpret their observation in a way that supports their expectations or preferred hypothesis (Kardish et al., 2015). As a relatively young scientific discipline that relies heavily on behavioural methods, applied ethology would appear to be at high risk of expectation bias. Moreover other predisposing factors for a high risk of expectancy effects are commonly present in applied ethology. For example, there seems little reason to assume that data collection staff in applied ethology have weak preconceptions or interests in the research outcome, or that underlying data to be assessed are less ambiguous, or that scoring methods are less subjective as compared to other scientific disciplines.

Despite these predisposing factors, and despite growing experimental evidence of expectation bias when using ethological research methods (Bohlen et al. 2014; Marsh and Hanlon, 2004, 2007; Tuyttens et al., 2014; van Wilgenburg and Elgar, 2013), blind observation is either grossly underreported or underutilized in animal behaviour studies (Burghardt et al., 2012; Kardish et al., 2015). Yet, blinding outcome assessors is widely accepted as an important aspect of a good study design to avoid expectation bias, in particular when outcome measures are not clearly defined, hard to perceive and require human judgement, and when the assessor has an interest in the outcome of the study (Rosenthal, 1966; Savović et al., 2012; Schulz and Grimes, 2002). Meta-analyses provide evidence that non-blinded studies result in mostly exaggerated (but sometimes also obscured) treatment effects compared to blinded studies, presumably due to expectancy effects (Bello et al., 2012; Hróbjartsson et al., 2012, 2013; Schulz et al., 1995). Blinding of relevant people involved in the study (including e.g. subjects, investigators, outcome assessors, data-managers and -analysts) is widely considered the gold standard study design that most effectively demonstrates the effectiveness of a product or intervention (Kaptchuk, 2001; Miller and Stewart, 2011). An increasing number of guidelines about what information should be provided in a research article (e.g. CONSORT, ARRIVE) demand a clear description of which members of the research staff were blinded (Kilkenny et al., 2010).

In order to improve the scientific credibility of applied ethology research, we ought to better understand the reasons why blind observation is not practiced more commonly. Possible reasons include, for example, a lack of awareness among applied ethologists about bias caused by expectancy effects, the conviction that their studies are not prone to expectation bias, a poor reporting of blinding methods in their publications, logistic constraints of blinding research staff, or the use of other debiasing techniques. The aim of this study was, therefore, to survey attendants of the International

Society for Applied Ethology (ISAE) 2014 congress on their opinion regarding the relevance and magnitude of expectation bias depending on the type of assessment outcome, and on the potential implications and solutions for their own research and for the field of applied ethology at large. The survey was conducted shortly before and after a combined congress plenary and workshop on expectation bias in applied ethology, so that its effect on the attendants' opinion could be evaluated.

2. Methods

2.1. Survey

A survey was conducted among delegates of the ISAE-2014 congress before and after a plenary talk and a workshop aimed at raising the issue of expectation bias among applied ethologists. The plenary talk was on expectation bias in applied ethology and covered three trials on veterinary students illustrating that subjective scorings of animal behaviour and welfare can be biased by context information about the conditions in which the animals were filmed (Tuyttens et al., 2014). Later on the same afternoon, this plenary talk was followed by a workshop entitled "Context and expectation effects in applied ethology: implications and solutions". During the workshop an overview was given of various debiasing techniques that had been suggested for other research fields such as forensic science (Reese, 2012) and clinical decision making (Croskerry et al., 2013a,b). Some techniques focused on debiasing the decision-making task, such as blinding, sequential unblinding, and reducing the complexity or ambiguity of the task. Other techniques focused on debiasing the decision-maker, such as training in bias awareness and critical thinking, perspective taking, augmenting accountability and reducing emotional context. This overview was followed by a group discussion on the applicability of these techniques for applied ethology research. At the end of the workshop participants were requested to fill out the same survey (irrespective of whether or not they had already filled out the survey before the plenary talk).

Congress delegates had received an email invitation to fill out the survey a couple of days before the start of the congress and were asked again to fill out the questionnaire during registration for the congress. Thirty-nine and 51 completed surveys were received before and after the plenary and workshop, respectively (Table 1). Although 13 anonymous surveys were received after the plenary and workshop, the vast majority of the respondents were researchers (including post-docs, research associates, professors) or (undergraduate or graduate) students. It was checked afterwards that a minority of the non-anonymous respondents had not yet published a scientific peer-reviewed paper on ethology that

was listen in the Web of Knowledge (Table 1). Fourteen attendants (11 researchers and three students) had filled out the survey both before and after the plenary and workshop.

The survey started with explaining that expectation bias refers to the possible influence of researchers' expectations or desires about the effect of certain (experimental) treatments or factors studied, and that such bias may (unconsciously) influence data recording and interpretation. It continued to explain that the survey probed into their opinion as an ethologist regarding the relevance and magnitude of this problem, the potential implications, and potential solutions. Although respondents were requested to fill out their name, it was promised that the results would be kept anonymous. It was also stated that there were no correct or incorrect answers as we were interested in their personal opinion.

The survey consisted of 8 questions. With the exception of two questions respondents could answer using a 10-point Likert scale ranging from 1 (not at all/very low) to 10 (very much/very high). They were asked about the perceived susceptibility to expectation bias of 5 types of observation-based recordings: occurrence of various behaviours (using an ethogram), interpreting the outcome of behavioural interactions, characterizing animal personalities, quantifying the severity of physical or clinical conditions, and assessing emotional state. The respondents were asked to what extent expectation bias may have influenced their own research outcomes in general and that of similar research performed by peers. Subsequently they were requested to consider their next experiment or study and asked about the perceived importance and feasibility to minimize/prevent expectation bias. Next they were asked to indicate how effective and feasible they considered various methods (listed in Fig. 6) for reducing expectation bias in their own research.

For the remaining questions they were asked to consider four theoretically different research situations in which the methods were objective versus subjective and in which the researcher/data-collector had no versus strong expectations about the outcome (Table 2). The respondents were asked what percentage of research they would categorize in the four situations in applied ethology and other scientific disciplines, e.g. animal physiology. Finally, they were asked to indicate for each of these four research situations, (i) how susceptible these are to expectation bias, and (ii) how much these would be influenced as an editor/reviewer of a respected peer reviewed journal by the fact that data-recorders were not blinded when deciding against or in favour of publication.

2.2. Statistical Analysis

The results of questions with a Likert scale were analysed using linear mixed regression models with, in case of repeated measures (paired observations or the multiple answers to sub questions), a random effect for respondent to

correct for this. All analyses were performed on both the full dataset and a subset of the dataset with only paired observations (respondents who filled out the questionnaire before and after the plenary and workshop). Fixed factors included in the models were job category, subquestion if applicable en answers to other questions if relevant. In the paired analyses time (before or after plenary and workshop) was also included as fixed effect as well as the relevant interactions with the other fixed factors in the model. Non-significant interactions were removed from the final models. All results presented are based on the full dataset, unless reported otherwise. In case of posthoc pairwise testing (e.g. between sub questions, before-after), p-values were corrected with the Tukey-Kramer adjustment for multiple comparisons. The analysed data were considered sufficiently normally distributed, based on the graphical evaluation (histogram and QQ-plot) of the residuals. All tests were two tailed at a significance level of 5% and all calculations were performed using the lme function from the nlme package in R 3.0.2. Interactions and fixed effects were removed from the model if the estimated effect was not significant.

