International Journal of Psychophysiology 119 (2017) 93–107

Contents lists available at ScienceDirect

International Journal of Psychophysiology

journal homepage: www.elsevier.com/locate/ijpsycho

Computational reproducibility of "Goal relevance and goal conduciveness appraisals lead to differential autonomic reactivity in emotional responding to performance feedback" (Kreibig, Gendolla, & Scherer, 2012): A guide and new evidence☆

Sylvia D. Kreibig

Department of Psychology, Stanford University, Stanford, USA

ARTICLE INFO

Keywords:
Computational reproducibility
Replication
Psychophysiology
Motivation
Emotion
Appraisal

ABSTRACT

The emerging field of the psychophysiology of motivation bears many new findings, but little replication. Using my own data (Kreibig, Gendolla, & Scherer, 2012), I test the reproducibility of this specific study, provide the necessary materials to make the study reproducible, and instantiate proper reproducibility practices that other researchers can use as a road map toward the same goal. In addition, based on re-analyses of the original data, I report new evidence for the motivational effects of emotional responding to performance feedback. Specifically, greater appraisal of goal relevance amplifies the emotional response to events appraised as conducive (i.e., effort mobilization), but not to those appraised as obstructive to a person's goals (i.e., effort withdrawal). I conclude by providing a ten-step road map of best practices to facilitate computational reproducibility for future studies.

> *An article about a computational result is advertising, not scholarship. The actual scholarship is the full software environment, code, and data, that produced the result.* (p. 385, Donoho, 2010, *paraphrasing* Claerbout and Karrenbach, 1992)

1. Introduction

The emerging field of the psychophysiology of motivation bears many new findings, but little replication. As one example, Kreibig et al. (2012) recently reported that emotions experienced in response to performance feedback after task execution have motivational effects that guide subsequent engagement with the environment. However, this finding has not been tested for reproducibility. The psychophysiology of motivation appears to lack tests for reproducibility, the necessary materials to make their studies reproducible, and the resources that may enable researchers to learn practices that optimize reproducibility. To fill this gap, I am using my own data to test the reproducibility of the above example (Kreibig et al., 2012, Sections 3–7 and 8.2), provide the necessary materials to make the study reproducible (Sections 2.2 and 8.3), and instantiate proper reproducibility practices that other researchers can use as a road map toward the same goal (Section 8.4). In addition, based on re-analyses of the original data, I report new evidence for the motivational effects of emotion in response to performance feedback (Sections 7 and 8.1).

Our data example (Kreibig et al., 2012) integrated predictions about how appraisals affect emotion (Scherer, 2001) with motivational interpretations of autonomic reactivity (Wright, 1996). Emotions are multicomponential responses that consist of coordinated changes, including subjective feeling and physiology (Scherer, 2001). Emotions are elicited by appraisals, which evaluate the meaning of a stimulus event for an individual's goals and determine the emotional response (Scherer, 2001). As two key appraisals, goal conduciveness evaluates whether a given event is consistent (i.e., conducive) or inconsistent (i.e., obstructive) with an individual's goals, and goal relevance evaluates the extent to which an event is relevant to an individual's goals (Scherer, 2001). Appraised conduciveness has been hypothesized to lead to positive emotions and appraised obstructiveness to negative emotions, whereas greater appraised relevance may cause the respective emotion to be experienced more intensely (e.g., Scherer, 2001). In short, goal relevance is hypothesized to moderate the effect of goal conduciveness on emotional responding.

☆ I thank Gunnar Schaefer for the helpful discussions and support in creating the Docker containerization of the computational environment. Research reported in this publication was supported by Université de Montréal BC00114115 and the National Institute Of Dental & Craniofacial Research of the National Institutes of Health under Award Number R56DE025321. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Address correspondence to: Sylvia Kreibig, Department of Psychology, Stanford University, 450 Serra Mall, Stanford, CA 94305, USA.
E-mail address: skreibig@stanford.edu.

http://dx.doi.org/10.1016/j.ijpsycho.2017.06.001
Received 31 January 2017; Received in revised form 30 May 2017; Accepted 1 June 2017
Available online 28 July 2017

Kreibig et al. (2012) tested this hypothesis using an achievement paradigm (Nummenmaa and Niemi, 2004) in a between-participants laboratory experiment. They manipulated goal conduciveness by varying the valence of performance feedback, i.e., success or failure. They manipulated goal relevance by presenting the task either as a validated psychological test of the participant's emotional intelligence (high relevance) or as the evaluation of a picture set in the context of a pilot study (low relevance). They measured emotion by assessing self-reported emotional feelings after and autonomic reactivity during feedback presentation.

Kreibig et al. (2012) reported that feeling self-report showed effects of conduciveness, with greater contentment, happiness, joy, and pride and less disappointment for conducive than obstructive conditions, but no interaction with relevance. Physiological reactivity showed the predicted interaction of goal relevance and goal conduciveness on cardiac autonomic regulation (CAR)—a key index of autonomic control—that indicated sympathetic–parasympathetic co-activation under high-relevance goal conduciveness and sympathetic–parasympathetic co-inhibition under high-relevance goal obstruction. Moreover, greater mean arterial pressure (MAP) and skin conductance level (SCL) were found for conducive than obstructive conditions and lower heart rate (HR) and greater SCL for high than low relevance conditions. This response pattern was interpreted as reflecting a physiologically activating engagement response under positive achievement emotions and a physiologically deactivating disengagement response under negative achievement emotions. The original study concluded that results suggested weak support for the interaction hypothesis of goal relevance and goal conduciveness. Given their novelty, these results need to be reproduced.

Reproducibility—as one of the strongest norms of the research community—states that new experimental findings are only accepted as true once they have been independently verified by another group of researchers (Collins, 1992; McKubre, 2008). Reproducibility consists of a three-layered structure: On the top layer lies *replication*: The original experiment is repeated by an independent researcher to generate results that allow one to reach the same conclusions as those reported in the original experiment (Ioannidis and Khoury, 2011; Peng, 2009). It requires new data collection with new research subjects and identical or similar methods and analyses (Schmidt, 2009). On the middle layer lies *empirical reproducibility*: The original experiment is described in sufficient detail to allow an independent researcher to reproduce the original experiment (Bollen et al., 2015; Schwalbe et al., 2016; Stodden, 2015). It requires specification of an experiment's conditions, parameters, and equipment. On the bottom layer lies *computational reproducibility*: Sufficient information is provided to allow an independent researcher to run the same analyses and obtain the exact same results as those reported in the original experiment (Ioannidis and Khoury, 2011). It requires making the original data, analysis scripts, and computational environment (i.e., the specific instantiation and configuration of the software application used to run the analyses) available.

While attempts of replication studies and the requirement of empirical reproducibility have been part of science for centuries, computational reproducibility is a newer construct that has received little attention. Because the specific way of computing analyses can significantly change the results obtained, computational reproducibility represents a fundamental requirement to ultimately achieve replication. I here focus on computational reproducibility.

I suggest that computational reproducibility comprises five steps. I introduce relevant terminology in relation to each step in what follows. As a convenient reference for readers, I present a glossary of terms for computational reproducibility in Table 1 (italicized words mark terms defined in the glossary). Using our data example (Kreibig et al., 2012), I then illustrate each step of computational reproducibility in the following sections.

The first step establishes *computational documentation and reproduction* (Chirigati et al., 2013; Leek and Peng, 2015; Peng et al., 2006; Peng, 2011; Stodden, 2015): A record of the exact type, implementation, and configuration of the computations performed to arrive at the original results is created (Donoho et al., 2009, Section 3.1). This includes the original results (as a necessary basis for comparing original and reproduced results) and the *computational compendium* (as a sufficient basis for regenerating original results). *Computational reproduction* is the actual test of *computational reproducibility*, i.e., the attempt to repeat and replicate the computations carried out to arrive at

Table 1
Glossary of terms for computational reproducibility. Italicized words in the definition mark terms further defined in the glossary.

Term	Definition
Analysis script	Record of the exact type, implementation, and configuration of the analyses carried out to arrive at the original results.
Computational accessibility	Capacity of the *analysis script* to be read by members of the scientific community, achieved by writing *analysis scripts* in a scripting language of common use in the scientific community.
Computational accuracy	The degree to which the correct result is represented in the computational outcome (i.e., precision). Cf. *computational correctness*.
Computational compendium	Collection of the data, *analysis scripts*, and *computational environment* used to arrive at the original results.
Computational correctness	Whether the right analyses were chosen and were configured and implemented in the right way to reach the right computational outcome (i.e., free from error). Cf. *computational accuracy*.
Computational documentation	Combination of the original results together with the *computational compendium*.
Computational environment	"A computer system that provides all the computational facilities necessary to solve a target class of problems" (Gallopoulos et al., 1994). It comprises a software application, typically supplemented by modular plug-ins or libraries, that provides the basic structure to analyze and visualize the problem. It supports a complete scientific work flow and enables scientists to design, build, and deliver scientific products all within one framework (Sloan et al., 2013).
Computational explorability	Capacity of the data to answer novel research questions that have not been addressed in prior analyses. This may involve applying a different type of analysis that brings together two or more variables not previously considered jointly.
Computational extensibility	Capacity of the *analysis script* to be expanded to include additional dependent variables or analyses to enhance the conclusions of existing analyses.
Computational flexibility	Capacity of an *analysis script* to be easily modified, such as reconfiguring analyses or figures according to other researchers' preferences or alternative foci. Requires *computational documentation* and *computational accessibility*.
Computational modifications	Minor or moderate changes in analytic procedures and assumptions.
Computational reproducibility	Provision of sufficient information to allow the replication of the computations carried out to arrive at the original results.
Computational reproduction	Actual test of *computational reproducibility*, i.e., the attempt to repeat and replicate the computations carried out to arrive at the identical results as originally reported.
Computational reviewability	Provision of complete *computational documentation* for *computational review*.
Computational review	Examination of the computational documents to evaluate *computational veracity*. Requires complete *computational documentation*.
Computational revisability	Capacity to incorporate changes in an existing *analysis script*. Requires *computational accessibility*.
Computational robustness	*Computational modifications* should be tolerated without significant change in the quantitative findings or qualitative conclusions.
Computational veracity	Evaluation of the original and reproduced results for *computational accuracy and correctness*. Discrepancies highlight deviations from *computational accuracy and correctness*, either in the original or reproduced results (or both), and allow for their correction.

the identical results as originally reported (Sections 3.2–3.5).

In a second step, provision of complete *computational documents* establishes *computational reviewability* (Section 4). *Computational review* evaluates *computational veracity* (Section 4.1). Authors, editors, reviewers, and peers compare original and reproduced results, examine the *computational compendium*, and document and correct deviations. Because *analysis scripts* need to be read by members of the scientific community, *computational accessibility* is a prerequisite, which is achieved by writing *analysis scripts* in a scripting language of common use in the scientific community. *Computational veracity* addresses both *computational correctness*, i.e., whether the right analyses were chosen and were configured and implemented in the right way to reach the right computational outcome, and *computational accuracy*, i.e., precision or the degree to which the correct result is represented in the computational outcome. Discrepancies between original and reproduced results highlight deviations from *computational correctness and accuracy*, either in the original or reproduced results (or both). I summarize different sources of computational errors in Table 2 (Baggerly, 2016; Donoho, 2010; Schwalbe et al., 2016; Suchard, 2016; Tibshirani et al., 2002).

In a third step, *computational revisability* translates correction and modification suggestions originating from *computational review* into concrete changes in the *analysis script* (Section 5). Openness of an existing *analysis script* to incorporate changes is particularly important for evaluating the effect of alternative computational settings and/or parameter choices. *Computational flexibility* suggests the capacity of an *analysis script* to be easily modified, such as reconfiguring analyses according to other researchers' preferences or alternative foci (Section 5). *Computational robustness* addresses the question of how analyses respond to (minor or moderate) changes in analytic procedures and assumptions while the main body of the analysis script remains unaltered (Schwalbe et al., 2016, Section 5.1 and 5.2). It implies that *computational modifications* should be tolerated without significant change in the quantitative findings or qualitative conclusions.

