Research Article
doi.org/10.1002/ansa.202000159
Chemistry Europe
European Chemical Societies Publishing

Received: 27 November 2020
Revised: 30 May 2022
Accepted: 30 May 2022

Helping reviewers assess statistical analysis: A case study from analytic methods

Ron S. Kenett[1] | **Bernard G. Francq**[2]

[1]The KPA group and the Samuel Neaman Institute, Technion, Haifa, Israel

[2]UCLouvain, ISBA (Institute of Statistics, Biostatistics and Actuarial Sciences), Louvain la Neuve, Belgium

Correspondence
Ron S. Kenett, The KPA group and the Samuel Neaman Institute, Technion, Israel.
Email: ron@kpa-group.com

Abstract

Analytic methods development, like many other disciplines, relies on experimentation and data analysis. Determining the contribution of a paper or report on a study incorporating data analysis is typically left to the reviewer's experience and good sense, without reliance on structured guidelines. This is amplified by the growing role of machine learning driven analysis, where results are based on computer intensive algorithm applications. The evaluation of a predictive model where cross validation was used to fit its parameters adds challenges to the evaluation of regression models, where the estimates can be easily reproduced. This lack of structure to support reviews increases uncertainty and variability in reviews. In this paper, aspects of statistical assessment are considered. We provide checklists for reviewers of applied statistics work with a focus on analytic method development. The checklist covers six aspects relevant to a review of statistical analysis, namely: (1) study design, (2) algorithmic and inferential methods in frequentism analysis, (3) Bayesian methods in Bayesian analysis (if relevant), (4) selective inference aspects, (5) severe testing properties and (6) presentation of findings. We provide a brief overview of these elements providing references for a more elaborate treatment. The robustness analysis of an analytical method is used to illustrate how an improvement can be achieved in response to questions in the checklist. The paper is aimed at both engineers and seasoned researchers.

KEYWORDS
bayesian analysis, data analysis, frequentism, information quality, review checklists, severe testing

1 | BACKGROUND

In the pharmaceutical industry, as well as in other contexts, reviewers provide feedback aimed at improving work based on statistical data analysis. A good reviewer is one who contributes to the analysis and constructively enhances its level. Some journals in medicine publish guidelines for such reviews.[1,2] We discuss here the review of papers or reports based on statistical analysis rather than mathematical modelling. We use, as a case study, a publication on the design of a high performance liquid chromatography (HPLC) analytic method. Section 2 provides a perspective on the review of statistical reports, Section 3 is a review of statistical analysis methods, and Section 4 presents two checklists. Section 5 is a case study based on the design of an HPLC analytic method. The paper concludes with a discussion.

Research Article
doi.org/10.1002/ansa.202000159
Chemistry Europe
European Chemical Societies Publishing

2 | SOME PERSPECTIVES ON THE STATISTICAL REVIEW OF APPLIED STATISTICS

An important aspect of reviewing applied statistics is related to the reproducibility of the research findings. Part of this has been addressed by a much-discussed American statistical association (ASA) statement on *p*-values.[3] While the conclusions of applied research papers must be supported by data statistical analysis, *p*-values (together with confidence intervals [CIs]) are, usually, mandatory in publications as evidence supporting alignment with the conclusions. The ASA statement formulates six principles for statistical analysis:

Principle 1: *p*-values can indicate how incompatible the data are with a specified statistical model.

Principle 2: *p*-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.

Principle 3: Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold.

Principle 4: Proper inference requires full reporting and transparency.

Principle 5: A *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result.

Principle 6: By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis.

Other approaches, mentioned in the ASA statement without critical appraisal, include (1) confidence intervals, (2) prediction intervals, (3) estimation, (4) likelihood ratios, (5) Bayesian methods, (6) Bayes factor and (7) credibility intervals.

Should these principles guide the review process of applied research papers in general? The answer is highly debated. A series of papers and blogs present contrarian and supporting views to these principles and approaches (e.g.,[4–8] some journals adopt opinionated guidelines affecting statistical analysis in papers they publish).[9,10] Specifically, they adopt a policy whereby null hypothesis statistical testing is not to be used. Mayo[11] characterized these debates as 'the statistics wars'. Much discussion has focused on misuses of the null hypothesis testing process and low powered studies. Daniel Kahneman, the Nobel Prize winner behavioral economist retracted from his book the mention of several studies retrospectively found to be based on underpowered studies (https://retractionwatch.com/2017/02/20/placed-much-faith-underpowered-studies-nobel-prize-winner-admits-mistakes/). The studies themselves were however not retracted from the journals that originally published them. The lack of retraction is another important element in the reproducibility discussion, see, for example, in medical research.[12] The rate of retraction has been estimated (from a study published in 2011) to 0.02% in biomedical fields with nearly half due to honest error or non-replicable findings.[13] Retraction is practically unknown in the industrial method application setting.

Given this background, how should a reviewer assess the statistical analysis of applied research? The next section maps out statistical characteristics of applied research. Later we focus on a case study in the development of analytic methods.

3 | Statistical analysis of applied research

Efron and Hastie[14] present a comprehensive review of statistical analysis over time. At the origin, classical statistics consists of an algorithmic and an inferential part. Frequentism (or 'objectivism') is based on the probabilistic properties of a procedure of interest, as derived and applied to observed data. This provides us with an assessment of bias and variance. The frequentists' interpretation is based on a scenario, where the same situation is repeated, endlessly. Within the frequentism framework, several methods can be applied: (1) the plug-in substitution principle, (2) the delta methods Taylor series approximation, (3) the application of parametric families and maximum likelihood theory, (4) the use of simulation and bootstrapping computer intensive numerical methods and (5) pivotal statistics.[15] These distinctions are important for reviewers to make. The Neyman–Pearson lemma provides an optimum hypothesis testing algorithm, where a black and white decision is made. With this approach you either reject the null hypothesis while testing for an alternative hypothesis, or not. This offers an apparently simple and effective way to conduct statistical inference that can be scaled up. On the other hand, conficence intervals (CIs) are considered by many as more informative. However, like *p*-value hacking, Barnett and Wren[16] demonstrate the wide prevalence of CI hacking. When a *p*-value is lower than the significance level (usually 5%), the test is said to be significant. When researchers strive to get significant (low) *p*-values, they hope to find CIs that do not overlap the null hypothesis. Specifically, 'the set of all CIs at different levels of probability . . . (yields a) confidence distribution',[17] (p. 363).