3. Results

3.1. Susceptibility to expectation bias

Of five different types of observation-based recordings, the respondents considered 'assessing emotional state' and 'characterizing animal personalities' as more susceptible to expectation bias, followed by 'interpreting the outcome of behavioural interactions' (Fig. 1). 'Occurrence of various behaviours (using an ethogram)' and 'quantifying the severity of physical/clinical conditions' were considered to be the least susceptible types of observation-based recordings. Across all five types of recordings, susceptibility to expectation bias was judged 8.1% higher after the plenary and workshop as compared to before ($t_{365} = 3.66$, $P < 0.001$, Fig. 1). This effect was even greater when only the 14 respondents who filled out the questionnaire before and after the plenary and workshop were considered (10.4%, $t_{121} = 5.36$, $P < 0.001$).

Fig. 2 shows that susceptibility to expectation bias was considered the lowest for research using objective methods and in which the researcher has no expectations about the outcome (situation Obj-NoE: mean score = 3.03 ± 0.22). Susceptibility was increased if the method is subjective (situation Sub-NoE: mean score = 5.06 ± 0.22) or when the researcher has strong expectations (situation Obj-Exp: mean score = 5.54 ± 0.22). If the method is subjective and the researcher has strong expectations, susceptibility to expectation bias was judged to be the highest (situation Sub-Exp: mean score = 8.19 ± 0.22). Across the various research situations, susceptibility was scored 6.2% higher after the plenary and workshop as compared to before ($t_{280} = 2.60$, $P = 0.010$). Again, this effect was even greater when only the 14

respondents who had filled out the questionnaire before and after the plenary and workshop were considered (7.1%, t_{94} = 2.61, P = 0.010).

Table 3 illustrates that the research situation that was considered the most susceptible to expectation bias (situation Sub-Exp) was judged more common in applied ethology as compared to other scientific disciplines such as animal physiology, whereas the least susceptible research situation (situation Obj-NoE) was judged less common in applied ethology. Respondents reported that more subjective methods are used in applied ethology as compared to other scientific disciplines (situations Sub-Exp + Sub-NoE: 43.7% versus 29.0%, t_{83} = 7.44, P < 0.001), and that the researchers are more likely to have strong expectations about the research outcome (situations Obj-Exp + Sub-Exp: 70.4% vs 65.2%, t_{83} = 2.30, P = 0.024).

3.2. Influence of non-blinded data collection on peer-review

The respondents indicated that the influence of non-blinded data collection on their decision as an editor or reviewer to accept or reject publication of a paper in a respected peer-reviewed journal would depend on the type of research ($F_{3,271}$ = 60.21, P < 0.001). Not blinding would be better tolerated if the data collector has no expectations about the outcome and uses objective methods (situation Obj-NoE) than if he uses subjective methods (situations Sub-Exp and Sub-NoE) (Fig. 3). The magnitude of these differences between research situations was more pronounced after than before the plenary and workshop (paired analysis, situation x moment, $F_{3,271}$ = 4.36, P = 0.005). After the plenary and workshop, the respondents were more critical about not-blinded data collection in research situation Sub-Exp than before (Fig. 3).

3.3. Influence of expectation bias on own and peers' research outcomes

Fig. 4 shows that the respondents perceived the influence of expectation bias to be smaller for their own research outcomes than for similar research performed by their peers ($F_{1,97}$ = 13.13, P < 0.001). The perceived influence was higher after than before the plenary and workshop irrespective of whether the research was performed by themselves or by their peers ($F_{1,97}$ = 9.52, P = 0.003).

3.4. Preventing expectation bias in respondents' next study

Considering their next experiment or study, the respondents believed it was highly important to take actions to minimize expectation bias (Fig. 5). The feasibility to effectively prevent expectation bias in their next study, however, was scored much lower ($F_{1,102} = 106.15$, $P < 0.001$). Importance and feasibility were both scored higher after the plenary and workshop as compared to before ($F_{1,102} = 4.01$, $P = 0.048$).

3.5. *Effectiveness and feasibility of debiasing techniques*

The perceived effectiveness of various methods for reducing expectation bias in the respondents' own ethological research before the plenary and workshop was different as compared to afterwards (method x moment: $F_{7,617} = 3.36$, $P = 0.002$) (Fig. 6a). Before the plenary and workshop, effectiveness was judged lowest (but still slightly above the neutral point of the scale) for 'randomly assigning different data-collectors to different experimental treatments' and 'balancing data-collectors with opposite expectations over different experimental treatments' and highest for 'blinding staff involved in data-recording'. After the plenary and workshop, there was a significant increase in the perceived effectiveness of the three debiasing techniques that had been scored lowest beforehand ('randomly assigning different data-collectors to different experimental treatments', 'balancing data-collectors with opposite expectations over different experimental treatments' and 'raising awareness of research staff about expectation bias'). The effectiveness of the other debiasing techniques was not scored significantly different before versus after the plenary and workshop.

With the exception of 'raising awareness of research staff about expectation bias' the various debiasing techniques generally received lower scores for their feasibility than their effectiveness (based on average scores, not tested statistically, Fig. 6). Before the plenary and workshop 'balancing data-collectors with opposite expectations over different experimental treatments' received the lowest feasibility scores, whereas 'raising awareness of research staff about expectation bias' followed by 'using unambiguous definitions when categorizing behaviours' were scored highest (Fig. 6b). The order of debiasing techniques from the least to the most feasible was similar before versus after the plenary and workshop. However, after the plenary and workshop all debiasing techniques were judged to be 6.8% more feasible than before ($F_{1,619} = 8.83$, $P = 0.003$) (Fig. 6b).

4. Discussion

We surveyed applied ethologists shortly before and after they had attended a plenary and workshop on expectation bias at the ISAE-2014 congress in order to better comprehend why blinding outcome assessors seems so

rarely practiced as a debiasing technique in this field of research. The vast majority (85%) of the non-anonymous respondents had published at least one peer-reviewed ethological paper, which suggests that our sample of respondents can be considered to have some expertise in applied ethology. As the survey was restricted to the congress delegates and as participation was voluntary (without any inclusion or exclusion criteria) it cannot be ruled out, however, that the respondents are a somewhat biased sample of the entire scientific community of applied ethologists. Another limitation of the study is that respondents before versus after the workshop and plenary may not be a comparable sample of the population. Indeed, the after sample was limited to those congress delegates that had chosen to participate in this particular workshop on expectation bias rather than in other simultaneous workshops or other activities. Nevertheless we believe that differences in responses before versus after the workshop reflect the effect of the combined plenary talk and workshop rather than a sampling bias. Indeed, these differences remained when the analyses were restricted to the 14 respondents who had filled out the questionnaire both before and after the plenary talk and workshop. Although caution is warranted to generalize findings to the field of applied ethology at large, the results suggest that a moderate awareness about expectancy effects among ethologists and the logistic constraints of blinded observations rather than a low susceptibility of the research field is the larger part of the explanation of why blinding is not more common.

