Should we test the model assumptions before running a model-based test?

Iqbal Shamsudheen[1] and Christian Hennig[2]

[1,2] *Department of Statistical Science, University College London, Gower Street, London, WC1E 6BT, United Kingdom,*
E-mail: iqbal.shamsudheen.14@ucl.ac.uk,
[2] *Dipartimento di Scienze Statistiche, Universita di Bologna, Via delle belle Arti, 41, 40126 Bologna, Italy,*
E-mail: christian.hennig@unibo.it

Abstract
Statistical methods are based on model assumptions, and it is statistical folklore that a method's model assumptions should be checked before applying it. This can be formally done by running one or more misspecification tests testing model assumptions before running a method that makes these assumptions; here we focus on model-based tests. A combined test procedure can be defined by specifying a protocol in which first model assumptions are tested and then, conditionally on the outcome, a test is run that requires or does not require the tested assumptions. Although such an approach is often taken in practice, much of the literature that investigated this is surprisingly critical of it, owing partly to the observation that conditionally on passing a misspecification test, the model assumptions are automatically violated ("misspecification paradox"). Our aim is to investigate conditions under which model checking is advisable or not advisable. For this, we review results regarding such "combined procedures" in the literature, we review and discuss controversial views on the role of model checking in statistics, and we present a general setup in which we can show that preliminary model checking is advantageous, which implies conditions for making model checking worthwhile.

Key words: Misspecification testing; Hypothesis test; Goodness of fit; Combined procedure; Misspecification paradox.

1 Introduction

Statistical methods are based on model assumptions, and it is statistical folklore that a method's model assumptions should be checked before applying it. Some authors believe that the invalidity of model assumptions and the failure to check them is at least partly to blame for what is currently discussed as "replication crisis" (Mayo (2018)), and indeed model checking is ignored in much applied work (Keselman et al. (1998), Strasak et al. (2007a,b), Wu et al. (2011), Sridharan & Gowri (2015), Nour-Eldein (2016)). Yet there is surprisingly little agreement in the literature about how to check the models. As will be seen later, several authors who investigated the statistical characteristics of running model checks before applying a model-based method comment rather critically on it. So is it sound advice to check model assumptions first? Our aim is to shed some light on the issue by collecting and commenting on relevant results and thoughts from the literature. We also present a new result that shows some conditions under which model checking is beneficial.

The amount of literature on certain specific problems that belong to this scope is quite large and we do not attempt to review it exhaustively. We restrict our focus to the problem of two-stage testing, i.e., hypothesis testing conditionally on the result of preliminary tests of model assumptions. More work exists on estimation after preliminary testing. For overviews see Bancroft & Han (1977), Giles & Giles (1993), Chatfield (1995), Saleh (2006). Almost all existing work focuses on analysing specific preliminary tests and specific conditional inference; here a more general view is provided.

To fix terminology, we assume a situation in which a researcher is interested in using a "main test" for testing a main hypothesis that is of substantial interest. There is a "model-based constrained (MC) test" involving certain model assumptions available for this. We will call "misspecification (MS) test" a test with the null hypothesis that a certain model assumption holds. We assume that this is not of primary interest, but rather only done in order to assess the validity of the model-based test, which is only carried out in case that the MS test does not reject (or "passes") the model assumption. In case that the MS test rejects the model assumption, there may or may not be an "alternative unconstrained (AU) test" that the researcher applies, which does not rely on the rejected model assumption, in order to test the main hypothesis. A "combined procedure" consists of the complete decision rule involving MS test, MC test, and AU test (if specified).

As an example consider a situation in which a psychiatrist wants to find out whether a new therapy is better than a placebo based on continuous measurements of improvement on two groups of patients, one group treated with the new therapy, the other with the placebo. The researcher may want to apply a two-sample t-test, which assumes normality (MC test). Normality can be tested by a Kolmogorov or Shapiro-Wilks test (MS test) in both groups, and in case normality is rejected, the researcher may decide to apply a Wilcoxon-Mann-Whitney (WMW) rank test (AU test) that does not rely on normality. Such a procedure is for example applied in Holman & Myers (2005), Kokosinska et al. (2018), and also at least implicitly endorsed in some textbooks, see, e.g., the flow chart Fig. 8.5 in Dowdy et al. (2004). There are some issues with this:

- The two-sample t-test has further assumptions apart from normality, namely that the data within each group are independently identically generated (i.i.d.), the groups are independent, and the variances are homogeneous. There are also assumptions regarding external validity, such as the sample being representative for the population of interest, and the measurements being valid. Neither is the WMW test assumption free, even though it does not assume normality. Using only a single MS test, not all of these assumptions are checked, and both the MC test and the AU test may be invalidated, e.g., by problems with the i.i.d. assumption. Using more than one MS test for checking model assumptions before running the MC test may be recommended. This could be formally defined within a more complex combined procedure, but for simplicity and in line with most of the existing literature we constrain ourselves mostly to situations in which only a single MS test is run, keeping in mind that there are further model assumptions that may require checking, see also Section 4.