Alternatively, statistical analysis can be conducted within a Bayesian framework by transforming a prior distribution on the parameters of interest, to a posterior, using the observed data. In this framework, one often invokes the Bayes factor, which is a likelihood ratio of the marginal likelihood of two competing hypotheses, usually a null and an alternative. The Bayes factor is a sort of Bayesian alternative to classical hypothesis testing. In computer age analytics one distinguishes between algorithms aiming at (1) estimation, (2) prediction or (3) explanations of structure in the data. Estimation is assessed by accuracy of estimators, prediction by prediction error, and explanations are based on variable selection using variance bias tradeoffs, penalized regression and regularization criteria.

Mayo[11] presents a perspective on statistical inference based on the concept of severe testing; she labels it 'error statistics philosophy'. For error statisticians, a claim, or research finding, is severely tested if it has been subjected to and passes a test that probably would have found flaws, were they present,[11] (p. xii). If little or nothing has been done to rule out flaws in inferring a claim, then it has not passed a severe test. Mayo identifies three types of models: primary models, experimental models and data models. Primary models break down a research question into a set of local hypotheses that can be investigated using reliable methods. Experimental models structure the particular models at hand

Research Article
doi.org/10.1002/ansa.202000159
Chemistry Europe
European Chemical Societies Publishing

and serve to link primary models to data models. Data models generate and model raw data, as well as checking whether the data satisfy the assumptions of the experimental models. Error statistical assessments pick up on the effects of data dredging, multiple testing, optional stopping and other biasing selection effects. Biasing selection effects are blocked in error statistical accounts because they preclude control of error probabilities. Error statistical accounts require a preregistration of the study.[18,19] Long-run performance requirements are only necessary and not sufficient for severity. Long-run behavior could be satisfied with error probabilities that do not reflect well-testedness. Tools that are typically justified, because they control the probability of erroneous inferences in the long-run, are given an inferential justification. It is only when long-run relative frequencies represent the method's capability to discern mistaken interpretations of data that the performance, and severe testing goals are reached. Mayo[11] presents a range of conceptual methods for severe testing: 'bad evidence, no test' (BENT), probabilism, performance and probativeness. Insevere tests yield BENT. *Performance* is about controlling the relative frequency of erroneous inferences in the long run of applications. *Probabilism* views probability as a means of assigning degrees of belief, support or plausibility to hypotheses. *Probativeness* is scrutinizing BENT science by the severity criterion. In interpreting CIs, one needs to connect actual experiments with hypothesized concepts. In general, the reported analysis should be able to pinpoint the sources of failed predictions and indicate what is/is not learned from negative results.[20] Every reported inference should include what cannot be reliably inferred, what potential mistakes were not probed or ruled out, and what gaps would need checking in order to avoid various misinterpretations of results, Mayo,[11] (p. 437). A podcast with Mayo on severe testing is available at https://mattasher.com/2020/11/23/ep-26-deborah-mayo-on-error-replication-and-severe-testing/. An applet for severity testing assessment is available in https://richarddmorey.shinyapps.io/severity/.

Another aspect, to be considered in reviewing an applied research paper, is study design. Some studies are based on observational data and some on interventions, or experiments, designed by the researchers. There are many publications on statistical methods to design experimental interventions. The following illustration is adapted from Kenett and Zacks.[21] Interventions are determined by factor level combinations, the effects measured through responses. One particular aspect in this methodology is the use of blocking and randomization, which aims at increasing the precision of the estimates and ensures the validity of the inference. As these aspects are ubiquitous in study design, we discuss them with some more details. Blocking is used to reduce errors. A block is a portion of the experimental material that is expected to be more homogeneous than the whole aggregate. An example of blocking is the boy's shoes example,[22] (p. 97). Two kinds of shoe soles' materials are to be tested by fixing the soles on *n* pairs of boys' shoes and measuring the amount of wear of the soles after a period of actively wearing the shoes. Since there is high variability in activity of boys, if *m* pairs will be with soles of one type and the rest of the other, it will not be clear whether any difference that might be observed in the degree of wear out is due to differences between the characteristics of the sole material or to the differences between the boys. By blocking by pair of shoes, we can reduce much of the variability. Each pair of shoes is assigned the two types of soles. The comparison within each block is free of the variability between boys. Furthermore, since boys use their right or left foot differently, one should assign the type of soles to the left or right shoes at random. Thus, the treatments (two types of soles) are assigned within each block at random. An analytic device with two columns is equivalent to the boys' feet. Other examples of blocks could be equipment, laboratory personnel or days of the week. Generally, if there are *t* treatments to compare, and *b* blocks, and if all *t* treatments can be performed within a single block, we assign all the *t* treatments to each block. The order of applying the treatments within each block should be randomized. Such a design is called a randomized complete block design. If not, all treatments can be applied within each block; it is desirable to assign treatments to blocks in some balanced fashion. Such designs are called balanced incomplete block designs. Randomization within each block is validating the assumption that the error components in the statistical model are independent. This assumption may not be valid if treatments are not assigned at random to the experimental units within each block. If factors are hard to change, a design based on split plots will prove more effective and accommodating to the logistic constraints. Of course, you can have a good experimental plan with attention to power, use of blocks, etc., but, overall, a bad experiment because the conditions were not chosen realistically, or because the wrong outcomes were measured, or you had the right outcomes but the wrong measurement instruments.

Yet another aspect of statistical analysis, with a potentially strong impact on the results, is selective inference. Selective inference is inference on a selected subset of the parameters that turned out to be of interest, *after* viewing the data. This selection leads to difficulties in reproducibility of results and needs to be accounted for and controlled in the statistical analysis. We can distinguish between out-of-study and in-study selection. The former is not evident in the published work and is due to publication bias, *p*-hacking or other forms of significance chasing. The in-study selection can be more evident in the published work. This is reflected by selection choices in abstract content, table, figure or in highlighting results passing a threshold.[23,24] Attentive reviewers of analytic work should be looking for such selective inference.