The moderate awareness is illustrated by the rather neutral scores (i.e. close to the mid-point of the scale) before the plenary and workshop for the perceived susceptibility to expectation bias of (1) the various observer-based recording methods commonly used in applied ethology, of (2) the respondents' own research outcomes and that of their peers, and of (3) the various research situations classified according to subjectivity and strength of a-priori expectations. Only the estimated 30% of the research situations where the data-collector uses subjective methods and has strong expectation about the outcome were perceived to be quite highly susceptible to expectancy effects. But even for such research situations the respondents seemed quite accommodating for non-blinding as a journal reviewer or editor.. Most respondents acknowledged, though, that it is important to take actions to minimize expectation bias in their next experiment or study (with an average score well above the neutral point of the scale). The significantly higher scores on all these questions after the plenary and workshop, illustrate the room and potential to increase the awareness among applied ethologists about the risk of expectancy. It should be emphasized, however, that we have demonstrated an immediate short-term effect of the plenary and workshop, but that it is not known how long this effect lasts. Probably, larger and more sustained concerted initiatives will be needed to raise the awareness of this scientific discipline in the long-term. These may include, for example, an increased emphasis on expectancy effects in textbooks for and the training

of applied ethologists, further illustrations of the existence and consequences of expectation bias in ethological research outcomes, and critical evaluations of the potential of such bias in the design of experiments, in the assessment of research proposals and in the reviewing process of publication of research papers (Kardish et al., 2015).

The apparent reluctancy to blind outcome assessors in applied ethological research seems at odds with the perceived high susceptibility of this scientific discipline to expectancy effects. The respondents reported that the methods used in applied ethology are on average more subjective, and the data-collectors are more likely to have strong expectations about the research outcome, as compared to other scientific disciplines such as animal physiology. Subjective recording methods (i.e. requiring human judgment) and strong research expectations are widely considered as risk factors for expectation bias (Bello et al., 2014; Bohlen et al., 2014; Marsh and Hanlon, 2007; Rosenthal, 1966; Savović et al., 2012; Tuyttens et al., 2014; van Wilgenburg and Elgar, 2013). The respondents estimated that only a small share (ca. 18%) of research in applied ethology involves objective methods and assessors without expectations about the outcome.

It is surprising, therefore, that only a minority of the published animal behaviour studies report to have avoided expectation bias by blinding outcome assessors (Burghardt et al., 2012; Kardish et al., 2015; von Wilgenburh and Elgar, 2014). The aforementioned limited awareness about expectancy effects (the so-called bias blind spot) could be part of the explanation. Before the plenary and workshop respondents indicated to be quite tolerant about non-blinded outcome assessors when reviewing research papers, and particularly so when objective methods are used. After the plenary and workshop, the respondents were less accommodating but only for studies using subjective methods and with research staff having strong expectations about the outcome. Another part of the explanation could relate to the poor feasibility of blinding outcome assessors in many types of ethological studies. Indeed blinding may be more easy to achieve in pharmacological trials (usually by using a placebo that looks similar to the drug being tested) than non-pharmacologic trials, partly because of the challenge of masking perceptible physical properties of the treatments (Bello et al., 2014; Boutron et al., 2004, 2007). Although modern multimedia techniques may enable blinding in most studies (e.g. using video recordings to observe behaviour), these techniques will usually increase cost and logistical complexity with no guarantee that it will make an important difference in any single study. This is reflected in the respondents' rather low score for the perceived feasibility to prevent expectation bias in their next experiment. Whereas blinding outcome assessors was regarded as the most effective method for reducing expectation bias in the respondents' own ethological research, its feasibility was scored much lower (in particular before the plenary and workshop).

When blinded outcome assessment is not feasible, it is important to ensure that outcomes are as robust as possible to the lack of blinding. This can be achieved by debiasing the outcome assessment method (or the decision-making task) or by debiasing the research staff (or decision-maker) (Reese, 2012). Methods to debias the assessment method may focus on modifying the outcome definition or method of assessment to minimize the subjective elements (Kahan et al., 2014; Savović et al., 2012), or on reducing the complexity, obscurity and ambiguity of the outcome assessment (Lerman et al., 2010; Nakhaeizadeh et al., 2014; Page et al., 2012). The respondents in the present study considered using objective methods and unambiguous definitions to categorize behaviours as nearly as effective, but more feasible, than blinding outcome assessors. The extent to which subjective assessments in ethological research can be fully replaced by objective measures likely depend on the nature of the study, and further clarification of the precise relation between expectancy effects and outcome assessment subjectivity, ambiguity, and obscurity is warranted.

One of the techniques for debiasing the decision-makers that was given a rather high score for both feasibility and effectiveness –when surveyed after the plenary and workshop in particular - concerned raising awareness of research staff about expectation bias. This technique aims to reduce decision-makers' vulnerability to biases through an improved understanding of underlying decision mechanisms and how biases can affect the decision-making process. Empirical tests of the effectiveness of this debiasing technique in other fields of research has had mixed results (Leddy et al., 2013; Kenyon, 2014; Maynes, 2015; Reese, 2012). Randomly assigning or balancing (according to outcome expectations) data-collectors to experimental treatments were considered the least effective and the least feasible debiasing techniques when surveyed before the plenary and workshop. After the plenary and workshop, both effectiveness and feasibility were scored higher. Depending on the type of study, comprehensively testing all possible research outcome expectations that may induce biased assessment outcomes is likely to be a daunting task indeed. Randomly assigning different data-collectors to different treatments is probably only feasible in large research teams with a substantial pool of data-collecting staff so that the amount of bias due to personal expectations from each single assessor is minimal in relation to the entire dataset . The effectiveness of these techniques is largely unknown.

Using highly experienced/trained data-collectors was ranked medium for both effectiveness and feasibility. There is some evidence from psychological research that assessor training can be effective in reducing bias, at least for outcome assessments that require human inference (Hoyt and Kerns, 1999). Caution may be warranted though because training/experience may increase the assessors' stake in the outcome of their assessments. Although inter-observer

agreement may be enhanced by experience and training, this improvement may be at the expense of accuracy (Bernardin and Pence, 1980) and is no guarantee of absence of expectation bias (Hróbjartsson et al., 2013; Tuyttens et al., 2014).