In a fourth step, *computational extensibility* enhances conclusions of reported results by calculating alternative variables and/or expanding existing analyses (Section 6). New code is added to existing analyses and the underlying data set is interacted with (more substantially than the mere modification of settings or parameters in existing *analysis scripts*, as is the case for *computational revisability*). This may generate new findings that go beyond what was known based on the original analyses.

In a final step, *computational explorability* reveals additional information contained in the data set by creating analyses that make use of features other than those used in the original analyses (Section 7). This may involve applying a different type of analysis that brings together two or more variables not previously considered jointly. *Computational explorability* predominantly makes use of the data, more so than *computational revisability and extensibility*, which primarily interact with the existing *analysis script*.

1.1. The present study

The present study aimed to computationally reproduce the results of analyses originally reported in Kreibig et al. (2012). I first identified the data set, based on which original analyses were computed. I decided to limit the present attempt of computational reproducibility to statistical analyses and thus base analyses on preprocessed data. I next distilled the core set of relevant statistical analyses from the original analysis scripts and adapted the code to run in the new computational environment. I then worked through an iterative process of re-generating the results of the original study, simplifying the code, and identifying and correcting coding errors. I compiled the result of this process into a compendium—comprising the original data, statistical analysis scripts, and computational environment—and shared it via an online collaboration tool that enables researchers to make their work public (Open Science Framework, OSF; https://osf.io/mhnr6/; developed by the Center for Open Science, Charlotteville, VA).

I first introduce the data example (Section 2). To demonstrate computational documentation and reproduction, I report the results originating from the reproduced computations using the same data set, but based on a new implementation of the analysis script and run in a new computational environment (Section 3). Given the expected outcome of deviations from the results as originally reported in Kreibig et al. (2012), under computational reviewability and veracity, I evaluate and quantify the degree of computational reproducibility between original and reproduced results and present an analysis of computational error types (Section 4). To illustrate some of the added benefits of a computationally reproducible study, I explore boundary conditions of original analyses in the context of computational revisability (Section 5). I show how to build on computational reproducibility in the context of computational extensibility (Section 6). Finally, I demonstrate how we can go beyond documented analyses in the context of computational explorability (Section 7). I illustrate these concepts using the present data example (Kreibig et al., 2012).

2. The data example

2.1. Data

Our data example has originally been described in Kreibig et al. (2012). This data set contains data from 119 participants, comprising demographic information (age, sex), physiological control variables (self-rated health status and consumption of caffeine, nicotine, and alcohol), and emotional response (self-reported feelings and autonomic responses). Emotional responses were collected while participants rested (habituation/wait period) and received feedback on their performance of a cognitive task, which they had just completed. Experimental tasks ensued in the order of habituation, task execution, wait period, and feedback presentation. Performance feedback was manipulated to be either of low or high relevance and either conducive or obstructive to the participant's goal of successful task performance.

Feeling self-report was collected immediately after the habituation and feedback periods with 6 positive (contentment, happiness, interest, joy, pride, surprise) and 9 negative (fear, anger, anxiety, shame, disappointment, embarrassment, repulsion, sadness, and scorn) feeling items. Autonomic activity was recorded during the wait and feedback periods. Data were scored to derive heart rate (HR), respiratory sinus

Table 2
Glossary of terms for computational errors.

Term	Definition
Analysis errors	Computation of the wrong kind of analysis, although the computational output of the computed analysis itself was correct.
Coding errors	Incorrect implementation of analysis, such as referencing incorrect variables or data ranges.
Conceptual errors	Inadequate conceptualization or inconsistent transfer of analysis framework to implemented analyses.
Downstream errors	Errors that result from upstream erroneous computations.
System specific/unclear errors	Deviations in results that may be due to different hardware, operating system, software, or package versions.
Transfer errors	Errors resulting from human inadequacies, such as result entries in the wrong cell of a table. This error type may be introduced as late as in the publishing step where, for example, data tables are newly typeset.

arrhythmia (RSA), pre-ejection period (PEP), mean arterial pressure (MAP), total peripheral resistance (TPR), skin conductance level (SCL), and skin conductance response amplitude (SRA). Kreibig et al. (2012) also calculated cardiac autonomic balance (CAB, defined as the difference of normalized RSA and inverse PEP) and cardiac autonomic regulation (CAR, defined as the sum of normalized RSA and inverse PEP; Berntson et al., 2008).

For analyses, physiological control variables were transformed into ordinal scales. Reactivity scores for feeling self-report were calculated by subtracting ratings collected after the habituation period from the ones collected after feedback presentation. For autonomic data, period averages were calculated for the 90-s wait period and the 120-s feedback period. For each participant and each measure, reactivity values during the feedback period were calculated by subtracting the average over the wait period.

2.2. Computational environment for statistical analyses

Statistical analyses for the present study were written in R-3.3.2 (R Development Core Team, 2007). `R` packages used included functions for bootstrapping (`boot`, version 1.3-18), multivariate analysis of variance (MANOVA, `car`, version 2.1-3), partial eta squared for linear models (`heplots`, version 1.3-1), parsing LATEX math formulas to `R` (latex2exp, version 0.4.0), linear discriminant analysis (`MASS`, version 7.3-45), and item analysis (`rela`, version 4.1). Computations were run on a 3.5 GHz Intel Core i7 iMac with OS X Version 10.10.5.

To enable reproducibility, the computational environment was emulated using the Docker software containerization platform (Docker Inc., San Francisco, CA; Boettiger, 2015). `R` scripts were ported into Jupyter Notebooks (https://jupyter.org/), a web application that allows one to create and share documents that contain interactive analysis code. The complete compendium (data, analysis scripts, and computational environment) is available via the OSF, https://osf.io/mhnr6/.

3. Computational documentation and reproduction

3.1. Statistical analyses for originally reported results

Computational results to be reproduced from the original study (Kreibig et al., 2012) were as follows: (1) demographic information of the sample and adjusted α-levels, which consider item intercorrelation (i.e., among subjective feeling variables and autonomic variables, respectively; Section 3.2); (2) preliminary analyses that tested *a priori* group differences using two-way analyses of variance (ANOVAs) or analyses of covariance (ANCOVAs) on (a) subjective feeling (Section 3.3.1) and (b) autonomic activation during baseline and autonomic reactivity during feedback presentation, including physiological control variables (Section 3.3.2); (3) main analyses involving two-way ANOVAs or ANCOVAs and—if significant—post-hoc Tukey Honestly Significant Differences (HSD) tests and decomposition of group means for inspection of residuals following a significant interaction effect (Rosnow and Rosenthal, 1989, 1995) on (a) subjective feelings (Section 3.4.1) (b) reactivity to feedback presentation (Section 3.4.2); and (4) supplementary analyses of the interaction hypothesis on composite scores of positive and negative feelings (Section 3.5). In what follows, I use underlining to mark deviating results between the original (Kreibig et al., 2012) and present study in text and associated tables (see Online Supplementary Material [SOM]).

3.2. Reproduction of demographic results and adjustment of significance level

As previously reported (Kreibig et al., 2012), 119 participants were considered in analyses, including 99 women (aged $M = 22.1$ years, $SD = 5.7$) and 20 men (aged $M = 25.0$ years, $SD = 5.6$). Unlike prior analyses, which only included α-adjustment for analyses of feedback reactivity, but not of baseline activation, I here included such adjustments (as summarized in SOM, Table A1). For tests of feedback reactivity, new calculations resulted in a less conservative α-level for autonomic indices (CAB, CAR), a more conservative α-level for cardiovascular indices (HR, MAP, TPR), and were otherwise identical.

3.3. Reproduction of preliminary analyses

3.3.1. Subjective feelings

Results from analyses of *a priori* group differences at baseline were numerically identical and supported the prior conclusion that "feeling self-report assessed after the habituation period did not differ between experimental groups" (p. 369, Kreibig et al., 2012; see SOM, Table A2).

3.3.2. Autonomic responses

Results from analyses of autonomic baseline activation replicated the prior finding that "experimental groups did not differ with respect to baseline autonomic activation," except on RSA and marginally on CAR, and that "autonomic activation during the wait period was … stable over the duration of the measurement" (p. 369, Kreibig et al., 2012; see SOM, Table A3). Still, I noted some minor numerical deviations in the second decimal place for *F*-values of the goal relevance effect reported from the ANOVA.

Two-way ANCOVAs with physiological control variables on autonomic activation during baseline and autonomic reactivity during feedback presentation indicated deviating results (summarized in SOM, Tables A4 and A5): Unlike the original finding of no influence of physiological control variables on autonomic baseline activation (Kreibig et al., 2012), present analyses indicated that nicotine consumption had a significant effect on MAP and caffeine consumption had a significant effect on TPR. Similarly, unlike the original finding of nicotine consumption influencing PEP and SRA reactivity during feedback presentation, present analyses indicated that general health had a significant effect on HR. Accordingly, nicotine consumption, caffeine consumption, and general health were included as covariates in main analyses of MAP, TPR, and HR, respectively.

3.4. Reproduction of main analyses

3.4.1. Subjective feelings

Consistent with original analyses, results from two-way ANOVAs on subjective feelings indicated significant main effects of goal conduciveness on contentment, happiness, joy, pride, and disappointment of medium to large effect size (SOM, Tables A6 and A7). Results of post-hoc Tukey HSD tests supported the prior conclusion that greater contentment was elicited by conducive than obstructive feedback; greater happiness and joy by low-relevance conducive than high- and low-relevance obstructive feedback; and lesser pride and greater disappointment by high-relevance obstructive than high- and low-relevance conducive feedback. There was one erroneous column entry in the original results table.

3.4.2. Autonomic responses

Analysis of goal relevance and goal conduciveness effects on autonomic reactivity during feedback presentation (summarized in SOM, Table A8 and Figure A1), produced conceptually identical results, albeit with slight numerical differences (given the above-noted modifications in covariate-adjusted analyses and deviations in the calculation of three effect size measures). As previously reported, I found "a significant relevance by conduciveness interaction … for the adjusted CAR measure" (p. 369, Kreibig et al., 2012). Post-hoc Tukey HSD tests showed that CAR was significantly greater for high-relevance conducive than high-relevance obstructive feedback. Decomposition of group means for inspection of residuals indicated an interaction ($g_{R\times C} = \pm 0.116$), which substantially deviated from the grand mean ($G = -0.129$) and made the largest contribution to group means compared to relevance

($g_R = \pm 0.007$) and conduciveness effects ($g_C = \pm 0.024$). As before, main effects for goal relevance were found on HR and SCL, with lower HR and greater SCL for the high vs. low relevance conditions, and main effects for goal conduciveness on MAP and SCL, with greater MAP and SCL for high-relevance conduciveness than the other conditions. Again, inclusion of physiological control variables in analysis did not change the pattern of results (see SOM, Table A8, unadjusted indices). These findings support the original conclusion that "effects of CAR, HR, MAP, and SCL differentiated high- and low-relevance conducive and obstructive experiences through interaction and main effects" (p. 370, Kreibig et al., 2012).

3.5. Reproduction of supplementary analyses

As in original analyses (Kreibig et al., 2012), interaction tests of sum scores of positive and negative feeling items remained non-significant, albeit the more liberally adjusted α-level of $p = .025$: positive, $F(1, 115) = 0.04$, $p = .84$, $\eta_G^2 < .001$; negative, $F(1, 115) = 4.04$, $p = .047$, $\eta_G^2 = .034$. I noted a report of one deviating F-value.

In conclusion, in the present section I generated the computational documentation—data, analysis scripts, and computational environment—necessary for computational reproducibility and made the full computational compendium available via the OSF. I also documented the results obtained from the computational reproduction of the statistical results originally reported in Kreibig et al. (2012).

4. Computational reviewability and veracity

Provision of complete computational documentation establishes computational reviewability.