Finally, findings have to be presented and generalized. Generalization can be achieved by a range of methods, some intuitive, some conceptual and some more formal, invoking, for example, causal arguments.[25,26] Findings can be presented in different ways. One approach is based on alternative verbal representations, some with meaning equivalence and some with surface similarity.[26] Verbal expression of research findings has been proposed in Greenland[27] and Yarkoni.[28] Alternative verbal statements, with meaning equivalence, represent the same conceptual statement. Alternatives with surface similarity seem similar to the target conceptual statement but have different meaning. This approach generates a table of alternative representations with a boundary of meaning (BOM).[26,29] The BOM is a demarcation line between claims, presented in alternative ways, and seemingly similar representations of findings not supported by

Research Article
doi.org/10.1002/ansa.202000159
Chemistry Europe
European Chemical Societies Publishing

Target statement	Meaning equivalence findings included in BOM	Surface similarity findings not included in BOM
Finding 1: The quality of life of patients and families affected with a food allergy to staple foods (milk, egg, sesame, peanut) is impaired	Food allergy in children impacts negatively on the day-to-day activities of the whole family	Educating patients on strict avoidance and carrying an epinephrine autoinjector is completely effective in avoiding accidental exposures in preschool children
	The incidence of accidental exposures to allergenic foods in preschool children is high	
	The currently recommended management of food allergy in children is patient education, strict avoidance, and carrying an epinephrine autoinjector	
Finding 2: All children suspected of an allergic reaction to foods should be referred to a center that includes appropriate facilities, medical, and support staff experienced in the diagnosis and treatment of children with food allergies as early as possible	The diagnosis of food allergy in children should be performed soon after the suspected event	Recommending strict avoidance of suspected allergenic foods is the best treatment for all young food allergic children
	There are no age limitations on the performance of diagnostic allergy tests, such as SPTs or observed food challenges, provided these are performed by well trained and experienced medical teams	Laboratory test such as sIgE to food can accurately diagnose food allergy in children
Finding 3: The natural history of CM allergy in children is still favorable as in most—it seems to resolve with time	The median age at resolution of CMA (by which time 50% of children have resolved their allergies) is between 6 and 8 years	Food allergy in children resolves in the first years of life
	Children with CMA and a positive family history of atopy, an initial anaphylactic reaction, recurrent wheezing or moderate/severe atopic dermatitis are less likely to resolve their CMA	Avoidance of allergenic foods is beneficial in preventing food allergy in children
Finding 4: A majority of children with IgE mediated CMA are capable of consuming certain amounts of EHBM proteins	Some children with CMA can develop immediate, life-threatening reactions to the ingestion of EHBM	Families of children with IgE-mediated CMA should be encouraged to try baked milk at home
	A minority of children with CMA are allergic also to heat denatured milk products. These are the most severely affected and least likely to resolve their allergies	All forms of heated and baked milk are similarly safe
Finding 5: In preschool children with CMA capable of ingesting EHBM safely, SGEP seems to promote earlier resolution	The median age at CMA resolution of preschool children, capable of ingesting EHBM safely and treated with SGEP including EHBM, seems to be significantly lower than in children treated with avoidance	Preschool children capable of ingesting EHBM safely and treated with SGEP including EHBM are developing true long-term tolerance to milk
	Most preschool children capable of ingesting EHBM safely and treated with SGEP including EHBM will be able to tolerate milk in their regular diet before entering school	EHBM is not a form of oral immunotherapy in food allergic children and therefore the follow- up recommended for these children is similar to patients with natural resolution of CMA (none)
Finding 6: A protocol of SGEP including EHBM, seems safe in children <4 years of age	A protocol of SGEP, including EHBM, performed by medical teams trained and experienced in the treatment of food allergy in children is safe	All children with IgE-mediated CMA should be treated with an SGEP with EHBM

FIGURE 1 Generalization of findings with alternative representation and a boundary of meaning (BOM) (adapted from Efron et al.[30])

the research. An example from Efron et al.[30] is shown in Figure 1. It describes findings from a study on the management of hypersensitivity reactions to non-steroidal anti-inflammatory drugs in children and adolescents and a structured gradual exposure protocol to baked and heated milk in the treatment of milk allergy. As an example, statements such as: 'The quality of life of patients and families affected with a food allergy to staple foods (milk, egg, sesame, peanut) is impaired' and 'Food allergy in children impacts negatively on day to day activities of the whole family' are considered equivalent in meaning. On the other hand, statements such as: 'Food allergy in children impacts negatively on day to day activities of the whole family' and 'Educating patients on strict avoidance and carrying an epinephrine autoinjector is completely effective in avoiding accidental exposures in preschool children activities of the whole family' carry only surface similarity. These alternatives were formulated by the researchers. The BOM is the demarcation line between the columns with meaning equivalence listings and the column with surface similarity listings. Other generalization methods are possible; the reviewer should identify what approach is used in a specific paper or report.

With this context, we formulate questions for reviewers of statistical analysis in applied research as checklists. These are listed in Table 1. The next section is about such questions. It is followed by a case study and a discussion.

4 | Statistical checklists for reviewing applied research

Our goal is to setup a checklist for a reviewer considering aspects related to the statistical analysis of a research paper. These are structured in six parts:

1. Study design
2. Algorithmic and inferential methods in frequentism analysis
3. Bayesian methods in Bayesian analysis
4. Selective inference aspects
5. Severe testing properties
6. Presentation of findings

Specific questions addressing these sections are listed in Table 1.