Not only outcome assessors, but also data-analysts may be prone to expectancy effects. The respondents considered blinding data-analysts as a fairly effective debiasing technique. The feasibility was considered rather low which is perhaps surprising because in principle data-analysts can almost always be blinded (Polit, 2011). Blinding data-analysts seldom occurs, perhaps because of misconceptions about the objectivity of statistical analysis (Polit, 2011).

5. Conclusions and recommendations

For the sake of the credibility of the scientific field, we recommend applied ethologists to become more aware of, and to prevent, expectancy bias affecting their research outcomes. Expectancy effects should be considered throughout all stages of a research project: from the evaluation of research proposals for funding, through the planning of the study design, to the collection, analyses, interpretation and reporting of data. In particular when outcome measures are subjective and when it's likely that outcome assessors and data-analysts have expectations or a vested interest in the research outcome, they should be blinded (and the effectiveness of the blinding should be tested) if possible. If blinding is not feasible, subjective and ambiguous elements requiring human judgement in the outcome assessment should be removed as much as possible. If that is not possible either, other – often less effective – debiasing techniques should be considered. Editors and reviewers ought to demand information and set standards regarding the likelihood of expectation bias when deciding whether or not to publish submitted manuscripts. Fellow ethologists and other users of scientific reports should give more credence to research outcomes that could not have been biased by expectancy effects.

Acknowledgments

We thank the organisers of ISAE-2014 for their support with organizing this survey and the congress attendants who filled out the survey and joined the workshop.

References

Begley, C.G. and Ellis, L.M. (2012). Drug developments: raise standards for preclinical cancer research. Nature, 483, 531-533.

Bello S., Grogsbøll, L.T., Gruber, J., Zizhuang, J.Z., Fisher, D. and Hróbjartsson, A. (2012) Lack of blinding of outcome assessors in animal model experiments implies risk of observer bias. Journal of Clinical Epidemiology, 67, 973-983.

Bello S., Moustgaard, H. and Hróbjartsson, A. (2014) The risk of unblinding was infrequently and incompletely reported in 300 randomized clinical trial publications. Journal of Clinical Epidemiology, 67, 1059-1068.

Bernardin, H.J. and Pence, E.C. (1980). Effects of rater training: creating new response sets and decreasing accuracy. Journal of Applied Psychology, 65, 60-66.

Bohlen, M., Hayes, E.R., Bohlen, B., Bailoo, J .D., Crabbe, J.C. and Wahlsten, D. (2014). Experimenter effects on behavioral test scores of eight inbred mouse strains under the influence of ethanol. Behavioural Brain Research, 272, 46-54.

Boutron, I., Guittet, L., Estellat, C., Mohar, D., Hróbjartsson, A., and Ravaud, P. (2007). Reporting methods of blinding in randomized trials assessing non-pharmacological treatments: A systematic review. PLoS Medicine, 4, e61.

Boutron, I., Tubach, F., Giraudeau, B. and Ravaud, P. (2004) Blinding was judged more difficult to achieve and maintain in nonpharmacologic than pharmacologic trials. Journal of Clinical Epidemiology, 57, 543-550.

Burghardt, G. M., Bartmess-LeVasseur, J. N., Browning, S. A., Morrsion, K. E., Stec, C. L., Zachau, C. E., et al. (2012). Perspectives – minimizing observer bias in behavioral studies: A review and recommendations. Ethology, 118, 511-517.

Croskerry, P., Singhal G. and Mamede, S. (2013a). Cognitive debiasing 1: origins of bias and theory of debiasing. British Medical Journal Quality and Safety, 22, i58-i64.

Croskerry, P., Singhal G. and Mamede, S. (2013b). Cognitive debiasing 2: impedements to and strategies for change. British Medical Journal Quality and Safety, 22, i65-i72.

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. PloS ONE, 5, e10068.

Fanelli, D. and Ioannidis (2013). US studies may overestimate effect sizes in softer research. Proceedings of the National Academy of Sciences, 110, 15031-36.

Hoyt, W. T., and Kerns, M.-D. (1999). Magnitude and moderators of bias in observer ratings: a meta-analysis. Psychological Methods, 4, 403-424.

Hróbjartsson, A., Thomsen, A. S. S., Emanuelsson, F., Tendal, B., Hilden, J., Boutron, I., et al. (2012). Observer bias in randomised clinical trials with binary outcomes: A systematic review of trails with both blinded and non-blinded outcome assessors. British Medical Journal, 344, e1119.

Hróbjartsson, A., Thomsen, A. S. S., Emanuelsson, F., Tendal, B., Hilden, J., Boutron, I., et al. (2013). Observer bias in randomized clinical trials with measurement scale outcomes: A systematic review of trails with blinded and non-blinded assessors. Canadian Medical Association Journal (early release: DOI:10.1503/cmaj.120744)

Ioannidis, J.P.A. (2005). Why most published research findings are false. Plos Medicine, 2, e124.

Ioannidis, J.P.A. (2008). Why most discovered true associations are inflated. Epidemiology, 19, 640-648.

Kahan, B.C., Cro, S., Doré, C., Bratton, D.J., Rehal, S., Maskell, N.A., Rahman, N. and Jairath, V. (2014). Reducing bias in open-label trials where blinded outcome assessment is not feasible: strategies from two randomised trials. Trials, 15, 456.

Kaptchuk, T. J. (2001). The double-blind, randomized, placebo-controlled trial: gold standard or golden calf? Journal of Clinical Epidemiology, 54, 541-549.

Kardish, M.R., Mueller, U.G., Amador-Vargas, S., Dietrich, E.I., Ma, R., Barrett, B. and Fang C-C. (2015). Blind trust in unblinded observation in ecology, evolution, and behavior. Frontiers in Ecology and Evolution, 3, 51.

Kenyon, T. (2014). Critical thinking education and debiasing. Informal Logic, 34, 341-363.

Kilkenny, C., Browne, W.J., Cuthill, I.C., Emerson, M., and Altman, D.G. (2010). Improving bioscience research reporting: the arrive guidelines for reporting animal research. PLoS Biology, 8, e1000412.

Leddy, M.A., Anderson, B.L., and Schulkin, J. (2013). Cognitive-behavioral therapy and decision science. New Ideas in Psychology, 31, 173-183.

Lerman, D. C., Tetreault, A., Hovanetz, A., Bellaci, E., Miller, J., Karp, H., et al. (2010). Applying signal-detection theory to the study of observer accuracy and bias in behavioural assessment. Journal of Applied Behaviour Analysis, 43, 195-213.

Marsh, D. M., and Hanlon, T. J. (2004). Observer gender and observer bias in animal behaviour research: Experimental tests with red-backed salamanders. Animal Behaviour, 68, 1425-1433.

Marsh, D. M., and Hanlon, T. J. (2007). Seeing what we want to see: confirmation bias in animal behavior research. Ethology, 113, 1089-1098.

Maynes, J. (2015). Critical thinking and cognitive bias. Informal Logic, 35, 183-203.