4.1. Evaluation of computational veracity

The present test of computational reproducibility involved the reproduction of 15 analyses and 999 computational results (Table 3). To evaluate computational veracity, I compared original and reproduced results against each other. Conceptually, core findings of the original study (Kreibig et al., 2012) remained unchanged. Numerically, computational reproduction resulted in a reproducibility rate of 91%. I identified 91 discrepancies between results reported in the original and the present study.

Various types of errors affected 10 of the 15 analyses reported in the original study (Kreibig et al., 2012). As detailed in Table 3, transfer errors were noticed in two cases: In one instance, results were incorrectly presented in a table, one column shifted over (original Table 3, reproduced SOM, Table A7) and in another instance, an incorrect value was reported for a numeric result presented in-text (original and reproduced in-text). System-specific/unclear errors were identified in four cases, affecting F-values of the ANOVA goal-relevance main-effect (original Table 1 and reproduced SOM, Table A3), η_G^2 and p-values (original Table 4 and SOM, reproduced Table A8), decomposition of the group mean conduciveness effect (original and reproduced in-text), and a p-value (original Footnote 5 and reproduced SOM, Table A8), presumably resulting from minor deviations given new hardware and operating systems and updated analysis software and library versions. Coding errors were observed in three cases, involving in one case incorrect range referencing of a variable set (original and reproduced in-text) and in two cases assuming a different data structure than present (original qualitative statement and reproduced in SOM, Tables A4 and A5). Downstream errors occurred in three cases, affecting calculation of adjusted α-levels given incorrect item intercorrelations (original in-text and reproduced SOM, Table A1) and calculation of AN(C)OVAs on autonomic reactivity variables (original Table 4 and reproduced SOM, Table A8) and related figures (original Figs. 1–3 and reproduced SOM, Figure A1) given incorrect identification of influence of physiological control variables. An analysis error was identified in one case, where a function for a diverging effect-size measure was called (original Table 4 and reproduced SOM, Table A8). A conceptual error was noted in that α-adjustment had only been applied to analyses of feedback reactivity but not to baseline activation (original in-text and SOM, reproduced Table A1).

Fig. 1 summarizes the relative contribution of each of the errors to the overall number of observed computational errors. While transfer errors were limited to a small percentage of cases (3 cases or 3%), and system-specific or unclear errors were also relatively infrequent (11 cases or 12%), it is notable that the relatively small number of coding errors (11 cases or 12%) gave rise to the large majority of downstream errors (53 cases or 58%). Analysis errors (3 cases or 3%) and conceptual errors (10 cases or 11%) again were rather infrequent and of limited impact.

Based on the above computational review, I conclude that while conceptually, core findings of the original study remained unchanged, numerically I found a reproducibility rate of 91%. Based on a detailed error analysis, I observed different types of deviations in computational correctness and accuracy. I observed that a small number of coding errors can lead to a large number of downstream errors. Companion publication of data, analysis scripts, and computational environment as well as code review and double checking of reported results can help eliminate errors, limiting them to those of system-specific/unclear sources.

5. Computational revisability: Flexibility and robustness

Journal article components, including figures and tables, are now commonly provided for download to support their reuse in research and teaching (Sandusky et al., 2007). However, depending on the focus of presentation, article components as displayed in the original publication may not be adequate for the new purpose. Providing the source code and data to reproduce the figures and tables allows for their flexible modification. As an example of computational flexibility, in SOM Figure A2, I present data on autonomic reactivity during feedback presentation (cf. SOM, Figure A1) with variables on the abscissa and trace exchanged.

To evaluate computational robustness, I here explore the impact of using different α-adjustment methods (e.g., Bonferroni method; Section 5.1) and different participant subsamples (e.g., exclusion of skin conductance non-responders; Section 5.2) on results and conclusions of the original study (Kreibig et al., 2012).

5.1. Evaluation of computational revisability: Alpha adjustment

Alpha adjustment continues to pose debate in many fields of research (Armstrong, 2014; Hewes, 2003; Hullett, 2007; Matsunaga, 2007; O'Keefe, 2003a; O'Keefe, 2003b; O'Keefe, 2007; Tutzauer, 2003, Weber, 2007), and different forms of alpha adjustment have been proposed and used (Blakesley et al., 2009; Hsu, 1996; Shaffer, 1995). The original study (Kreibig et al., 2012) used a single-step Šidák derivative method, referred to as Dubey/Armitage-Parmar (D/AP) α-adjustment, which incorporates a measure of correlation among dependent variables (Sankoh et al., 1997).

Using a more conservative approach (Blakesley et al., 2009), with the Bonferroni method being the most conservative (Bonferroni, 1935, 1936), would affect only one result, which is significant under the D/AP method but not under the Bonferroni method. This concerns the interaction effect of goal relevance and goal conduciveness on CAR reactivity in response to feedback presentation. Though just affecting one result, this represents the core result of the original study (Kreibig et al., 2012) and would eliminate the already weak support for the interaction hypothesis of goal relevance and goal conduciveness.

Using a more liberal approach, such as the variously advocated abandonment of α-adjustment procedures (e.g., O'Keefe, 2003a), would

Table 3
Detailed analysis of computational veracity based on original results (Kreibig et al., 2012) and reproduced results from the present study.

	Reported analysis	Original results	Reproduced results	Variables	No. var.	Derived measures	No. derived measures	No. data points	No. deviations	Error type	Reproduced data points
Demographic information and adjustment of significance level											
1	Demographics	In-text	In-text	no. subjects	1	*N*	1	1	0	–	100%
2				gender, age	2	*N*,*M*,*SD*	3	6	0	–	100%
3	Adjustment of significance level	In-text	SOM, Table A1	POS, NEG, DA, CV, EDA at bsl. and fdbk.	10	*α*,*r*	2	20	14	CE, CNE, DSE	30%
Preliminary analyses											
4	Subjective feelings (bsl. diff.)	Qualitative	SOM, Table A2	SF	15	*p*	1	15	0	–	100%
5	Autonomic activation (bsl. diff.)	Table 1	SOM, Table A3	ANS	9	*M*,*SE*,*F*,*p* for R, C, IA; Cronbach's *α*	15	135	7	S/U	95%
6	Autonomic activation (phys. ctr. var.)	Qualitative	SOM, Table A4	ANS	9	*df*s, *F*,*p* for alc., caf., nic., g.h.	12	108	4	CE	96%
7	Autonomic reactivity (phys. ctrl. var.)	Qualitative	SOM, Table A5	ANS	9	*df*s, *F*,*p* for alc., caf., nic., g.h.	12	108	6	CE	94%
Main analyses											
8	Subjective feelings (exp. effects)	Table 2	SOM, Table A6	SF	15	*M*,*SE* for R × C	8	120	0	–	100%
9		Table 3	SOM, Table A7	SF	15	F, p, η_G^2 for R, C, IA; p for each contrast	15	225	2	TE	99%
10	Autonomic reactivity (exp. effects)	Table 4	SOM, Table A8	ANS	9	F, p, η_G^2 for R, C, IA; p for each contrast	15	135	35	AE, CE, DSE, S/U	74%
11		Figs. 1–3	SOM, Fig. A1	ANS	9	*M*,*SE* for R × C	8	72	20	DSE	72%
12		In-text	In-text	CAR	1	IA, grand mean, main effects	4	4	1	S/U	75%
13		Footnote 4	SOM, Table A8	CAR, RSA	2	*df*s, F, p, η_G^2 for IA	5	10	0	–	100%
14		Footnote 5	SOM, Table A8	PEP, SRA	2	*df*s, F, p, η_G^2 for R, C, IA	15	30	1	S/U	99%
Supplementary analyses											
15	Subjective feelings composites	In-text	In-text	POS, NEG	2	*df*s, F, p, η_G^2 for IA	5	10	1	TE	90%
Total								**999**	**91**		**91%**

Note. "Original results" refers to the data source from Kreibig et al. (2012); "Reproduced results" refers to the data source in the present study. Abbreviations: No. var. – number of variables; bsl. diff. – baseline differences; bsl. – baseline; fdbk. – feedback; phys. ctr. var. – physiological control variables; exp. effects – experimental effects; POS – positive; NEG – negative; DA – derived autonomic indices; CV – cardiovascular indices; EDA – electrodermal indices; SF – subjective feelings; ANS – autonomic variables; R – relevance; C – conduciveness; IA – interaction; alc. – alcohol consumption; caf. – caffeine consumption; nic. – nicotine consumption; g.h. – general health; AE – analysis error; CE – coding error; CNE – conceptual error; DSE – down-stream error; S/U – system-specific/unclear; TE – transfer error.

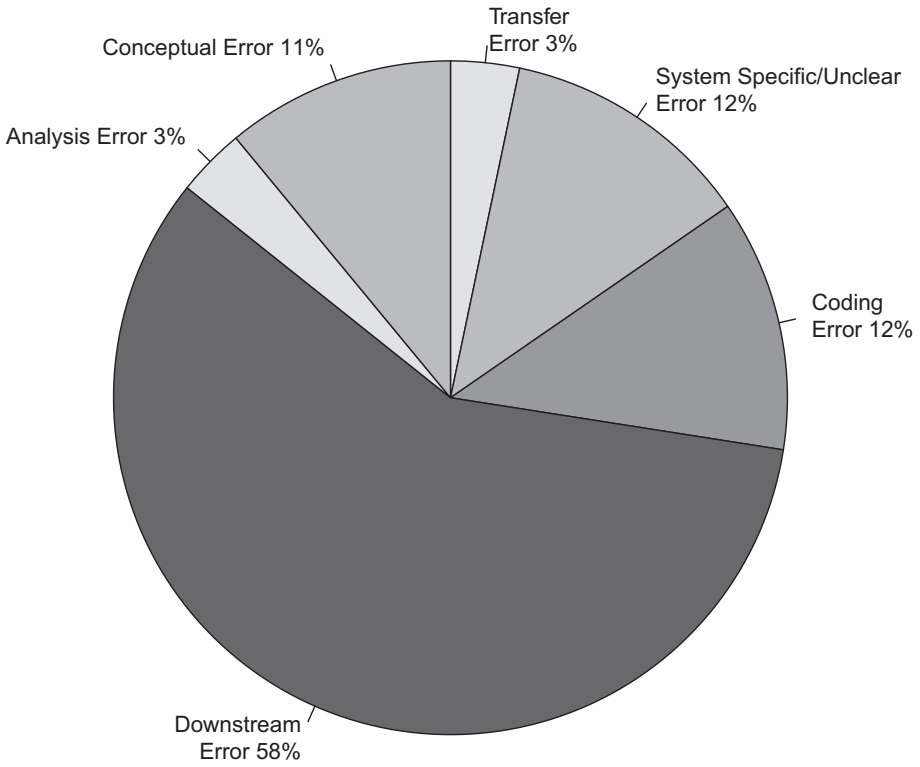

Fig. 1. Break-down of computational errors identified in the computational reproduction of Kreibig et al. (2012) in the present study.

result in no differences for subjective feelings and autonomic activation at baseline or covariance analyses. For main analyses of self-reported feelings, this approach would identify a significant relevance main effect on shame and repulsion, with greater shame and repulsion for high-relevance feedback. It would identify a broader effect of goal conduciveness not only on contentment, happiness, joy, and pride, but also on interest, with greater interest for conducive than obstructive feedback. It would also identify a broader effect of goal obstruction not only on disappointment but also on fear, anger, embarrassment, repulsion, and sadness, with greater negative feelings for obstructive than conducive feedback. In addition, this approach would result in a predicted interaction effect of goal relevance and goal conduciveness on repulsion, with greater repulsion for high-relevance obstructive feedback than for the other feedback conditions, as shown in SOM, Figure A3.