These questions provide checklists to reviewers assigned the task of assessing the statistical analysis of an applied research paper. They are

Research Article
doi.org/10.1002/ansa.202000159
Chemistry Europe
European Chemical Societies Publishing

TABLE 1 Questions for reviewing statistical analysis in applied research

Part	Questions
1. Study design	1.1 Is the experimental set up clearly presented? 1.2 Have aliasing and power consideration been taken into account? 1.3 Is there reference to blocking, split plots and randomization? 1.4 Was an IRB required, and if so, was it obtained? (if relevant) 1.5 Are there any data ethics issues to consider?
2. Algorithmic and inferential methods	2.1 Are the algorithmic and inferential methods uses clearly stated? 2.2 Is the analysis aiming at estimation, predictive or explanatory goals? 2.3 Are data and code available to replicate the analysis? 2.4 Are outcomes of inferential analysis properly interpreted?
3. Bayesian analysis	3.1 Are prior distributions justified using prior experience or data? 3.2 What are the Bayesian methods used in the analysis? 3.3 How are Bayes factors interpreted?
4. Selective inference	4.1 Has the study been pre-registered? 4.2 Have any false discovery rate corrections been made? 4.3 Is the presentation of findings affected by selective inference?
5. Severe testing	5.1 Have the findings been tested with an option of failing the test? 5.2 Is the study a first or is it replicating previous studies? 5.3 Have probabilism, performance and probativeness criteria been considered? 5.4 What type of model is used in the analysis: primary models, experimental models or and data models? 5.5 If used, how are confidence interval (CI) interpreted?
6. Presentation of findings	6.1 How are the research findings presented? 6.2 Have the research findings been generalized? 6.3 Are there any causality arguments presented? 6.4 In a causal study, are there issues of endogeneity (reverse-causation)?

not meant to be prescriptive and are only designed as a sort of review checklist.

In this paper, we focus on evaluating studies presenting results in the development of analytic methods.[31] As background to such applications, we propose the checklist in Table 2. A reviewer should consider the checklist questions to help characterise the study under consideration.

5 | A CASE STUDY

The case study concerns the development of an HPLC method analyzed by Romero et al.[32] The specific system consists of an Agilent 1050, with a variable-wavelength ultra violet (UV) detector and a model 3396-A integrator. Table 3 lists the factors and their levels used in the designed experiments of this case study. The original experimental array was a 2^{7-4} fractional factorial experiment with three center points (see Table 4). The levels '−1' and '1' correspond to the lower and upper levels listed in Table 3, and '0' corresponds to the nominal level. The lower and upper levels are chosen to reflect variation that might naturally occur about the nominal setting during regular operation. The fractional factorial experiment consists of 11 runs that combine the design factor levels in a balanced set of combinations, including three center points.

What do we learn from this fractional factorial experiment?

The following paragraphs illustrate the use of the checklist in Table 1 on this robustness study.

1. The study design is clearly described (fractional factorial design). The aliases in such designs are well-known from the statistical literature. The authors propose a main effects only model (due to the narrow ranges of the investigated parameters):

$$\text{Predicted Peak Height} = \hat{Y} = \hat{\alpha} + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_7 X_7$$

Where $\hat{\alpha}$ is the estimated intercept, and the $\hat{\beta}_i$'s are the estimated coefficients (slopes) of each independent variable (X_1 being the gradient, …, X_7 the dimethyl-formamide Percentage, see Table 4). The intercept, $\hat{\alpha}$, is the peak height predicted at the nominal levels of each input variables as the design is coded between −1 and +1. The goal of the robustness study is then to study the impact of changes in the input variables to the nominal level (target) of peak height.
2. A power analysis is not given. The experiments were run in a randomized order (no blocking or split-plot design is used). The use of an institutional review board (IRB) is not required in this non-clinical study, and there are no data ethics issues.
3. The methodology and the inferential method are well described. The multiple linear regression fit is summarized here in Table 5 (with the estimates, 95% CIs and *p*-values given by the use of *t*-test).
4. As explained by Kenett,[33] these experiments help to answer the following important questions:

   - How sensitive is the method to natural variation in the input settings?

   - Which inputs have the largest effect on the outputs from the method?

Research Article
doi.org/10.1002/ansa.202000159
Chemistry Europe
European Chemical Societies Publishing

TABLE 2 Checklist for analytic methods

Analytic method element	Description and question (Q)
Precision	This requirement makes sure that method variability is only a small proportion of the specifications range (upper specification limit – lower specification limit). This is also called gage reproducibility and repeatability (GR&R). *Q: Does the study address precision? How?*
Selectivity	Determination of impurities to monitor at each production step and specification of design methods that adequately discriminate the relative proportions of each impurity. *Q: Does the study address selectivity? How?*
Sensitivity	The achievement with the method of effective process control, by accurately reflecting changes in CQA's that are important relative to the specification limits. *Q: Does the study address sensitivity? How?*
Method Design Intent	Identification and specification of the analytical method performance *Q: Is the method design intent stated?*
Method Design Selection	Approach to the selection of the method work conditions to achieve the design intent *Q: Is the study design described?*
Method Control	Establishment and definition of appropriate controls for the components with the largest contributions to performance variability. *Q: Is the application of the method discussed?*
Method Control Validation	Demonstration of acceptable method performance with robust and effective controls. *Q: Is the method validation demonstrated?*
Method robustness	Testing robustness of analytical methods involves evaluating the influence of small changes in the operating conditions. *Q: Is the method robustness evaluated?*
Method ruggedness	Ruggedness testing identifies the degree of reproducibility of test results obtained by the analysis of the same sample under various normal test conditions such as different laboratories, analysts, and instruments *Q: Is the method ruggedness evaluated?*

TABLE 3 Factors and levels in high performance liquid chromatography (HPLC) experiments

Factor	Nominal value	Lower level (−1)	Upper level (+1)
Gradient profile	1	0	2
Column temp (^{0}C)	40	38	42
Buffer conc (mM)	40	36	44
Mobile-phase buffer pH	5	4.8	5.2
Detection wavelength (nm)	446	441	451
Triethylamine (%)	0.23	0.21	0.25
Dimethylformamide (%)	10	9.5	10.5

- Are there different inputs that dominate the sensitivity of different responses?
- Is the variation transmitted from factor variation large relative to natural run-to-run variation?

The input variable with the lowest *p*-value is the column temperature while the highest *p*-value is given for the buffer pH. However, none of the parameters is significant. The authors do not address the possibility of improving robustness by possibly moving the nominal setting to one that is less sensitive to factor variation.