Miller, L. E., and Stewart, M. E. (2011). The blind leading the blind: use and misuse of blinding in randomized controlled trials. Contemporary Clinical Trials, 32, 240-243.

Nakhaeizadeh, S., Dror, I.E., and Morgan, R.M. (2014). Cognitive bias in forensic anthropology: visual assessment of skeletal remains is susceptible to confirmation bias. Science & Justice, 54, 208-214.

Page, M., Taylor, J., and Blenking, M. (2012). Context effects and observer bias: Implications for forensic odontology. Journal of Forensic Sciences, 57, 108-112.

Polit, D.E. (2011). Blinding during the analysis of research data. International Journal of Nursing Studies, 48, 636-641.

Reese, E.J. (2012). Techniques for mitigating cognitive biases in fingerprint identification. UCLA Law Review, 59, 1252-1290.

Rosenthal, R. (1966). Experimenter effects in behavioral research. New York, USA: Appleton-Century-Crofts.

Savović, J., Jones, H.E., Altman, D.G., Harris, R.J., Jüni, P., Pildal, J. et al. (2012). Influence of reported study design characteristics on intervention effect estimates from randomized, controlled trials. Annals of Internal Medicine, 157, 429-438.

Schulz, K. F., Chalmers I., Hayes, R. J., and Altman, D. G. (1995). Empirical evidence of bias: Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. Journal of the American Medical Association, 273, 408-412.

Schulz, K. F. and Grimes, D.A. (2002) Blinding in randomized trials: hiding who got what. The Lancet, 359, 969-700.

Tuyttens, F.A.M., de Graaf, S., Heerkens, J.L.T., Jacobs, L., Nalon, E., Ott, S., Stadig, L., Van laer, E. and Ampe, B. (2014) Observer bias in animal behaviour research: can we believe what we score , if we score what we believe? Animal Behaviour, 60, 273-280.

Van Wilgenburg, E., and Elgar, M.A. (2013) Confirmation bias in studies of nestmate recognition: a cautionary note for research into the behaviour of animals. PLoS One, 8, e53548.

Table 1

The number of respondents per profession who had filled out the survey either before, after or before and after the plenary and workshop on expectation bias. The number of non-anonymous respondents that have published a scientific peer-reviewed paper on ethology listed in the Web of Knowledge is indicated as well.

Moment	Profession	Number	Number with at least 1 ethological A1-publication
Before	Researcher[1]	23	22
	Students[2]	13	10
	Other[3]	3	0
	Unknown[4]	0	NA
After	Researcher	25	21
	Students	13	8
	Other	0	0
	Unknown	13	NA
Before & After	Researcher	11	11
	Students	3	3
	Other	0	0
	Unknown	0	NA

[1] Includes post-docs, research associates, (assistant) professors

[2] Includes MSc and PhD students

[3] Includes a veterinarian, a humane standards officer and a director of an animal shelter

[4] Includes anonymous respondents who had not filled out their name and profession

Table 2

Four different theoretical research situations based on whether or not the researcher/data-collector uses objective (Obj) versus subjective (Sub) methods and has strong (Exp) versus no expectations (NoE) about the research outcome.

	Objective methods	Subjective methods
Data-collector has strong expectations	Obj-Exp	Sub-Exp
Data collector has no expectations	Obj-NoE	Sub-NoE

Table 3

Respondents' perceived prevalence of research situations (Obj-Exp, Sub-Exp, Obj-NoE, Sub-NoE, see Table 1) in applied ethology versus other scientific disciplines.

	Applied Ethology least square mean (SD)	Other least square mean (SD)	$\lvert t_{83} \rvert$	P
Obj-Exp	40.6 (19.8)	48.4 (19.6)	4.2	<0.001
Sub-Exp	29.8 (17.7)	16.9 (10.6)	7.8	<0.001
Obj-NoE	17.9 (14.1)	23.3 (14.3)	3.9	<0.001
Sub-NoE	13.9 (8.9)	12.1 (10.6)	1.9	0.058

Fig. 1. Least squares mean (± SE) scores on a 1-10 scale of the respondents' perceived susceptibility to expectation bias of five types of observation-based recordings before (n=39) and after (n=51) the plenary and workshop: occurrence of various behaviours (ethogram), quantifying the severity of physical or clinical conditions (physical), interpreting the outcome of behavioural interactions (outcome), characterizing animal personalities (personality), and assessing emotional state (emotion). Scores for observation-based recordings without a common letter (a-c) in between brackets on the Y-axis differ significantly. Asterisks indicate scores that differ significantly before versus after the plenary and workshop.

Fig. 2. Least squares mean (± SE) scores on a 1-10 scale of the respondents' perceived susceptibility to expectation bias of research situations (Obj-Exp, Sub-Exp, Obj-NoE, Sub-NoE, see table 1) before (n=39) and after (n=51) the plenary and workshop. Scores for research situations without a common letter (a-c) in between brackets on the Y-axis differ significantly. Asterisks indicate scores that differ significantly before versus after the plenary and workshop.

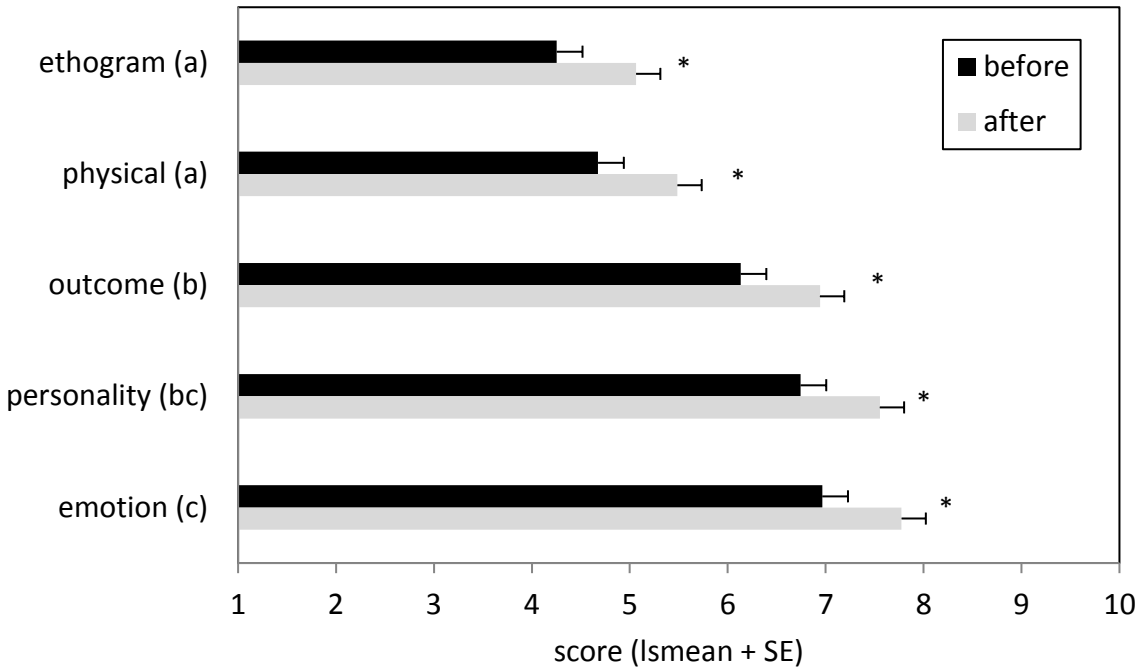
Fig. 3. Least squares mean (± SE) scores on a 1-10 scale of the influence of not blinding data-recorders on the respondents' reported decision as an editor/reviewed against or in favour of publication depending on (i) the research situation (Obj-Exp, Sub-Exp, Obj-NoE, Sub-NoE, see table 1), and on (ii) the timing of the survey (before or after the plenary and workshop). Scores for research situations without a common letter to the right of the black (a-b) and grey histograms (x-z) differ significantly between research situations. Asterisks indicate scores that differ significantly before (n=39) versus after (n=51) the plenary and workshop.