For autonomic reactivity, in addition to the relevance main effects on HR and SCL, this analysis approach would identify a main effect of goal relevance on CAB, with greater CAB reactivity for high than low relevance. In addition to conduciveness main effects on MAP and SCL, it would identify main effects of goal conduciveness on TPR and SRA, with greater TPR and greater SRA for conducive than obstructive feedback. In addition to an interaction effect of goal relevance and goal conduciveness on CAR, it would identify predicted interaction effects on CAB, RSA, and MAP (see SOM, Figure A2), with greater CAB reactivity for high-relevance conducive than low-relevance conducive feedback; greater RSA reactivity for high-relevance conducive than low-relevance conducive and high-relevance obstructive feedback; and greater MAP for high-relevance conducive than high- or low-relevance obstructive feedback.

Decomposition of group means indicated that the largest contribution to group means was made by the interaction effect for CAB and RSA, but by the conduciveness effect for MAP. This suggests primary contribution of the predicted interaction effect to the effect observed on CAB and RSA, but not on MAP (SOM, Table A9). Thus, while results obtained under the Bonferroni method would *eliminate* the already weak support for the interaction hypothesis of goal relevance and goal conduciveness, results obtained under abandonment of α-adjustment procedures would *strengthen* the support for the interaction hypothesis with new evidence on the level of self-reported feelings (repulsion) and additional evidence on the level of autonomic reactivity (CAB and RSA).

5.2. Evaluation of computational revisability: Skin conductance non-responders

Skin conductance non-responders are individuals who show no or little electrodermal activity, operationalized as those who show a response amplitude of 0 μSiemens in all (Venables and Christie, 1980) or $\geq$ 90% of recording periods (Lake et al., 2016). They make up approximately 5–10% of participants (Straube, 1979; Venables, 1978). Different approaches to handling skin conductance response data with zero amplitudes have been discussed (Boucsein, 1992, 2012; Venables and Christie, 1980): Analyses may include SCR data of all recording epochs, including those epochs of zero amplitude (Armel and Ramachandran, 2003; Lake et al., 2016) or only SCR data without recording epochs of zero amplitude (i.e., using SCR magnitude; Spottiswoode and May, 2003). Alternatively, analyses may include SCR data of all subjects, including those who exhibited zero amplitude to the majority of stimuli (i.e., nonresponders; Siller et al., 1995) or only SCR data from subjects who did not meet criteria for skin conductance non-responders (Armel and Ramachandran, 2003; Lake et al., 2016; van der Zwaag et al., 2011). The original study (Kreibig et al., 2012) included all subjects, including potential skin conductance non-responders, because these data may represent genuine zero responses during a neutral baseline condition and under low-relevance performance feedback (Boucsein, 2012).

To evaluate the impact of skin conductance non-responders on the results, I re-analyzed SRA data using two different definitions of skin conductance non-responders. I defined skin conductance non-responders as participants who showed no skin conductance response either in 100% (more stringent definition) or in $\geq$ 90% (less stringent definition) of trials, which included the five 20-s baseline epochs and six 20-s feedback epochs. According to the more stringent definition, I identified nine participants (eight women) as non-responders, as summarized in SOM, Table A10 (a). For the less stringent definition, I

identified 15 participants (12 women) as non-responders, see SOM, Table A10 (b). Rerunning the analyses with these participants excluded showed no difference in results for any of the analyses or conclusions (SOM, Table A10 (c) and Figure A4).

Excluding zero amplitude SCR data did not make the evidence in the data for the interaction hypothesis of goal relevance and goal conduciveness stronger. It may rather have weakened it given the reduced number of subjects considered in analyses. This is especially true if zero amplitude SCR data represented valid responses in the present data set.

In summary, I evaluated the modifiability of statistical analyses by exploring the impact of different α-adjustment methods and participant subsamples on results and conclusions of the original study (Kreibig et al., 2012). Minor to moderate computational modifications left quantitative findings and qualitative conclusions relatively unchanged.

6. Computational extensibility

I illustrate computational extensibility, first, by reporting results from feeding newly derived composite scores of subjective feelings to the set of main analyses (Section 6.1). This approach allows an evaluation of whether composite scores may be more sensitive to indicate the predicted interaction effect of goal relevance and goal conduciveness. Second, I report results from wrapping a bootstrap analysis around the set of main analyses to run an internal replicability analysis (Section 6.2). This allows calculation of confidence intervals for effect size measures, based on which we can estimate replicability of reported results in future studies.

6.1. Evaluation of computational extensibility: Composite scores of self-reported feelings

In recent years, there has been increased interest in evaluating the sensitivity of composite measures of self-reported feelings, including positive and negative affect (Ito et al., 1998; Kron et al., 2015; Larsen et al., 2003), subjective valence and arousal (Betella and Verschure, 2016; Bradley and Lang, 1994), and mixed emotional feelings (Larsen et al., 2016; Schimmack, 2001). Composite measures may reveal aspects not shown by individual measures, particularly if they are based on nonlinear integration of individual items. Additionally, the smaller number of tests necessitated by the smaller number of composite scores allows for a more liberal significance level for each test than when testing each individual measure.

I formed composite scores for positive feelings (sum of contentment, happiness, interest, joy, and pride), negative feelings (sum of fear, anger, anxiety, shame, disappointment, embarrassment, repulsion, sadness, and scorn)[1], subjective bipolar valence (difference between positive and negative feelings), subjective arousal (sum of positive and negative feelings; Kron et al., 2015), and mixed emotional feelings (minimum of positive and negative feelings; Schimmack, 2001).

As summarized in SOM, Table A11, composite indices revealed strong goal conduciveness main effects: Positive feelings were greater for conducive than obstructive feedback. Negative feelings were greater for high- and low-relevance obstructive than high-relevance conducive feedback. More positive and less negative feelings were present for conducive than obstructive feedback. Arousal was comparable between feedback conditions other than being greater for low-relevance conducive than low-relevance obstructive feedback. Finally, mixed emotional feelings were greater for high-relevance obstructive than high- or low-relevance conducive feedback.

None of the composite scores showed improved sensitivity to the hypothesized interaction effect of goal relevance and goal conduciveness. Results of subjective arousal and mixed emotional feelings extended findings beyond what was known about the effects of experimental feedback presentation on subjective emotional feelings based on the original study (Kreibig et al., 2012): Results on arousal suggested that feedback conditions were comparable with respect to subjective arousal level. Results on mixed emotional feelings suggested that high-relevance obstructive feedback elicited greater co-activation of positive and negative feelings than any of the other conditions.

[1] This approach excluded surprise from positive and negative composite scores, because surprise may be experienced in response to both goal conducive and obstructive conditions and thus cannot be unambiguously assigned a positive or negative feeling quality.

6.2. Evaluation of computational extensibility: Internal replicability analysis

Because conventional statistical significance tests do not evaluate result replicability (Cumming, 2008; Cumming and Maillardet, 2006), it is useful to employ replicability analysis. Internal replicability addresses the probability that future replications of the current study would yield similar results (cf. Kreibig et al., 2015). The bootstrap approach (Efron, 1979; Wasserman and Bockenholt, 1989) represents a particularly powerful internal replication method: It generates a very large number of additional samples from the current sample and recalculates relevant parameters for each sample, based on which confidence intervals (CIs) are constructed to express the uncertainty associated with point estimates.

Replicability of results from the original study (Kreibig et al., 2012) was tested based on CIs of effect sizes derived from non-parametric bootstrap analysis. Given the between-subject design of the study, subjects within groups were used as units for resampling with replacement for constructing bootstrap samples. One thousand samples of the original four-group sample were created. For each bootstrap sample, the set of main analyses was computed and resulting effect sizes were saved. Bootstrap distributions were generated by compiling respective effect sizes across bootstrap samples, from which 90% CIs for η_G^2 were constructed. The conventional definition of small effect sizes ($\eta_G^2 = 0.01$ [small], 0.09 [medium], 0.25 [large]; Bakeman, 2005; Cohen, 1988) was used as lower bound for evaluating replicability of similar-sized effects in future studies.

SOM, Figure A5 shows 90% CIs of effect sizes (η_G^2) from ANOVAs on subjective feelings among high- and low-relevance conducive and obstructive feedback. Results indicated for the conduciveness effect large effects on contentment (0.253, 0.414) and small-to-medium sized effects on happiness (0.073, 0.246), joy (0.022, 0.186), pride (0.042, 0.203), anger (0.011, 0.122), disappointment (0.048, 0.205), and embarrassment (0.017, 0.138). Medium-to-large sized effects were found for composite scores of positive feelings (0.159, 0.338), negative feelings (0.092, 0.241), and subjective valence (0.231, 0.402), while small-to-medium sized effects were found for arousal (0.015, 0.157) and mixed emotional feelings (0.058, 0.221).

For autonomic reactivity (see SOM, Figure A6), results indicated small-to-medium sized effects for the relevance effect on HR (0.034, 0.207) and SCL (0.013, 0.152), for the conduciveness effect on MAP (0.016, 0.175) and SCL (0.037, 0.200), and for the interaction effect on CAR (0.010, 0.132).

Taken together, results of internal replicability analysis indicated strong likelihood of replication in future studies of goal conduciveness main effects on a considerable number of subjective feeling variables. In particular, results underscored an important role of contentment in distinguishing conducive and obstructive conditions and the usefulness and increased reliability of forming composite indices (positive and negative feelings, subjective valence). Results furthermore indicated likely replication of goal relevance and goal conduciveness main and interaction effects on the set of autonomic variables identified in the original study (CAR, HR, MAP, and SCL; Kreibig et al., 2012) and supported sensibility of the α-adjustment originally applied. In summary, in the above section, I explored extensibility of statistical

Table 4
Loadings and percentage of explained variance of Principal Component Analysis on autonomic reactivity.

Principal Components	PC1	PC2	PC3	PC4	PC5	PC6	PC7
	Peripheral vaso-constriction	Cardiac chronotropic response	Sympathetic cholin.-/ adren. response	Cardiac autonomic regulation	Electro- dermal activation	Cardiac chronotropic response	Peripheral vaso-constriction
Variables							
RSA	0.19	−0.36	0.02	**0.76**	0.24	0.22	0.39
PEP	0.15	0.40	**-0.51**	**0.51**	−0.25	−0.35	−0.33
HR	0.10	**0.67**	−0.27	−0.11	0.31	**0.52**	0.31
MAP	**0.55**	−0.14	−0.30	−0.33	−0.05	−0.43	**0.54**
TPR	**0.54**	−0.33	−0.25	−0.16	0.23	0.36	**-0.57**
SCL	0.46	0.19	0.44	0.07	**-0.68**	0.30	0.04
SRA	0.35	0.32	**0.56**	0.08	**0.52**	−0.39	−0.17
Explained variance							
Percentage	26.8	21.6	16.5	15.7	8.4	6.8	4.2
Cumulative	26.8	48.4	64.8	80.6	89.0	95.8	100

Note. For abbreviations of variable names, see Figure B1. PC – principal component. Loadings typeset in boldface mark those which met the criterion of $r \geq 0.5$ and were used for interpreting PCs. Percentage – percentage of variance explained; cumulative – cumulative percentage of variance explained.

Table 5
Results of two-way Analysis of Variance (ANOVA) with $df = 1115$, generalized Eta-squared measure of effect size (η_G^2), and p-Values of Tukey HSD post-hoc tests for Principal Components (PCs) of autonomic reactivity during feedback presentation ordered by relative importance.