The data are available, and the analysis can be reproduced (even though no computer code is given in the original paper). The outcomes of the analysis are interpreted and visualized by means of graphs.

5. No Bayesian analysis is reported in this study.
6. Non-clinical studies do not need to be pre-registered. No false discovery rate corrections were made, which means that the five response variables in the original paper must be interpreted separately (only the peak height is considered here). The joint analysis of the different response variables could be further discussed. The presentation of findings was comprehensive without undue emphasis on specific findings. However, there is no evidence that the selected model is the right one. The authors chose a main effects only model but robustness has a close link to nonlinearity. Kenett[33] has shown that this (simplified) model suffers from a lack of fit when analyzing the height of the peak and encourages the use of quadratic and/or interaction terms in robustness study. The *p*-value

Research Article
doi.org/10.1002/ansa.202000159
Chemistry Europe
European Chemical Societies Publishing

TABLE 4 Original fractional factorial experimental array for high performance liquid chromatography (HPLC) experiment (seven independent variables and one response variable (peak height)

Gradient (X1)	Column temperature (X2)	Buffer Concentration (X3)	Buffer pH (X4)	Detection Wavelength (X5)	Triethylamine percentage (X6)	Dimethyl-formamide Percentage (X7)	Peak Height (Y)
1	1	1	1	1	1	1	221.351
1	1	−1	−1	1	−1	−1	226.029
1	−1	1	1	−1	−1	−1	226.136
1	−1	−1	−1	−1	1	1	225.052
−1	1	1	−1	−1	1	−1	221.835
−1	1	−1	1	−1	−1	1	224.268
−1	−1	1	−1	1	−1	1	234.957
−1	−1	−1	1	1	1	−1	234.699
0	0	0	0	0	0	0	221.249
0	0	0	0	0	0	0	218.445
0	0	0	0	0	0	0	219.921

TABLE 5 Estimated coefficients on the original fractional factorial design for high performance liquid chromatography (HPLC) experiment

	Estimate (95% confidence interval [CI]), *p*-value
Intercept	224.9 (219.14, 230.67), $p < 00001$
Gradient	−2.15 (−8.91, 4.61), $p = 0.39$
Col Temp	−3.42 (−10.18, 3.34), $p = 0.21$
Buf Conc	−0.72 (−7.48, 6.04), $p = 0.76$
Buf pH	−0.18 (−6.94, 6.59), $p = 0.94$
Det Wave	2.47 (−4.3, 9.23), $p = 0.33$
Trie perc	−1.06 (−7.82, 5.71), $p = 0.65$
Dim Perc	−0.38 (−7.15, 6.38), $p = 0.87$

of the lack of fit test is indeed $p = 0.018$, which indicates that the form of the model is not adequate. One can notice that the three replicates at the center point of the design are much lower than their predictions. The model predicts the peak height at 224.90 when all the parameters are set to their nominal level, while the observed mean is 219.87

Kenett[33] shows that a definite screening design is appropriate to evaluate the robustness of a chemical process by estimating linear and quadratic terms. Table 6 shows the 17 runs of such a design with the corresponding peak height.

The quadratic effect of 'gradient' is then significant (Table 7). The main effects and the quadratic effect of 'gradient' are statistically significant, the adjusted R^2 is 82%, and the run-to-run variation has an estimated standard deviation of 2.498. This results in a curvature of the response variable around the nominal level of the 'gradient' factor, which is important when interpreting the results (while this was neglected in the original analysis). Thus, this quadratic term gives valuable information about where to set the gradient to achieve a robust method (typically at the minimum of this curvature where potential variation of the gradient will have minimum impact). In order to improve robustness, we need to identify nonlinear effects. Here, the only nonlinear effect is for gradient. The effect of each input variables in the peak height is illustrated on Figure 3. This shows us that the quadratic response curve for gradient reaches a minimum quite close to the nominal value (0 in the coded units of Figure 2). Consequently, setting the nominal level of Gradient to that level is a good choice for robustness. The other factors can also be kept at their nominal settings. They have only minor quadratic effects, so moving them to other settings will have no effect on method robustness. The level of variation on the response variable can then be assessed by simulating a noise from normal distributions around the nominal levels (using the simulator in JMP statistical disovery (JMP) with a normal distribution standard deviation (SD) = 0.4 in coded units for each input). Figure 3 shows the results of this simulation. The standard deviation of peakHeight associated with variation in the factor levels is 2.832, very similar in magnitude to the SD for run-to-run variation from the experimental data. The estimate of the overall method SD is then 3.776 (the square root of $2.498^2 + 2.832^2$). Figure 4 shows the histogram and density of the peak height obtained by simulations with noise on each of the seven input variables plus the run-to-run variability. By calculating quantiles 2.5 and 97.5, one can be 95% confident that the peak height will lie between 212.15 and 227.12. Dividing these values by the intercept (the target peak height estimated by the model), one can claim with 95% confidence that the peak height should not deviate more than 4.5% from its target value.

7. No option of failing the severe test approach is made. The study does not aim to replicate any previous studies. Probabilism is assessed by means of *p*-values for the significance of each parameter (no *p*-values are given in the original paper but significant

Research Article
doi.org/10.1002/ansa.202000159
Chemistry Europe
European Chemical Societies Publishing

TABLE 6 Definitive screening design for high performance liquid chromatography (HPLC) experiment (17 runs, seven independent variables and one response variable [peak height])

Gradient	Column temperature	Buffer concentration	Buffer pH	Detection wavelength	Triethylamine percentage	Dimethyl-formamide Percentage	Peak height
−1	−1	−1	1	−1	1	1	232.873
−1	−1	1	−1	1	1	0	228.823
−1	−1	1	1	0	−1	−1	231.756
−1	0	−1	−1	1	−1	1	234.056
−1	1	−1	1	1	0	−1	226.949
−1	1	0	−1	−1	1	−1	221.77
−1	1	1	0	−1	−1	1	223.008
0	−1	−1	−1	−1	−1	−1	220.459
0	0	0	0	0	0	0	214.52
0	1	1	1	1	1	1	216.927
1	−1	−1	0	1	1	−1	225.315
1	-1	0	1	1	−1	1	234.211
1	−1	1	−1	−1	0	1	226.512
1	0	1	1	−1	1	−1	221.193
1	1	−1	−1	0	1	1	220.424
1	1	−1	1	−1	−1	0	222.251
1	1	1	−1	1	−1	−1	226.226