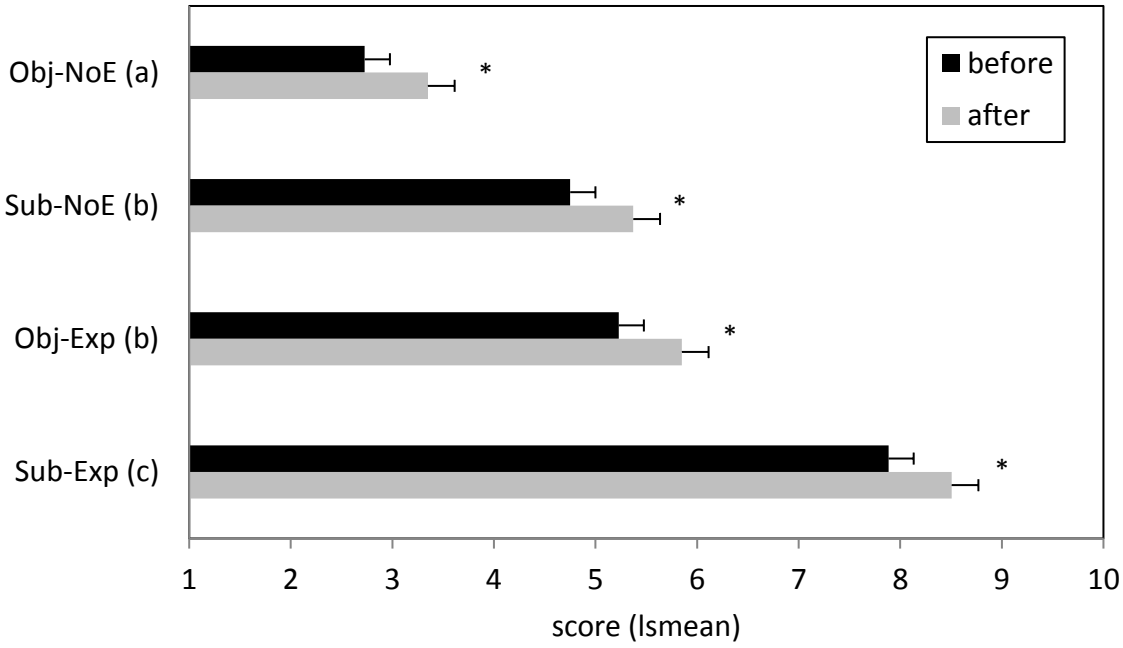
Fig. 4.. Least squares mean (± SE) scores on a 1-10 scale of the perceived influence of expectation bias on the outcomes of the respondents' own research and that by their peers depending on the timing of the survey (before and after the plenary and workshop). Lack of a common letter (a-b) in between brackets on the Y-axis indicates significant difference between scores for the respondents' own research versus that of peers. Asterisks indicate scores that differ significantly before (n=39) versus after (n=51) the plenary and workshop.

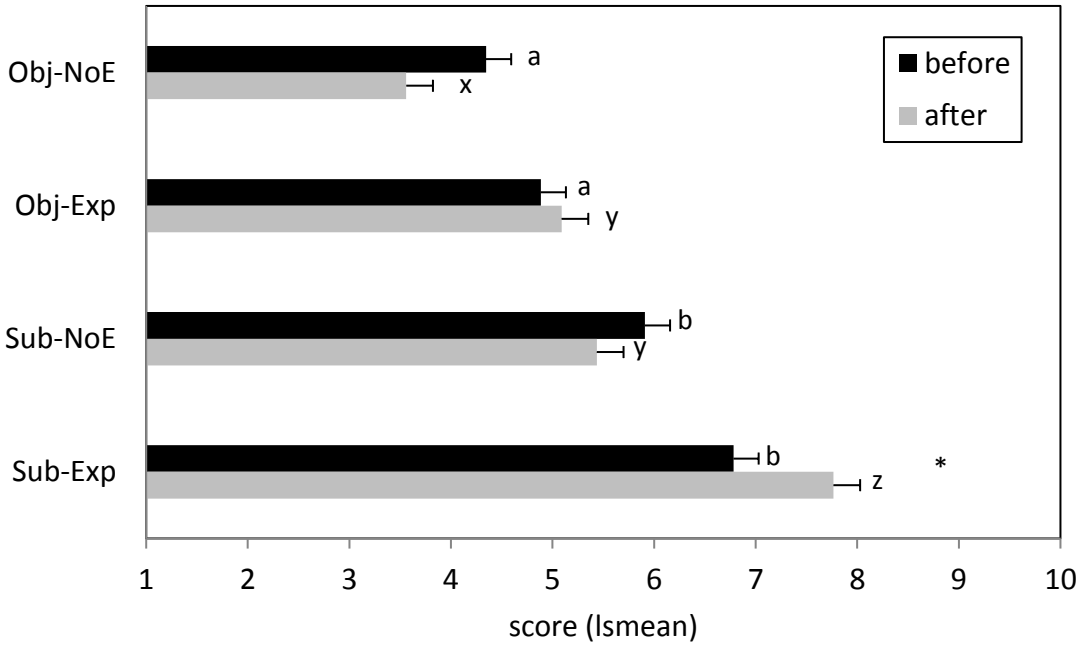
Fig. 5. Least squares mean (± SE) scores on a 1-10 scale of the respondents' perceived importance to take actions to minimize, and perceived feasibility to effectively reduce expectation bias in their next experiment depending on the timing of the survey (before and after the plenary and workshop). Lack of a common letter (a-b) in between brackets on the Y-axis indicates significant difference between scores for perceived importance and feasibility. Asterisks indicate scores that differ significantly before (n=39) versus after (n=51) the plenary and workshop.