		ANOVA						Tukey HSD					
Principal components	Relative importance	Relevance		Conduciveness		Rel. × Cond.							
		F	η_G^2	F	η_G^2	F	η_G^2	C:c	O:o	C:O	c:o	C:o	O:c
PC1	0.252	3.51.	0.030	**22.73*** **	**0.165**	**7.00** **	**0.057**	.010	–	–	< .001	–	< .001
PC2	0.087	**7.19** **	**0.059**	1.34	0.011	1.98	0.017	.023	–	–	–	.037	–
PC5	0.069	**6.81***	**0.056**	0.03	< 0.001	1.44	0.012	.039	–	–	–	–	–
PC3	0.053	4.49(*)	0.038	0.38	0.003	1.40	0.012	–	–	–	–	–	–
PC4	0.045	0.17	0.001	0.31	0.003	4.91(*)	0.041	–	–	–	–	–	–
PC7	0.031	0.14	0.001	1.34	0.012	2.17	0.019	–	–	–	–	–	–
PC6	0.007	0.37	0.003	0.36	0.003	0.07	0.001	–	–	–	–	–	–

Note. Relative importance – sum of η_G^2 of goal relevance, goal conduciveness, and their interaction. Adjusted significance level for multiple testing is $\alpha = .0167$, indicated by the following symbols: (.) $p < .10$; (*) $p < .05$; * $p < .0167$; ** $p < .01$; *** $p < .001$. Values typeset in boldface remain significant after correcting α for multiple tests using the Bonferroni method for uncorrelated PCs. C – high-relevance conducive; c – low-relevance conducive; O – high-relevance obstructive; o – low-relevance obstructive.

analyses, first, by subjecting a new set of variables to the set of main analyses and, second, by wrapping a bootstrap analysis around the set of main analyses to run an internal replicability analysis.

7. Computational explorability

I illustrate computational explorability first by using a multivariate analysis approach for testing the interaction hypothesis of goal relevance and goal conduciveness on autonomic reactivity during feedback presentation (Section 7.1). This type of analysis draws on the multivariate structure of multiple autonomic variables that has not been leveraged in the exclusively univariate analysis approach of the original study (Kreibig et al., 2012). Second, I illustrate computational extensibility by applying classification analysis to autonomic reactivity during feedback presentation (Section 7.2). This analysis tests whether for previously unseen cases group membership can be predicted reliably from the set of predictors. That is, can we do better than chance in predicting if the subject experienced conducive or obstructive feedback of high or low relevance?

7.1. Evaluation of computational explorability: Multivariate test

Multivariate analysis of variance (MANOVA) tests whether experimental groups show statistically significant mean differences on a combination of dependent variables (Tabachnick and Fidell, 2013). The important variables contributing to the multivariate group separation can be identified by using ANOVA on uncorrelated dependent variables (as formed by principal component analysis, PCA), with the dependent variables that have significant univariate F-values being the important ones, ranked in importance by effect size. To interpret the pattern of differences among the dependent variables identified by MANOVA, descriptive discriminant analysis (DDA) is used. DDA finds a linear combination of the dependent variables that gives maximum separation between group centroids while at the same time minimizing the variation within each group (Venables and Ripley, 2002).

I submitted autonomic variables to PCA[2] to create the seven uncorrelated autonomic principal components (PCs) shown in Table 4. The MANOVA on this set of PCs identified a goal relevance main effect, $F(7,115) = 3.58$, $p < .01$, $\eta_G^2 = 0.179$, a goal conduciveness main effect, $F(7,115) = 3.72$, $p < .01$, $\eta_G^2 = 0.185$, and—as hypothesized—an interaction effect of goal relevance and goal conduciveness, $F(7,155) = 2.94$, $p < .01$, $\eta_G^2 = 0.152$. To identify which PCs contributed to the effects identified by MANOVA, I next ran separate ANOVAs on each PC. To avoid inflation of Type I error, I used a Bonferroni-type adjustment to keep the overall α below .05 (setting α to .015 for the first three PCs to give more important variables more liberal alphas and to .001 for the other four PCs for an overall α of .049; p. 272; Tabachnick and Fidell, 2013). In decreasing order of importance, significant effects were shown by PCs 1 (goal conduciveness and interaction effects), 2, and 5 (goal relevance effects; see Table 5).

I next subjected the subset of three PCs, which showed significant effects in the ANOVA, to DDA to interpret the dimensions along which feedback conditions differed. As reported in Table 6, the first two linear

[2] PCA did not include derived autonomic variables because this would have introduced multicollinearity.

Table 6
Linear discriminant analysis of three-Principal Component (PC) top-down solution identified as significant in univariate analysis of variance, ordered by relative importance (see Table 6).

	Linear discriminants		
	LD1	LD2	LD3
Principal Components			
PC1	**0.74**	−0.34	0.12
PC2	−0.25	**-0.64**	−0.48
PC5	**-0.59**	**-0.66**	**1.01**
Variance explained			
Percentage	72.9	26.9	0.2
Cumulative	72.9	99.8	100.0

Note. Percentage – percentage of variance explained; cumulative – cumulative percentage of variance explained. Loadings typeset in boldface mark those which met the criterion of $r \geq .5$ and were used for interpreting LDs.

discriminant functions accounted for 72.9% and 26.9%, respectively (explaining together 99.8% of the between-group variance, suggesting that the contribution to group separation of the third discriminant function was negligible). Fig. 2 (a) shows that the first discriminant function maximally separated high-relevance conducive feedback from the other three feedback conditions. The second discriminant function discriminated low-relevance conducive feedback from high and low-relevance obstructive feedback.

The structure (loading) matrix of correlations between predictors and discriminant functions, as seen in Table 6, suggested that the best predictors for distinguishing between high-relevance conducive feedback and the other feedback conditions (first function) were PCs 1 and 5:[3] High-relevance conducive feedback elicited greater vasoconstriction (MAP and TPR) and electrodermal activity (SCL) than the other feedback conditions (cf. SOM, Figure A2). The best predictors for separating low-relevance conducive from high and low-relevance obstructive feedback were PCs 2 and 5: Low relevance conducive feedback elicited less cardiac slowing (HR) and electrodermal deactivation (SCL and SRA) than high and low-relevance obstructive feedback (cf. SOM, Figure A2).

In summary, using this top-down multivariate analysis approach for testing the interaction hypothesis of goal relevance and goal conduciveness on autonomic reactivity during feedback presentation suggested considerable separation of group cell means: Feedback conditions formed three clusters, separating high-relevance conducive, low-relevance conducive, and (high and low-relevance) obstructive feedback. Multivariate response patterns of cardiac chronotropic, peripheral vascular, and electrodermal reactivity were important variables for distinguishing among feedback conditions. These results contribute new insights to the interaction hypothesis of goal relevance and goal conduciveness on emotional responding (Kreibig et al., 2012).

7.2. Evaluation of computational explorability: Predictive classification

Predictive discriminant analysis (PDA) is used in situations where prior designation of groups exists. In the present data example, these were groups of high or low-relevance conducive or obstructive feedback. The aim of PDA is to classify observations into these known groups. PDA generates a model or classification rule according to which observations are assigned their likely group membership. To avoid over-fitting the model to the training data, cross-validation can be used for estimating how the results of a statistical analysis will generalize to an independent data set. Leave-one-out cross-validation uses $N-1$ observations as the training set and the remaining observation as the validation set. This is repeated N times to create all possible samples for calculating the average correct classification rate. Again, DDA is used to interpret the dimensions identified by PDA along which groups differ.

I ran predictive PDA with cross-validation using the leave-one-out approach over sets of PCs of decreasing relative importance, i.e., sum of two-way ANOVA effect sizes (see Table 5). Fig. 3 summarizes the percentage correct classification for each of the seven different PDAs. Maximal group separation was achieved with the subset of four most important PCs: For the total usable sample of 119 subjects, 53 subjects (44%) were classified correctly (see Table 8), compared with 29.75 subjects (25%) by chance alone. The 44%-classification rate was achieved by misclassifying a number of cases of low-relevance obstructive feedback as high-relevance obstructive or low-relevance conducive feedback.

To interpret the classification rule according to which feedback conditions were differentiated, I subjected this four-PC bottom-up solution to DDA. While the first two linear discriminant functions again accounted for a substantial percentage of the between-group variance, the third linear discriminant accounted for an extra 7% (see Table 7). Fig. 2 (b) shows that group separation was similar as for the three-PC top-down solution: The first discriminant function maximally separated high-relevance conducive feedback from the other three feedback conditions through PCs 1 (greater vasoconstriction) and 5 (greater electrodermal activity). However, with the present solution, the second discriminant function separated low-relevance conducive from high-relevance obstructive feedback, with low-relevance obstructive feedback falling in between, through PCs 2 (lesser cardiac slowing) and 5 (lesser electrodermal deactivation). The new third discriminant function separated low-relevance conducive and high-relevance obstructive feedback from low-relevance obstructive feedback through PC3. Low-relevance conducive and high-relevance obstructive feedback elicited lesser inotropic increases (PEP) and electrodermal deactivation (SRA) than low-relevance obstructive feedback (cf. SOM, Figure A2).

This bottom-up multivariate analysis approach showed that—on average—groups of different feedback experiences can be classified at a better-than-chance rate, with the exception of low-relevance obstructive feedback, which showed a high confusion rate with low-relevance conducive and high-relevance obstructive feedback. Classification seemed to work best for high-relevance conducive feedback. This corroborates findings of the top-down multivariate analysis approach, which suggested three clusters, separating high-relevance conducive, low-relevance conducive, and (high and low-relevance) obstructive feedback. Notably, even though different PC subsets (three vs. four PCs) were used for MANOVA/DDA and PDA/DDA, obtained results were quite consistent between analyses. These results suggest that greater appraisal of goal relevance amplifies the emotional response to events appraised as conducive but not to those appraised as obstructive to one's goals. Events appraised as conducive may lead to a mobilization of effort relative to the appraised relevance of the event. However, events appraised as obstructive may lead to a withdrawal of effort irrespective of the appraised relevance of the event. This finding has not been previously reported, demonstrating how a different type of analysis (univariate vs. multivariate) can generate new insights.

Taken together, in the above section I examined computational explorability by applying top-down and bottom-up analysis approaches for testing multivariate separation among groups of high or low relevance conducive or obstructive feedback on autonomic reactivity during feedback presentation.

8. Discussion

With the present study, I aimed to computationally reproduce the results of analyses from Kreibig et al. (2012). To demonstrate computational documentation and reproduction, I reported the results originating from reproduced computations (Section 3) and compiled a compendium comprising the data, statistical analysis scripts, and computational environment, available via the Open Science Framework (https://osf.io/mhnr6/). Under computational review and veracity, I

[3] Loadings less than .50 were not interpreted.

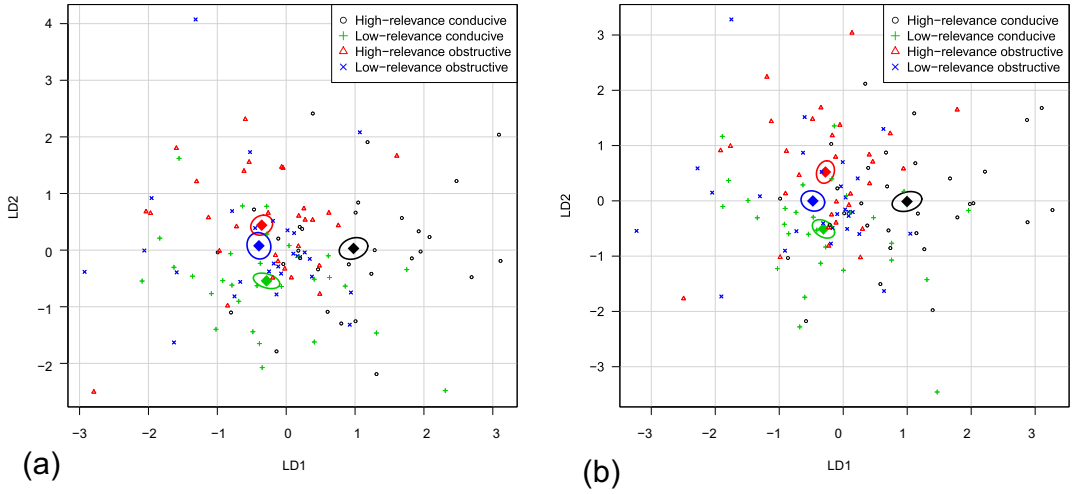

Fig. 2. Differentiation of experimental groups along linear discriminants (LD) 1 and 2 using the three-PC top-down (a) or four-PC bottom-up (b) multivariate analysis approaches. Diamonds show group centroids, dots individual subjects, and ellipses 1 standard error of the mean.