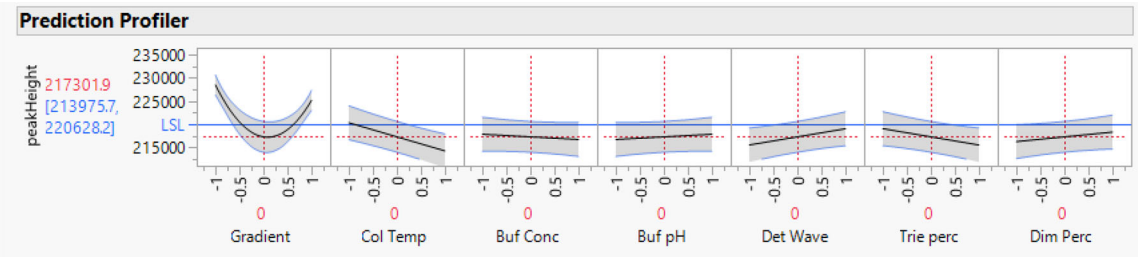

FIGURE 2 Profiler of peak Height at nominal levels (grey areas and blue curves are the 95% confidence intervals)

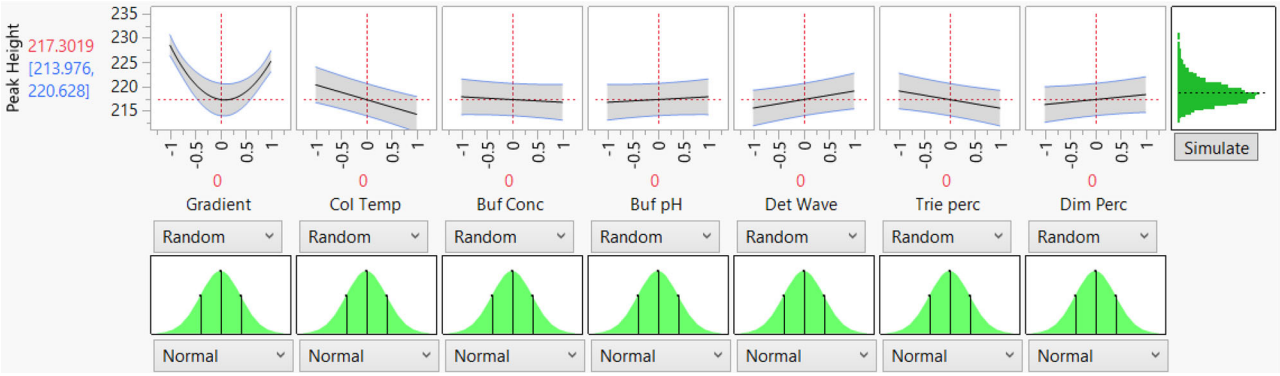

FIGURE 3 Profiler of peak Height at nominal levels (grey areas and blue curves are the 95% confidence intervals), with added noise from a normal distribution (mean equal to the nominal level) on the input variables and the impact on the peak height (histogram on the right) (JMP ver. 15.2)

Research Article
doi.org/10.1002/ansa.202000159
Chemistry Europe
European Chemical Societies Publishing

TABLE 7 Parameter estimates for high performance liquid chromatography (HPLC) experiment (with quadratic effect(s) from the definitive screening design)

	Estimate (95% confidence interval [CI]), *p*-value
Intercept	217.3 (213.98, 220.63), $p < .0001$
Gradient	−1.65 (−3.19, −0.11), $p = 0.04$
Col temp	−3.03 (−4.57, −1.49), $p = 0.002$
Buf Conc	−0.56 (−2.1, 0.98), $p = 0.42$
Buf pH	0.56 (−0.98, 2.1), $p = 0.42$
Det wave	1.75 (0.21, 3.29), $p = 0.03$
Trie perc	−1.76 (−3.3, −0.22), $p = 0.03$
Dim Perc	1.02 (-0.52, 2.56), $p = 0.16$
Gradient*Gradient	9.51 (5.84, 13.18), $p = 0.0003$

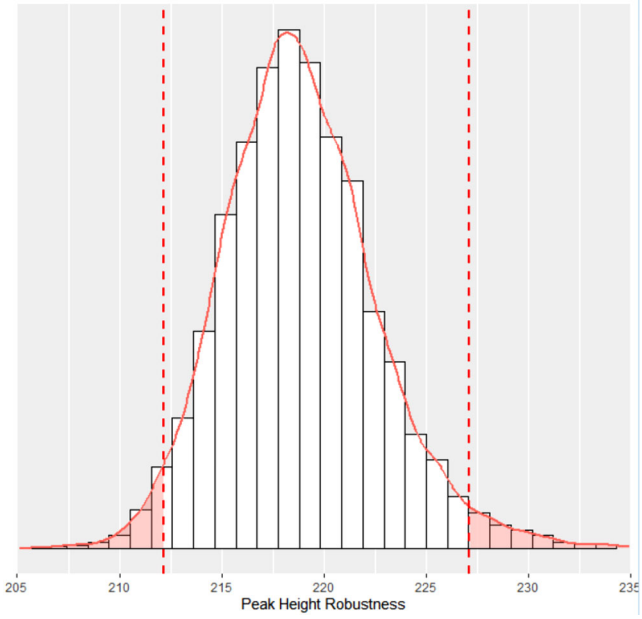

FIGURE 4 Histogram and density of peak Height at nominal levels for the seven input variables with added noise and run-to-run variability. Dashed vertical lines are the quantiles 2.5% and 97.5% (212.15 and 227.12)

parameters are highlighted). CIs are not given in the original paper but are provided here in Tables 5 and 7.