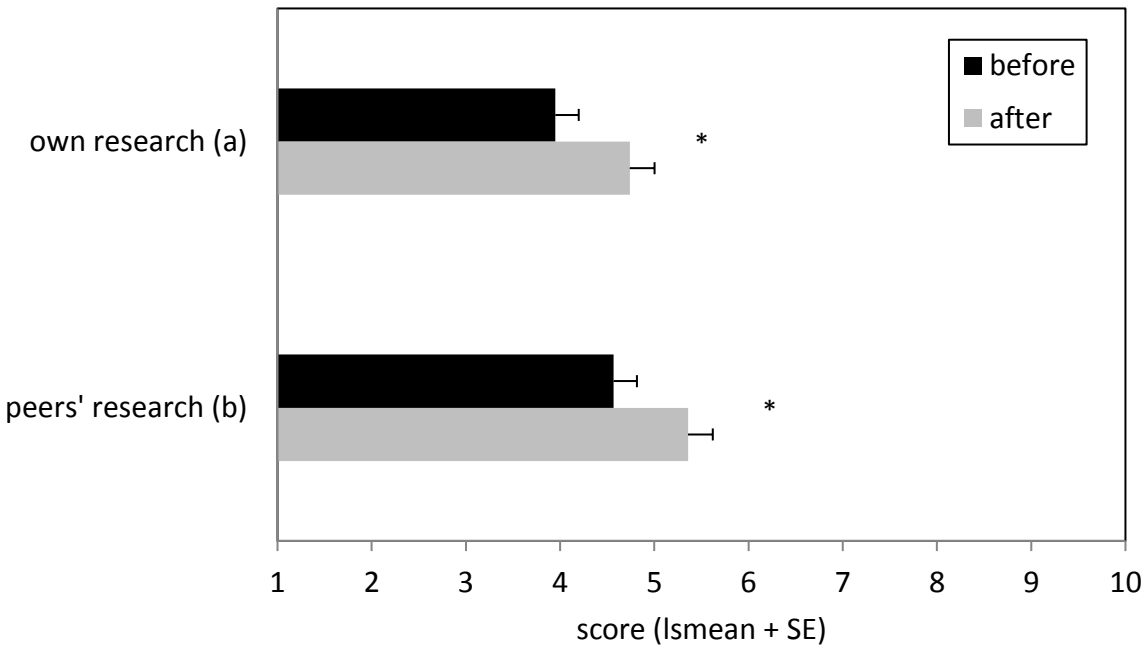
Fig. 6. Least squares mean (± SE) scores on a 1-10 scale of the perceived (a) effectiveness and (b) feasibility of various methods for reducing expectation bias in the respondents' own applied ethological research depending on the timing of

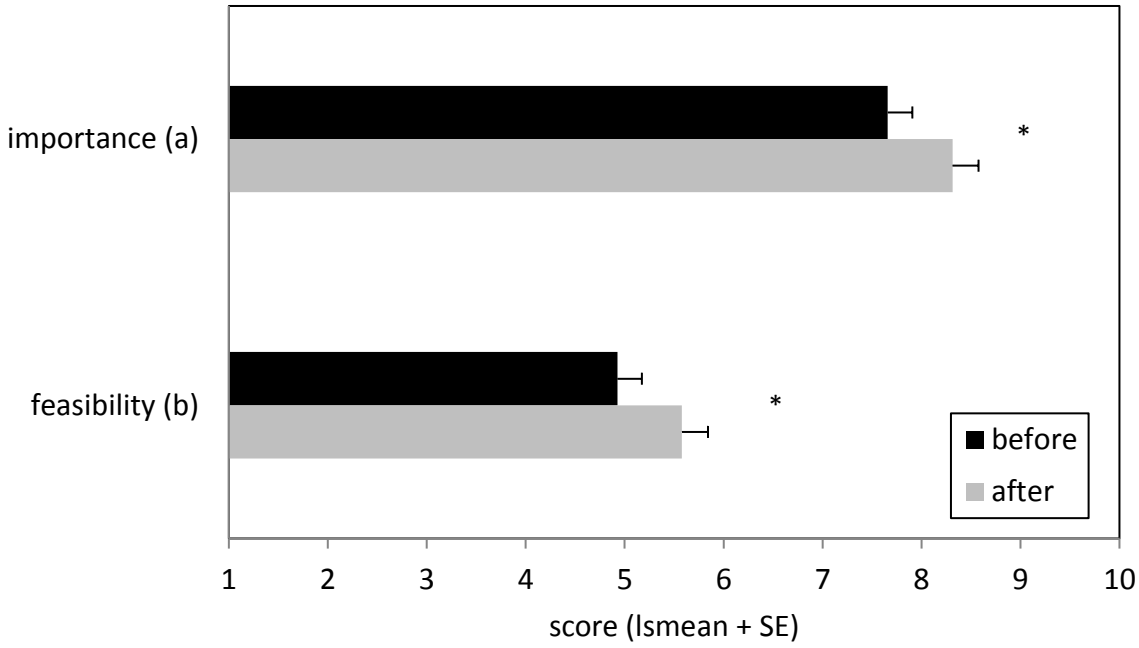
the survey (before or after the plenary and workshop). Scores for debiasing methods without a common letter to the right of the black (a-c) and grey bars (w-z) differ significantly. Asterisks indicate scores that differ significantly before (n=39) versus after (n=51) the plenary and workshop.


ethogram (a)
physical (a)
outcome (b)
personality (bc)
emotion (c)
before
after
*
*
*
*
*
1
2
3
4
5
6
7
8
9
10
score (lsmean + SE)