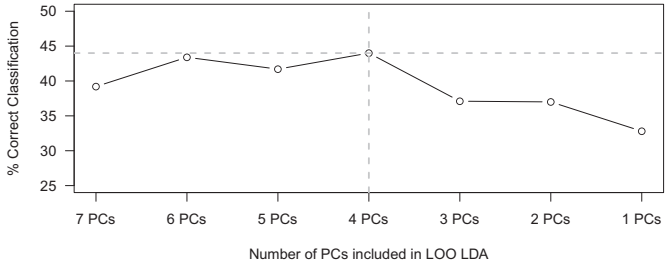

Fig. 3. Percentage correct classification for leave-one-out (LOO) linear discriminant analysis based on decreasing importance and number of Principal Components (PCs) included (see Table 5). Gray horizontal and vertical lines mark the value of maximal correct classification, 44%, achieved by the predictive discriminant analysis using the 4-PC solution.

Table 7
Linear discriminant analysis of four-Principle Component (PC) bottom-up solution maximizing group separation based on relative importance of PCs in univariate analysis of variance (see Table 6).

	Linear discriminants		
	LD1	LD2	LD3
Principal Components			
PC1	**0.71**	−0.39	−0.05
PC2	−0.27	**-0.59**	0.30
PC5	**-0.61**	**-0.60**	0.11
PC3	0.22	0.27	**0.87**
Variance explained			
Percentage	67.70	25.50	6.90
Cumulative	67.70	93.20	100.00

Note. Percentage – percentage of variance explained; cumulative – cumulative percentage of variance explained. Loadings typeset in boldface mark those which met the criterion of $r \geq .5$ and were used for interpreting LDs.

evaluated and quantified the degree of computational reproducibility between original and reproduced results and presented an analysis of computational error types (Section 4). To illustrate some of the added benefits of building on a computationally reproducible study, I explored implications of computational reproducibility using my data example: I evaluated the revisability of statistical analyses by exploring the impact of different α-adjustment methods and different participant subsamples on results and conclusions of the original study (Section 5). I investigated extensibility of statistical analyses by evaluating the sensitivity of newly derived composite feeling variables for indicating the predicted interaction effect of goal relevance and goal conduciveness and by wrapping a bootstrapping procedure around main analyses to estimate the replicability of reported results (Section 6). I examined computational explorability by applying top-down and bottom-up analysis approaches for testing multivariate group separation of autonomic predictors (Section 7). Based on analyses run in the context of computational explorability, in the below, I discuss new evidence for the motivational effects of emotion in response to performance feedback (Section 8.1). I also discuss computational reproducibility of this particular data example (Kreibig et al., 2012; Section 8.2), materials for computational reproducibility (Section 8.3), and a road map toward proper computational reproducibility practices (Section 8.4).

8.1. New evidence for the motivational effects of emotion in response to performance feedback

Under computational explorability (Section 7), I leveraged the multivariate structure of multiple autonomic variables for testing the interaction hypothesis of goal relevance and goal conduciveness on emotional reactivity in terms of autonomic reactivity. Both top-down (MANOVA) and bottom-up (PCA) multivariate analysis approaches demonstrated separation of feedback conditions into three clusters, separating high-relevance conducive, low-relevance conducive, and (high and low-relevance) obstructive feedback.

These results suggest that greater appraisal of goal relevance amplifies the emotional response to events appraised as conducive but not to those appraised as obstructive to one's goals. Events appraised as conducive may lead to a mobilization of effort relative to the appraised relevance of the event. However, events appraised as obstructive may lead to a withdrawal of effort irrespective of the appraised relevance of the event. This may be taken to suggest that more or less effort may be mobilized for engaging in a given task. However, if effort is withdrawn, all effort is withdrawn rather than withdrawing more—or less?—effort. Quitting may be an all-or-none rather than a gradual process. This differentiation is consistent with motivational intensity theory, which assumes a resource conservation approach to effort investment in task engagement depending on appraised task difficulty and self-relevance (Brehm and Self, 1989; Gendolla and Richter, 2010). It suggests a linear increase of effort mobilization with increasing task difficulty and self-relevance up to a level of maximally justified effort beyond which all effort is withdrawn (Gendolla and Krüsken, 2001; Silvestrini and Gendolla, 2013; Silvia et al., 2014). While effort mobilization and withdrawal have been demonstrated before (Annis et al., 2001) and during (Silvestrini and Gendolla, 2013) task execution, it has not been demonstrated during performance feedback after task execution.

Table 8
Classification Table based on predictive discriminant analysis with percentages (and case count in parentheses).

	Predicted group membership				
Actual group membership	High-relevance conducive	High-relevance obstructive	Low-relevance conducive	Low-relevance obstructive	**Sum**
High-relevance conducive	**58.1**	16.1	16.1	9.7	**26.1**
	(18)	(5)	(5)	(3)	**(31)**
High-relevance obstructive	13.3	**53.3**	26.7	6.7	**25.2**
	(4)	**(16)**	(8)	(2)	**(30)**
Low-relevance conducive	23.3	16.7	**43.3**	16.7	**25.2**
	(7)	(5)	**(13)**	(5)	**(30)**
Low-relevance obstructive	7.1	35.7	35.7	**21.4**	**23.5**
	(2)	(10)	(10)	**(6)**	**(28)**
Sum	**25.5**	**30.5**	**30.5**	**13.6**	**100.0**
	(31)	**(36)**	**(36)**	**(16)**	**(119)**

Note. Entries typeset in boldface in the diagonal represent successful classification, i.e., matching of actual and predicted group membership. Off-diagonal entries represent classification mistakes given the current classification rule.

Appraised task difficulty and goal conduciveness may play similar roles in determining effort in these contexts: Easy, intermediate, and difficult tasks or goal conduciveness justify (a certain degree of) effort investment; however, impossible tasks or goal obstruction do not justify any effort investment as they do not contribute to obtaining a goal.

Given the exemplary nature of analyses reported from the present study, results reported herein and conclusions drawn from them are limited by the fact that I did not systematically explore the full scope of different multivariate analysis strategies, including testing different classifiers (e.g., linear discriminant analysis versus quadratic discriminant analysis, and non-parametric algorithms such as *k*-nearest neighbor). Neither did I apply multivariate analyses to self-reported feelings. While univariate analyses indicated only the presence of conduciveness main effects, it would be very interesting to see whether and—if so—what pattern of interaction would be revealed if considering the multivariate structure of multiple feeling variables. There remains much to be explored in the context of the current data example.

8.2. Computational reproducibility of Kreibig et al. (2012)

Conceptualizing this study was an interesting exercise of retracing my own analyses and revisiting this data set. Based on applying computational reproducibility methods to my own work and exploring their implications, I introduced new terminology along the way (Section 1). Only time will tell how useful these concepts will be for introducing reproducibility practices to the field of the psychophysiology of motivation in particular and the scientific community at large. For procedural clarity for the present work, it seemed useful to delineate five different steps of computational reproducibility, of which I illustrated each with examples. To recapitulate, (1) *computational reproduction* generated results that left main findings unchanged (Section 3); (2) *computational review and veracity* indicated a reproducibility rate of 91% (Section 4); (3) *computational revisability* suggested robustness of results to minor to moderate modifications in computational methods (e.g., alternative α-adjustment methods, skin conductance non-responder definition; Section 5); (4) *computational extensibility* further underscored the stability of previously reported findings, both for alternative approaches to quantifying emotional reactivity (i.e., composite scores of subjective feelings) as well as for future replications with different subject samples (i.e., internal replicability analysis; Section 6); and (5) *computational explorability* illustrated the potential value of applying alternative analyses to the present data set that explored different features of it than previous analyses (i.e., multivariate analysis methods; Sections 7 and 8.1).

Different limitations apply to each of the steps of computational reproducibility: Computational reproduction, review, and veracity are first and foremost limited by the fact that I—as the first author of the prior study (Kreibig et al., 2012) rather than an independent researcher—reproduced analyses, albeit on different hardware and updated software. Second, for the purpose of the present paper, computational reproduction, review, and veracity only evaluated reproducibility of statistical analyses, not of preprocessing, although it is important to make the *entire* work flow reproducible.

For computational revisability, extensibility, and explorability, presented analyses are of exemplary nature only. Given the scope of this report, I was not able to address every possible issue. For example, in the context of computational revisability it remains for future research to examine more widespread issues like outlier exclusion criteria, which are often arbitrary and can notably change results.

More broadly, the present study itself only represents an example of evaluating computational reproducibility. Other issues of computational reproducibility may be found for other studies. Still, the present evaluation may be useful in identifying more general issues. For example, the present study singled out computational errors of different origin (Section 4). Knowledge of these different error sources may raise awareness of their potential presence and/or influence in future work.

8.3. Materials for computational reproducibility of Kreibig et al. (2012)

The materials to achieve computational reproducibility, of course, go beyond what is provided through the present report. I document analysis outcomes in detail in supplementary online materials (SOM, Section A). I provide the computational compendium through the Open Science Framework (https://osf.io/mhnr6/). It comprises the raw data, which have been deidentified and formated for data sharing. Making these available opens the possibility for future collaborations and to address additional research questions. The compendium also comprises the analysis scripts that expose the computations taken to arrive at the reported results, step by step. The compendium also comprises the computational environment, which must be scriptable and should ideally be freely available, open-source, and expandable (e.g., R Development Core Team, 2007). However, beyond this, there were still additional tools needed to achieve containerization of the computational environment (e.g., Docker; Boettiger, 2015) and appropriate versioning control (e.g., github).

There are many benefits to producing and using computationally reproducible work: Not only is this level of transparency the best way to reduce computational errors, conserve and share data, and document analyses, but it also increases research efficacy, impact, visibility, and portability (Vandewalle et al., 2009). Computational reproducibility improves work habits—merely knowing that analysis scripts will be available to others will motivate researchers to polish and improve these scripts to raise them to a higher level of quality than they would have been if "no one were looking," increasing confidence in the results

and improving collaboration (Donoho, 2010). Computational reproducibility also provides the best way to teach the computational methods and skills to our students by verified examples of real data from their respective fields of research. Computational reproducibility allows one to build on existing research and to create entirely new artifacts (Vandewalle et al., 2009). Computational reproducibility generates greater continuity and cumulative impact, as work products persist when students and postdocs move on to new labs. Finally, computational reproducibility allows researchers to make their product of publicly sponsored work available to everyone, as is required by responsible stewardship of public goods.

8.4. Road map toward proper computational reproducibility practices

Reproducible research is still in its infancy—and under-appreciated—in the domain of psychophysiology. A search for the keywords "computational reproducibility" and "psychophysiology" as of January 2017 returned no hits in the *International Journal of Psychophysiology*, *Biological Psychology*, *Psychophysiology*, on GoogleScholar, or Google. The term "reproducibility" alone has in the present journal been predominantly used in the context of measurement reliability: Internal consistency of forming composite scores of repetitions of items or trials (e.g., Narici, 1997); intra-individual reproducibility or test–retest reliability, evaluating correlation coefficients between repeated measurements separated by hours, weeks, or months (e.g., Lal and Craig, 2016); inter-individual reproducibility comparing the results of two independently conducted experiments with different samples (e.g., Burgess and Gruzelier, 1997); and inter-site reproducibility, examining results obtained from different measurement equipment and multiple laboratories (Sutton et al., 2009). The situation is similar in other psychophysiology journals (Kappenman and Keil, 2017). Tests of computational reproducibility are much more common in journals of other disciplines, including computer science (Črepinšek et al., 2014), biology (Garijo et al., 2013; Gentleman, 2005; Tibshirani et al., 2002), medicine (Kanwal et al., 2015), economics (Herndon et al., 2014), and even archeology (Marwick, 2016).