8. The research findings are well described and presented in summary tables and visualized by means of (3D) graphs. The paper concludes with a recommendation to set the ranges of the different parameters for the HPLC results to be robust. The causality issue is less important in this study as the original fractional factorial design is orthogonal for the main effects.

The checklist in Table 1 aims to improve the quality of the review as it clearly highlights that the study design, the goals, the statistical methodology and the data are clearly described in this HPLC robustness study. It also shows some points to improve (i.e., few words about the power of the study are missing, the authors could elaborate on the multiplicity issues when analyzing several response variables, the model adequacy is not discussed). CIs could help to better understand the impact and the importance of each parameter effect. In addition, the checklist in Table 2, specific for analytic methods, gives an overall summary of different important elements to consider when developing and analyzing analytical methods. The case study focuses on the HPLC's method robustness. The precision can be estimated with the three replicates at the nominal level of each of seven parameters (usually called in pharmaceutical industry critical process parameters). Different response variables (usually called critical quality attribute (CQA), critical quality attributes) are measured (this section focuses on the peak height).

6 | DISCUSSION

To evaluate the checklist in Table 1, we conducted an experiment by asking several researchers to review a paper by Smith et al.[34] before and after seeing the checklist table. They were then asked to comment on the checklist and more precisely address the question: 'do the guidelines provided by the checklist improve the quality of the review'? Their comments are hereby summarized.

'While many manuscripts are sent for review by editors or peers in industry, there is a lack of consistency in reviewing the innovations, due to the early development stage of the research and the lack of commonly shared views. How to evaluate papers regarding their innovation in interdisciplinary fields is usually not very clear'.

'I felt a bit dumb without the checklist as no clear guidelines were given in the first round of review. There are some weak or missing points in the paper that would not be highlighted or even not spotted without the checklist'.

'This checklist is very useful to be sure that some important points are adequately addressed in the paper. It might be good to send the statistical review to help the subject matter expert reviewer as well'.

In Francois,[35] the author analyzes data from an experiment design to assess the effect of variability in the review process. The paper described an experiment where 10% of submitted manuscripts (166 items) submitted for publication in a conference proceeding went through the review process twice. Arbitrariness was measured as the conditional probability for an accepted submission to get rejected if examined by the second committee. This number was equal to 60%, for a total acceptance rate equal to 22.5%. The author applies a Bayesian analysis to these two numbers, by introducing a hidden parameter, which measures the probability that a submission meets basic quality criteria. The standard quality criteria considered in this study include novelty, clarity, reproducibility, correctness and no form of misconduct. These were met by a large proportion of submitted items. The Bayesian estimate for the hidden parameter was equal to 56% (95% CI: [0.34, 0.83]). As a result of this analysis, the author suggests that the total acceptance rate should be increased in order to decrease arbitrariness estimates in future review processes.

Yet another approach for reviewing applied research is based on the information quality framework introduced in Kenett and Shmueli.[36,43] This framework involves four components (study utility (U), the data (X), the data analysis (f) and the analysis goal (g)) and eight dimensions (data resolution, data structure, data integration, temporal relevance, generalizability, chronology of data and goal, operationalization and communication). Information quality is defined as the utility of a particular data set for achieving a given analysis goal by employing statistical analysis or machine learning algorithms.[36,37]

Data analysis pipelines affect the outcomes of statistical analysis, Botvinik-Nezer et al.[38] These are usually not documented. Part of this is the handling of missing data and outliers. For an exception see openml.org Vanschoren et al.[39] where open access is given to the data and its analysis platform. Reviewers of data analysis uploaded to this platform should be able to fully replicate the study under review. Popp and Biskup[40] has proposed a framework in Pythin for the analysis of spectroscopic data focussing on reproducibility and good scientific practice. We therefore anticipate that future publications will require a documentation of the data analysis pipeline, beyond current requests to make data and code publicly available.

In conclusions, several areas in science have set checklists tailored to their needs, see for example, Feng et al.[41] and Aczel et al.[42] Our goal here is to provide such support in the context of statistical analysis in studies focused on the development of analytic methods.

ACKNOWLEDGEMENT

The authors are grateful to Dr Sylvie Scolas and Dominique Derreumaux (GSK), and Professors Ran Jin, and Xinwei Deng (Virginia Tech) for their helpful comments on the checklist table.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