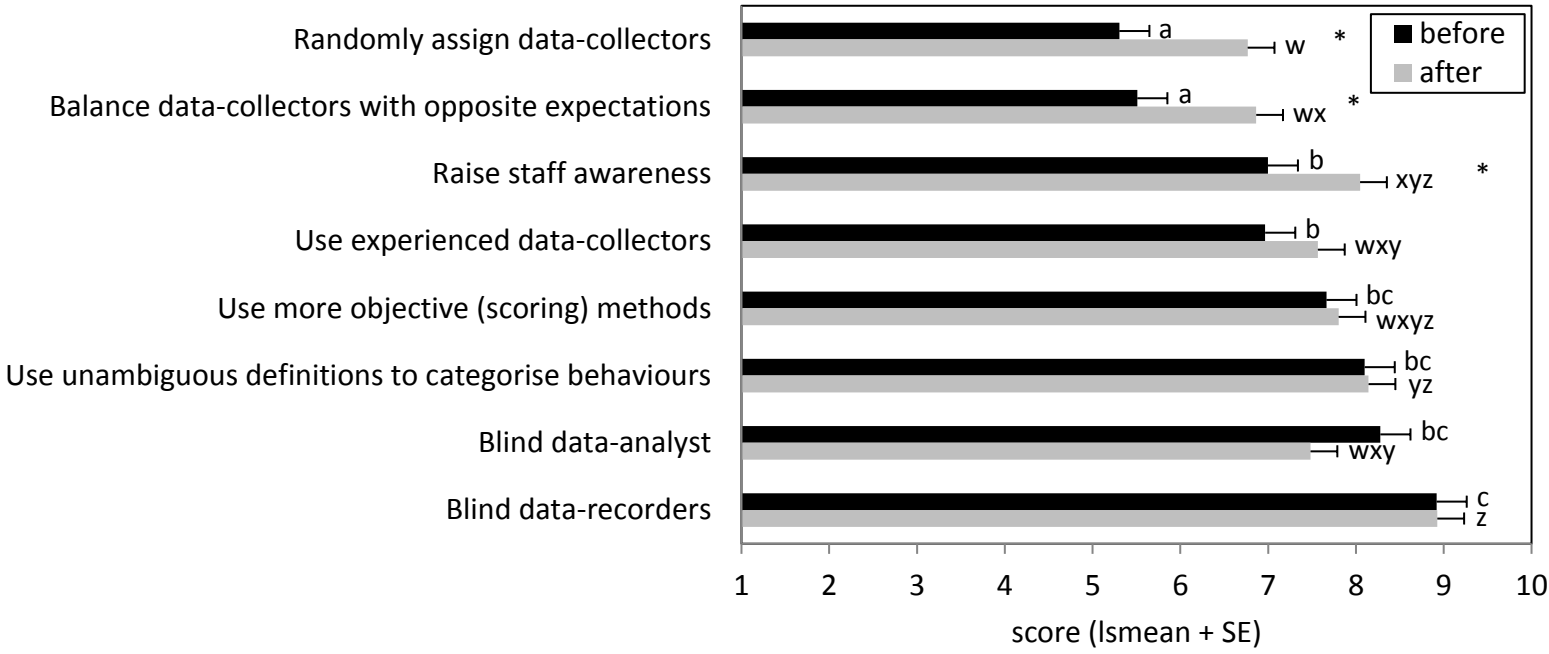

Obj-NoE (a)
Sub-NoE (b)
Obj-Exp (b)
Sub-Exp (c)
before
after
*
*
*
*
1
2
3
4
5
6
7
8
9
10
score (lsmean)


Obj-NoE
Obj-Exp
Sub-NoE
Sub-Exp
a
x
a
y
b
y
b
z
*
before
after
1
2
3
4
5
6
7
8
9
10
score (lsmean)

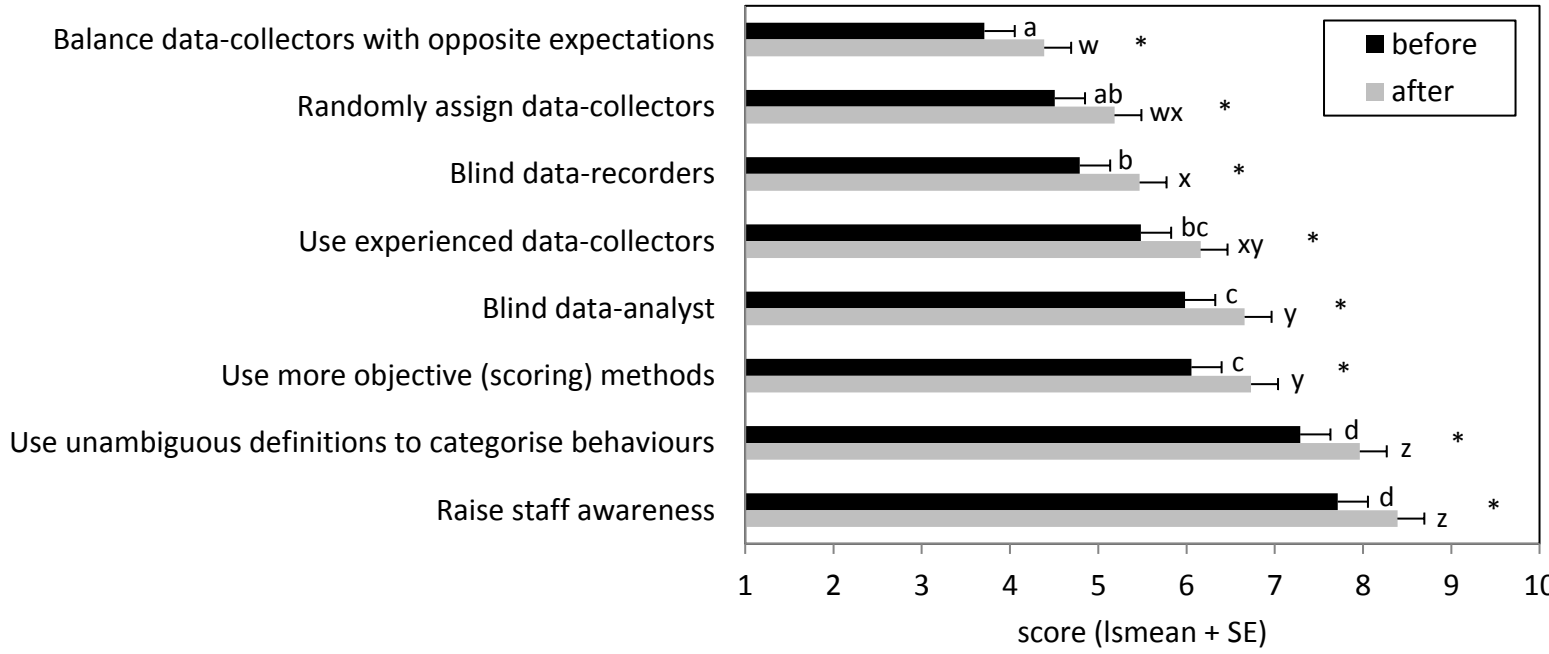

before
after
own research (a)
peers' research (b)
*
*
1
2
3
4
5
6
7
8
9
10
score (lsmean + SE)


importance (a)
feasibility (b)
*
*
before
after
1
2
3
4
5
6
7
8
9
10
score (lsmean + SE)

a


Randomly assign data-collectors
Balance data-collectors with opposite expectations
Raise staff awareness
Use experienced data-collectors
Use more objective (scoring) methods
Use unambiguous definitions to categorise behaviours
Blind data-analyst
Blind data-recorders
a
w
*
a
wx
*
b
xyz
*
b
wxy
bc
wxyz
bc
yz
bc
wxy
c
z
before
after
1
2
3
4
5
6
7
8
9
10
score (lsmean + SE)

b


Balance data-collectors with opposite expectations
Randomly assign data-collectors
Blind data-recorders
Use experienced data-collectors
Blind data-analyst
Use more objective (scoring) methods
Use unambiguous definitions to categorise behaviours
Raise staff awareness
a
w
*
ab
wx
*
b
x
*
bc
xy
*
c
y
*
c
y
*
d
z
*
d
z
*
before
after
1
2
3
4
5
6
7
8
9
10
score (lsmean + SE)