The Open Science Framework is slowly being adopted in order to improve transparency and facilitate data sharing in psychophysiological research: While two articles in this Journal refer to and endorse the OSF (Baldwin, 2017; Larson and Moser, 2017), none actually use it to share their data, code, and/or computational environment. Even methods publications still report mathematical procedures and algorithms in an appendix format, without routinely sharing their code (e.g., Cecotti and Ries, 2017); others link to institutional websites that host the described software (Groppe, 2017; Small et al., 2009). The situation is similar for *Biological Psychology*. In *Psychophysiology*, five reports from three different laboratories refer to the OSF for data sharing (Bradford et al., 2015; Elkins-Brown et al., 2016; Hefner et al., 2016; Kaye et al., 2016; klein Selle et al., 2015). Without use of a common archive or repository, it is impossible to search past publications for those which share their data, if there is no explicit indexing (American Psychological Association, 2017). Only reading and manual indexing can identify the few publications that do (e.g., Boekel et al., 2017). Data sharing seems much more common in other domains of psychology and science at large (Kidwell et al., 2016; Perrino et al., 2013).

The present study strives to provide a vision and concrete example of data analytic and reporting practices that may optimize computational reproducibility. It is my hope that future research can learn and benefit from the examples provided and, in the spirit of reproducibility, build on the concepts and procedures illustrated in the presented work.

Based on the above-presented work on computationally reproducing the data example from Kreibig et al. (2012), I conclude by providing a ten-step road map of best practices to facilitate computational reproducibility for future studies: (1) Write intelligible code: Write verbose rather than terse code that may be easily read and understood and use named constants rather than literal values (i.e., magic numbers). (2) Annotate your code: The more you include explanations in your code, the easier it will be for yourself or others to understand your code. (3) Use a version control system: When revising your code, you should not be afraid of overwriting old or outdated versions of your code. A version control system allows one to label, store, and retrieve prior versions of the code. (4) Keep a meticulous analysis log: Create an analysis script that, when run, produces the exact set of results presented in your manuscript. No manual fiddling allowed. (5) Practice code review to identify and eliminate computational errors: Explain your code to yourself or someone else. (6) Build on existing and validated code: Make use of open source projects wherever possible to allow maximum accessibility. (7) Write sustainable code: Explicitly document dependencies (and their versions) that are necessary for your code to run to allow others to build on your code. (8) Adopt software platforms that support reproducible research: Limit use of closed-source proprietary software and favor software that was built and is maintained by an open source community to make your code and computational environment accessible to the largest possible audience. (9) Avoid proprietary data formats in favor of accessible ones: To make your data accessible to other researchers, use simple and flexible file formats for data storage and exchange. (10) Capture a snapshot of your raw data, code, and computational environment and preserve it in a long-term archive: Ensurance of indefinite hosting of the computational compendium is critical to allow its future access. Several archives hosted by nonprofit foundations have evolved over the past decade, such as the OSF.

The field of the psychophysiology of motivation is seeing many new findings but their systematic replication is still few in numbers. Given the rapid pace with which it is moving forward, it behooves the field to pause and look back to evaluate itself. It is appropriate to question, revise, and update its procedures. I hope that I could demonstrate that computational reproducibility is a concept worth adopting, as it creates a strong foundation of research findings for future research to build on.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ijpsycho.2017.06.001.

References

American Psychological Association, Links to data sets and repositories. Retrieved Jan 30, 2017, from. http://www.apa.org/research/responsible/data-links.aspx.
Annis, S.S., Wright, R.A.R.A., Williams, B.J.B.J., 2001. Interactional influence of ability perception and task demand on cardiovascular response: appetitive effects at three levels of challenge. J. Appl. Behav. Res. 6, 82–107.
Armel, K.C.K.C., Ramachandran, V.S.V.S., 2003. Projecting sensations to external objects: evidence from skin conductance response. Proc. R. Soc. Lond. B Biol. Sci. 270 (1523), 1499–1506.
Armstrong, R.A.R.A., 2014. When to use the Bonferroni correction. Ophthalmic Physiol. Opt. 34 (5), 502–508.
Baggerly, K.K., 2016. Reporting scientific results and sharing scientific study data. In: Schwalbe, M.M. (Ed.), Statistical Challenges in Assessing and Fostering the Reproducibility of Scientific Results: Summary of a Workshop. National Academies Press, Washington, DC.
Bakeman, R.R., 2005. Recommended effect size statistics for repeated measures designs. Behav. Res. Methods 37, 379–384.
Baldwin, S.A.S.A., 2017. Improving the rigor of psychophysiology research. Int. J. Psychophysiol. 111, 5–16.
Berntson, G.G.G.G., Norman, G.J.G.J., Hawkley, L.C.L.C., Cacioppo, J.T.J.T., 2008. Cardiac autonomic balance versus cardiac regulatory capacity. Psychophysiology 45 (4), 643–652.
Betella, A.A., Verschure, P.F.P.F., 2016. The affective slider: a digital self-assessment scale for the measurement of human emotions. PloS one 11 (2), e0148037.
Blakesley, R.E.R.E., Mazumdar, S.S., Dew, M.A.M.A., Houck, P.R.P.R., Tang, G.G., Reynolds IIIC.F., C.F., Butters, M.A.M.A., 2009. Comparisons of methods for multiple hypothesis testing in neuropsychological research. Neuropsychology 23 (2), 255–264.
Boekel, W.W., Forstmann, B.B., Keuken, M.M., 2017. A test-retest reliability analysis of diffusion measures of white matter tracts relevant for cognitive control. Psychophysiology 54 (1), 24–33.
Boettiger, C.C., 2015. An introduction to Docker for reproducible research. ACM SIGOPS Oper. Syst. Rev. 49 (1), 71–79.