REFERENCES

1. Altman DG, Gore SM, Gardner MJ, Pocock SJ. Statistical guidelines for contributors to medical journals. *Ann Clin Biochem*. 1992;29:1-8.
2. Lang T, Altman D. Basic statistical reporting for articles published in clinical medical journals: the SAMPL Guidelines. In: Smart P, Maisonneuve H, Polderman A, eds. *Science Editors' Handbook*. European Association of Science Editors; 2013.
3. Wasserstein R, Lazar N. The ASA's statement on p-values: context, process, and purpose. *Am Stat*. 2016;70:129-133.
4. Gelman A, Loken E. The statistical crisis in science. *Am Sci*. 2014;2:460-465.
5. Amrhein V, Greenland S. Remove, rather than redefine, statistical significance. *Nat Hum Behav*. 2018;2:4.
6. Mayo D. P-value thresholds: forfeit at your peril. *Eur J Clin Invest*. 2019;49:e13170.
7. Ioannidis J. The importance of predefined rules and prespecified statistical analyses: do not abandon significance. *JAMA*. 2019a;321:2067-2068.
8. Anjum RL, Copeland S, Rocca E. BMJ evidence-based medicine. 2020;25:6-8.
9. Mertens W, Jenkens R. New guidelines for null hypothesis significance testing in hypothetico-deductive IS research. *J Assoc Inf Syst*. 2019. https://osf.io/preprints/socarxiv/5qr7v/
10. Trafimow D, Michael M. Editorial. *Basic Appl Soc Psychol*. 2015;37(1):1-2.
11. Mayo D. *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*. Cambridge University Press; 2018.
12. Harris RS. *Rigor Mortis: How Sloppy Science Creates Worthless Cures, Crushes Hope, and Wastes Billions*. Basic Books; 2017.
13. Wager E, Williams P. Why and how do journals retract articles? An analysis of Medline retractions 1988-2008. *J Med Ethics*. 2011;37:567-570.
14. Efron B, Hastie T. *Computer Age Statistical Inference: Algorithms, Evidence and Data Science*. Cambridge University Press; 2016.
15. Cox DR. Theory and general principle in statistics. *JRSS(A)*. 1981;144:289-297.
16. Barnett AG, Wren JD. Examination of CIs in health and medical journals from 1976 to 2019: an observational study. *BMJ Open*. 2019;9:e032506. Accessed October 20, 2020. https://doi.org/10.1136/bmjopen-2019-032506
17. Cox DR. Some problems connected with statistical inference. *Ann Mat Stat*. 1958;29(2):357-372. Accessed October 20, 2020.
18. Gelman A. 2020. https://statmodeling.stat.columbia.edu/2020/03/26/the-value-or-lack-of-value-of-preregistration-in-the-absence-of-scientific-theory/
19. Gelman A. 2020. https://statmodeling.stat.columbia.edu/2020/03/27/lets-do-preregistered-replication-studies-of-the-cognitive-effects-of-air-pollution-not-because-we-think-existing-studies-are-bad-but-because-we-think-the-topic-is-important-and-we-want-to-unders/
20. Haig BD. What can psychology's statistics reformers learn from the error-statistical perspective?. *Methods Psychol*. 2020;2:100020. https://doi.org/10.1016/j.metip.2020.100020
21. Kenett RS, Zacks S. *Modern Industrial Statistics: with applications in R, MINITAB and JMP*. 3rd ed. John Wiley and Sons; 2021.
22. Box GEP, Hunter W, Hunter S. *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*. John Wiley and Sons; 1978.
23. Ioannidis J. What have we (not) learnt from millions of scientific papers with p-values?. *Am Stat*. 2019b;73(sup1):20-25. Accessed October 20, 2020.
24. Benjamini Y. Selective inference: the silent killer of replicability, Rietz Lecture, joint statistical meetings. Paper presented at: Colorado Convention Center; July 27–August 1, 2019; Denver, Colorado.
25. Pearl J. Generalizing experimental findings. *J Causal Inference*. 2015;3(2):259-66.
26. Kenett RS, Rubinstein A. Generalizing research findings for enhanced reproducibility: a translational medicine case study. 2017. https://ssrn.com/abstract=3035070
27. Greenland S. Invited commentary: the need for cognitive science in methodology. *Am J Epidemiol*. 2017;186(6):639-645.
28. Yarkoni T. The generalizability crisis, PsyArXiv. 2019. https://psyarxiv.com/jqw35
29. Kidon M, Blanca-Lopez MN, Gomes E, et al. Diagnosis and management of hypersensitivity reactions to non-steroidal anti-inflammatory drugs (NSAIDs) in children and adolescents. *Pediatric Allergy Immunol*. 2018;29(5):469-480. https://onlinelibrary.wiley.com/doi/full/10.1111/pai.12915
30. Efron A, Zeldin Y, Gotesdyner L et al. A structured gradual exposure protocol to baked and heated milk in the treatment of milk allergy. *J Pediatr*. 2018;203:204-209.e2. Accessed October 20, 2020. https://www.jpeds.com/article/S0022-3476(18)31090-4/abstract

Research Article
doi.org/10.1002/ansa.202000159
Chemistry Europe
European Chemical Societies Publishing

31. Kenett RS, Kenett DA. *Quality by Design Applications in Biosimilar Technological Products, Accreditation and Quality Assurance*. Springer Verlag; 2008.
32. Romero R, Gázquez D, Cuadros-Rodríguez L, et al. A geometric approach to robustness testing in analytical HPLC. *LCGC North Am*. 2002;20(1):72-80.
33. Kenett R. The QbD Column: applying QbD to make analytic methods robust. The QbD Column, JMP. 2017. https://community.jmp.com/t5/JMP-Blog/The-QbD-Column-Applying-QbD-to-make-analytic-methods-robust/ba-p/33206
34. Smith M, Francq B, McConnachie A, Wetherall K, Pelosi A, Morrison J. Clinical judgement, case complexity and symptom scores as predictors of outcome in depression: an exploratory analysis, *BMC Psychiatry*. 2020;20:125.
35. Francois O. Arbitrariness of peer review: a Bayesian analysis of the NIPS experiment. 2015. Accessed October 20, 2020. http://arxiv.org/abs/1507.06411
36. Kenett RS, Shmueli G. On information quality. *J Royal Statistical Society, Series A*. 2014;177(1):3-38.
37. Kenett RS, Shmueli G. Helping authors and reviewers ask the right questions: the InfoQ framework for reviewing applied research. *J Int Assoc Official Stat*. 2016;32:11-35.
38. Botvinik-Nezer R, Holzmeister F, Camerer CF, et al. Variability in the analysis of a single neuroimaging dataset by many teams, bioRxiv. 2019. https://www.biorxiv.org/content/10.1101/843193v1
39. Vanschoren J, van Rijn JN, Bischl B, Torgo L. OpenML: networked science in machine learning. *SIGKDD Explor*. 2013;15(2):49-60.
40. Popp J, Biskup T. ASpecD: a modular framework for the analysis of spectroscopic data focussing on reproducibility and good scientific practice. *Chem Methods*. 2022;2:e202100097.
41. Feng X, Park DS, Walker C, et al. A checklist for maximizing reproducibility of ecological niche models. *Nat Ecol Evol*. 2019;3:1382-1395. https://doi.org/10.1038/s41559-019-0972-5
42. Aczel B, Szaszi B, Sarafoglou A et al. A consensus-based transparency checklist. *Nat Hum Behav*. 2020;4:4-6. Accessed October 20, 2020. https://doi.org/10.1038/s41562-019-0772-6
43. Kenett RS, Shmueli G. *Information Quality: The Potential of Data and Analytics to Generate Knowledge*. John Wiley and Sons; 2016.

How to cite this article: Kenett RS, Francq BG. Helping reviewers assess statistical analysis: A case study from analytic methods. *Anal Sci Adv*. 2022;3:212-222. https://doi.org/10.1002/ansa.202000159