Bollen, K.K., Cacioppo, J.J., Kaplan, R.R., Krosnick, J.J., Olds, J.J., 2015. Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science: Report of the Subcommittee on Replicability in Science, Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences. Retrieved from the National Science Foundation Web site: www.nsf.gov/sbe/AC_Materials/SBE_Robust_and_Reliable_Research_Report.pdf.
Bonferroni, C.E.C.E., 1935. Studi in Onore del Professore Salvatore Ortu Carboni. Tipografia del Senatopp. 13–60.
Bonferroni, C.E.C.E., 1936. Teoria statistica delle classi e calcolo delle probabilita. In: Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze. 8. pp. 3–62.
Boucsein, W.W., 1992. Electrodermal Activity. Plenum Press, New York.
Boucsein, W.W., 2012. Electrodermal Activity. Springer Science & Business Media.
Bradford, D.E.D.E., Starr, M.J.M.J., Shackman, A.J.A.J., Curtin, J.J.J.J., 2015. Empirically based comparisons of the reliability and validity of common quantification approaches for eyeblink startle potentiation in humans. Psychophysiology 52 (12), 1669–1681.
Bradley, M.M.M.M., Lang, P.J.P.J., 1994. Measuring emotion: the self-assessment manikin and the semantic differential. J. Behav. Ther. Exp. Psychiatry 25, 49–59.
Brehm, J.W.J.W., Self, E.A.E.A., 1989. The intensity of motivation. Annu. Rev. Psychol. 40, 109–131.
Burgess, A.A., Gruzelier, J.J., 1997. How reproducible is the topographical distribution of EEG amplitude? Int. J. Psychophysiol. 26 (1), 113–119.
Cecotti, H.H., Ries, A.J.A.J., 2017. Best practice for single-trial detection of event-related potentials: application to brain-computer interfaces. Int. J. Psychophysiol. 111, 156–169.
Chirigati, F.F., Shasha, D.D., Freire, J.J., 2013. Reprozip: using provenance to support computational reproducibility. In: Presented as part of the 5th USENIX Workshop on the Theory and Practice of Provenance. USENIX, Berkeley, CA. Retrieved from https://www.usenix.org/conference/tapp13/reprozip-usingprovenancesupportcomputationalreproducibilitythe.
Claerbout, J.J., Karrenbach, M.M., 1992. Electronic documents give reproducible research a new meaning. In: Proceedings of the Society of Exploration Geophysics, October.
Cohen, J.J., 1988. Statistical power analysis for the behavioral sciences, 2nd ed. Erlbaum, Hillsdale, N. J.
Collins, H.H., 1992. Changing Order: Replication and Induction in Scientific Practice. University of Chicago Press.
Črepinšek, M.M., Liu, S.-H.S.-H., Mernik, M.M., 2014. Replication and comparison of computational experiments in applied evolutionary computing: common pitfalls and guidelines to avoid them. Appl. Soft Comput. 19, 161–170.
Cumming, G.G., 2008. Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. Perspect. Psychol. Sci. 3 (4), 286–300.
Cumming, G.G., Maillardet, R.R., 2006. Confidence intervals and replication: where will the next mean fall? Psychol. Methods 11 (3), 217–227.
Donoho, D.L.D.L., 2010. An invitation to reproducible computational research. Biostatistics 11 (3), 385–388.
Donoho, D.L.D.L., Maleki, A.A., Rahman, I.U.I.U., Shahram, M.M., Stodden, V.V., 2009. Reproducible research in computational harmonic analysis. Comput. Sci. Eng. 1, 8–18.
Efron, B.B., 1979. Bootstrap methods: another look at the jackknife. Ann. Stat. 7, 1–26.
Elkins-Brown, N.N., Saunders, B.B., Inzlicht, M.M., 2016. Error-related electromyographic activity over the corrugator supercilii is associated with neural performance monitoring. Psychophysiology 53 (2), 159–170.
Gallopoulos, E.E., Houstis, E.E., Rice, J.J., 1994. Computer as thinker/doer: problem-solving environments for computational science. IEEE Comput. Sci. Eng. 1 (2), 11–23.
Garijo, D.D., Kinnings, S.S., Xie, L.L., Xie, L.L., Zhang, Y.Y., Bourne, P.E.P.E., Gil, Y.Y., 2013. Quantifying reproducibility in computational biology: the case of the tuberculosis drugome. PloS one 8 (11), e80278.
Gendolla, G.H.E.G.H.E., Krüsken, J.J., 2001. The joint impact of mood state and task difficulty on cardiovascular and electrodermal reactivity in active coping. Psychophysiology 38, 548–556.
Gendolla, G.H.E.G.H.E., Richter, M.M., 2010. Effort mobilization when the self is involved: some lessons from the cardiovascular system. Rev. Gen. Psychol. 14, 212–226.
Gentleman, R.R., 2005. Reproducible research: a bioinformatics case study. Stat. Appl. Genet. Mol. Biol. 4 (1), 1–23.
Groppe, D.M.D.M., 2017. Combating the scientific decline effect with confidence (intervals). Psychophysiology 54 (1), 139–145.
Hefner, K.R.K.R., Verona, E.E., Curtin, J.J., et al., 2016. Emotion regulation during threat: parsing the time course and consequences of safety signal processing. Psychophysiology 53 (8), 1193–1202.
Herndon, T.T., Ash, M.M., Pollin, R.R., 2014. Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. Camb. J. Econ. 38 (2), 257–279.
Hewes, D.E.D.E., 2003. Methods as tools. Hum. Commun. Res. 29 (3), 448–454.
Hsu, J.J., 1996. Multiple Comparisons. Theory and Methods. Chapman & Hall, London.
Hullett, C.R.C.R., 2007. Concerns about error go beyond cumulating Type I error: a response to Matsunaga. Commun. Methods and Meas. 1 (4), 275–279.
Ioannidis, J.P.J.P., Khoury, M.J.M.J., 2011. Improving validation practices in “omics” research. Science 334 (6060), 1230–1232.
Ito, T.T., Larsen, J.T.J.T., Smith, N.K.N.K., Cacioppo, J.T.J.T., 1998. Negative information weighs more heavily on the brain: the negativity bias in evaluative categorization. J. Pers. Soc. Psychol. 75, 887–900.
Kanwal, S.S., Lonie, A.A., Sinnott, R.O.R.O., Anderson, C.C., 2015. Challenges of large-scale biomedical workflows on the cloud—A case study on the need for reproducibility of results. In: Computer-Based Medical Systems (CBMS), 2015 IEEE 28th International Symposium on, pp. 220–225.
Kappenman, E.S.E.S., Keil, A.A., 2017. Introduction to the special issue on recentering science: replication, robustness, and reproducibility in psychophysiology. Psychophysiology 54 (1), 3–5.
Kaye, J.T.J.T., Bradford, D.E.D.E., Curtin, J.J.J.J., 2016. Psychometric properties of startle and corrugator response in NPU, affective picture viewing, and resting state tasks. Psychophysiology 53 (8), 1241–1255.
Kidwell, M.C.M.C., Lazarević, L.B.L.B., Baranski, E.E., Hardwicke, T.E.T.E., Piechowski, S.S., Falkenberg, L.-S.L.-S., Kennett, C.C., Slowik, A.A., Sonnleitner, C.C., Hess-Holden, C.C., et al., 2016. Badges to acknowledge open practices: a simple, low-cost, effective method for increasing transparency. PLoS Biol 14 (5), e1002456.
klein Selle, N.N., Verschuere, B.B., Kindt, M.M., Meijer, E.E., Ben-Shakhar, G.G., et al., 2015. Orienting versus inhibition in the concealed information test: different cognitive processes drive different physiological measures. Psychophysiology 53 (4), 579–590.
Kreibig, S.D.S.D., Gendolla, G.H.E.G.H.E., Scherer, K.R.K.R., 2012. Goal relevance and goal conduciveness appraisals lead to differential autonomic reactivity in emotional responding to performance feedback. Biol. Psychol. 91, 365–375. http://dx.doi.org/10.1016/j.biopsycho.2012.08.007.
Kreibig, S.D.S.D., Samson, A.C.A.C., Gross, J.J.J.J., 2015. The psychophysiology of mixed emotional states: internal and external replicability analysis of a direct replication study. Psychophysiology 52, 873–886.
Kron, A.A., Pilkiw, M.M., Banaei, J.J., Goldstein, A.A., Anderson, A.K.A.K., 2015. Are valence and arousal separable in emotional experience? Emotion 15 (1), 35–44.
Lake, J.I.J.I., Meck, W.H.W.H., LaBar, K.S.K.S., 2016. Discriminative fear learners are resilient to temporal distortions during threat anticipation. Timing Time Percept. 4 (1), 63–78.
Lal, S.K.S.K., Craig, A.A., 2005. Reproducibility of the spectral components of the electroencephalogram during driver fatigue. Int. J. Psychophysiol. 55 (2), 137–143.
Larsen, J.T.J.T., Norris, C.J.C.J., Cacioppo, J.T.J.T., 2003. Effects of positive and negative affect on electromyographic activity over *zygomaticus major* and *corrugator supercilii*. Psychophysiology 40, 776–785.
Larsen, J.T.J.T., Hershfield, H.E.H.E., Stastny, B.J.B.J., Hester, N.N., 2016. On the relationship between positive and negative affect: their correlation and their co-occurrence. Emotion 17, 323–336.
Larson, M.J.M.J., Moser, J.S.J.S., 2017. Rigor and replication: towards improved best practices in human electrophysiology research. Int. J. Psychophysiol. 111, 1–4.
Leek, J.T.J.T., Peng, R.D.R.D., 2015. Opinion: reproducible research can still be wrong: adopting a prevention approach. Proc. Natl. Acad. Sci. 112 (6), 1645–1646.
Marwick, B.B., 2016. Computational reproducibility in archaeological research: basic principles and a case study of their implementation. J. Archaeol. Method Theory 1–27.
Matsunaga, M.M., 2007. Familywise error in multiple comparisons: disentangling a knot through a critique of O'Keefe's arguments against alpha adjustment. Commun. Methods Meas. 1 (4), 243–265.
McKubre, M.C.M.C., 2008. The importance of replication. In: 14th International Conference on Cold Fusion—International Conference on Condensed Matter Nuclear Science, Washington, DC.
Narici, L.L., 1997. Driven and synchronized brain activities in the α band: a neuromagnetic test for frequency responsiveness. Int. J. Psychophysiol. 26 (1), 137–148.
Nummenmaa, L.L., Niemi, P.P., 2004. Inducing affective states with success-failure manipulations: a meta-analysis. Emotion 4, 207–214.
O'Keefe, D.J.D.J., 2003a. Colloquy: should familywise alpha be adjusted? Hum. Commun. Res. 29 (3), 431–447.
O'Keefe, D.J.D.J., 2003b. Searching for a defensible application of alpha-adjustment tools. Hum. Commun. Res. 29 (3), 464–468.
O'Keefe, D.J.D.J., 2007. Responses to Matsunaga: it takes a family—a well-defined family—to underwrite familywise corrections. Commun. Methods Meas. 1 (4), 267–273.
Peng, R.D.R.D., 2009. Reproducible research and biostatistics. Biostatistics 10 (3), 405–408.
Peng, R.D.R.D., 2011. Reproducible research in computational science. Science 334 (6060), 1226–1227.
Peng, R.D.R.D., Dominici, F.F., Zeger, S.L.S.L., 2006. Reproducible epidemiological research. Am. J. Epidemiol. 163, 783–789.
Perrino, T.T., Howe, G.G., Sperling, A.A., Beardslee, W.W., Sandler, I.I., Shern, D.D., Pantin, H.H., Kaupert, S.S., Cano, N.N., Cruden, G.G., et al., 2013. Advancing science through collaborative data sharing and synthesis. Perspect. Psychol. Sci. 8 (4), 433–444.
Development Core Team, R.R., 2007. R: a language and environment for statistical computing [Computer software and manual]. In: R Foundation for Statistical Computing, Vienna, Austria (Retrieved from www.r-project.org).
Rosnow, R.L.R.L., Rosenthal, R.R., 1989. Definition and interpretation of interaction effects. Psychol. Bull. 105, 143–146.
Rosnow, R.L.R.L., Rosenthal, R.R., 1995. “Some things you learn aren't so”: Cohen's paradox, Asch's paradigm, and the interpretation of interaction. Psychol. Sci. 6, 3–9.
Sandusky, R.J.R.J., Tenopir, C.C., Casado, M.M.M.M., 2007. Uses of figures and tables from scholarly journal articles in teaching and research. Proc. Am. Soc. Inform. Sci. Tech. 44 (1), 1–13.
Sankoh, A.J.A.J., Huque, M.F.M.F., Dubey, S.D.S.D., 1997. Some comments on frequently used multiple endpoint adjustment methods in clinical trials. Stat. Med. 16, 2529–2542.
Scherer, K.R.K.R., 2001. Appraisal considered as a process of multi-level sequential checking. In: Scherer, K.R.K.R., Schorr, A.A., Johnstone, T.T. (Eds.), Appraisal processes in emotion: Theory, Methods, Research. Oxford University Press, New York

and Oxford, pp. 92–120.
Schimmack, U.U., 2001. Pleasure, displeasure, and mixed feelings: are semantic opposites mutually exclusive? Cognit. Emot. 15, 81–97.
Schmidt, S.S., 2009. Shall we really do it again? The powerful concept of replication is neglected in the social sciences. Rev. Gen. Psychol. 13 (2), 90–100.
Schwalbe, M.M., et al., 2016. Statistical Challenges in Assessing and Fostering the Reproducibility of Scientific Results: Summary of a Workshop. National Academies Press, Washington, DC. Retrieved from https://www.nap.edu/read/21915/chapter/1.
Shaffer, J.P.J.P., 1995. Multiple hypothesis testing. Annu. Rev. Psychol. 46, 561–584.
Siller, A.A., Ambach, W.W., Vaitl, D.D., 2015. Investigating expectation effects using multiple physiological measures. Front. Psychol. 6.
Silvestrini, N.N., Gendolla, G.H.G.H., 2013. Automatic effort mobilization and the principle of resource conservation: one can only prime the possible and justified. J. Pers. Soc. Psychol. 104 (5), 803–816.
Silvia, P.J.P.J., Moore, L.C.L.C., Nardello, J.L.J.L., 2014. Trying and quitting: how self-focused attention influences effort during difficult and impossible tasks. Self Identity 13 (2), 231–242.
Sloan, B.M.B.M., McCorkle, D.S.D.S., Bryden, K.M.K.M., 2013. An overview of computational environments for engineering. In: Ames Laboratory Conference Papers, Posters, and Presentations, Paper 1, . Retrieved from http://lib.dr.iastate.edu/ameslab_conf/1.
Small, S.L.S.L., Wilde, M.M., Kenny, S.S., Andric, M.M., Hasson, U.U., 2009. Database-managed Grid-enabled analysis of neuroimaging data: the CNARI framework. Int. J. Psychophysiol. 73 (1), 62–72.
Spottiswoode, S.J.P.S.J.P., May, E.E., 2003. Skin conductance prestimulus response: Analyses, artifacts and a pilot study. J. Sci. Explor. 17 (4), 617–641.
Stodden, V.V., 2015. Reproducibility. In: Brockman, M.J.M.J. (Ed.), This Idea Must Die: Scientific Theories that are Blocking Progress. Harper Collins.
Straube, E.R.E.R., 1979. On the meaning of electrodermal nonresponding in schizophrenia. J. Nerv. Ment. Dis. 167 (10), 601–611.
Suchard, M.M., 2016. Assessment of factors affecting reproducibility. In: Schwalbe, M.M. (Ed.), Statistical Challenges in Assessing and Fostering the Reproducibility of Scientific Results: Summary of a Workshop. National Academies Press, Washington, DC, pp. 55–57.
Sutton, B.P.B.P., Ouyang, C.C., Karampinos, D.C.D.C., Miller, G.A.G.A., 2009. Current trends and challenges in MRI acquisitions to investigate brain function. Int. J. Psychophysiol. 73 (1), 33–42.
Tabachnick, B.G.B.G., Fidell, L.S.L.S., 2013. Using Multivariate Statistics, 6th ed. Pearson, Boston, MA.
Tibshirani, R.J.R.J., Efron, B.B., et al., 2002. Pre-validation and inference in microarrays. Stat. Appl. Genet. Mol. Biol. 1 (1), 1000.
Tutzauer, F.F., 2003. On the sensible application of familywise alpha adjustment. Hum. Commun. Res. 29 (3), 455–463.
van der Zwaag, M.D.M.D., Westerink, J.H.J.H., van den Broek, E.L.E.L., 2011. Emotional and psychophysiological responses to tempo, mode, and percussiveness. Music. Sci. 15 (2), 250–269.
Vandewalle, P.P., Kovacevic, J.J., Vetterli, M.M., 2009. Reproducible research in signal processing: what, why, and how. IEEE Signal Process. Mag. 26, 37–47.
Venables, P.H.P.H., 1978. Psychophysiology and psychometrics. Psychophysiology 15 (4), 302–315.
Venables, P.H.P.H., Christie, M.J.M.J., 1980. Electrodermal activity. In: Martin, I.I., Venables, P.H.P.H. (Eds.), Techniques in Psychophysiology. Wiley, Chichester, UK, pp. 3–67.
Venables, W.N.W.N., Ripley, B.D.B.D., 2002. Modern Applied Statistics with S. Spinger-Verlag, New York.
Wasserman, S.S., Bockenholt, U.U., 1989. Bootstrapping: applications to psychophysiology. Psychophysiology 26 (2), 208–221.
Weber, R.R., 2007. Responses to Matsunaga: to adjust or not to adjust alpha in multiple testing: that is the question. Guidelines for alpha adjustment as response to O'Keefe's and Matsunaga's critiques. Commun. Methods Meas. 1 (4), 281–289.
Wright, R.A.R.A., 1996. Brehm's theory of motivation as a model of effort and cardiovascular response. In: Gollwitzer, P.M.P.M., Bargh, J.A.J.A. (Eds.), The Psychology of Action: Linking Cognition and Motivation to Behavior. Guilford, New York, pp. 424–453.