- The two-sample t-test tests the null hypothesis $H_0: \mu_1 = \mu_2$ against $H_1: \mu_1 \neq \mu_2$ (or larger, or smaller), where μ_1 and μ_2 are the means of the two normal distributions within the two groups. H_0 and H_1 are defined within the normal model, and more generally, H_0 and H_1 of the MC test are defined within the assumed model. H_0 and H_1 tested by the AU test will not in general be equivalent, so there needs to be an explicit definition of the hypotheses tested by a procedure that depending on the result of the MS test will either run the MC or the AU test. In the example, in case that the variances are indeed homogeneous, the H_0 and H_1 tested by the t-test are a special case of H_0 and H_1 tested by the WMW test, namely that the two within-groups distributions are equal (H_0) or that one is stochastically larger or smaller than the other (H_1). See Fay & Proschan (2010) for a discussion of different "perspectives" of what the WMW- and t-test actually test. The combined procedure delivers a test of these more general H_0 and H_1, which sometimes may not be so easy to achieve. The key issue is how the scientific research question (whether the new therapy is equivalent to a placebo) translates into the specific model assumed by the MC test and the more general model assumed by the AU test.

The AU test may rely on fewer assumptions by being nonparametric as above, or by being based on a more general parametric model (such as involving an autoregressive component in case of violation of independence). It does not necessarily have to be based on more general assumptions than the MC test, it could also for example apply the original model with a transformed variable.

It is well known, going back to Bancroft (1944), that the performance characteristics of a combined procedure such as type 1 and type 2 error probabilities (size and one minus the power) in general differ from the characteristics of the MC test run unconditionally, even if the model assumptions of the MC test are fulfilled. This is a special case of data-dependent analysis, called "garden of forking paths" by Gelman & Loken (2014), who suggest that such analyses contribute to the fact that "reported statistically significant claims in scientific publications are routinely mistaken".

The issue of interest here is whether the performance characteristics of the combined procedure under various models (with model assumptions of the MC test fulfilled or violated) are good enough to recommend it, compared to running either the MC or the AU test unconditionally. If this is the case, model checking is advisable; if this is not the case, the main test to be run should be decided without checking the model by running the MS test. We will also comment on informal (visual) model checking.

We generally assume that the MS test is carried out on the same data as the main test. Some of the issues discussed here can be avoided by checking the model on independent data, however such data may not be available, or this approach may not be preferred for reasons of potential waste of information and lack of power. See Chatfield (1995) for a discussion of the case the "independent" data are obtained by splitting the available dataset. In any case it would leave open the question whether the data used for MS testing are really independent of the data used for the main test, and whether they do really follow the same model. If possible, this is however a valuable option.

The situation is confusing for the user in the sense that checking model assumptions is recommended in many places (e.g., Spanos (1999), Cox (2006), Kass et al. (2016)), but an exact formal specification how to do this in any given situation is hardly ever given. On the other hand, tests are routinely used in applied research to decide about model assumptions in all kinds of setups, often for deciding how to proceed further (e.g., Gambichler et al. (2002), Maydeu-Olivares et al. (2009), Hoekstra et al. (2012), Ravichandran (2012), Abdulhafedh (2017), Wu et al. (2019), Hasler et al. (2020)). Regarding the setup above, Fay & Proschan (2010), reviewing the literature, state that there are some true distributions under which the two-sample t-test is better than the WMW test, and some (non-normal) others for which the WMW test is better than the t-test, but they explicitly advise against normality testing or any data dependent method to decide between these, and prefer considerations based on the sample size and prior knowledge about the data ("if there is a small possibility of gross errors"). If in doubt, they prefer the WMW test, whereas Rochon et al. (2012), also advising against data dependent decisions, prefer the t-test, based on simulations that focused on different non-normal distributions than the heavy tailed ones on which Fay & Proschan (2010) base their recommendation. The problem is that there are very many possible non-normal distributions (and in general many possible violations of the model assumptions), for some of which the MC test is still better than the AU test, even though for some others the AU test is clearly preferable. Many users however will not know, before seeing the data, which of these distributions is more relevant in their situation. Surely there is a demand for a test or any formal rule to distinguish between situations in which the WMW test (or any other specific alternative to the t-test) is better, and situations in which the t-test is better, based on the observed data. But this problem is different from distinguishing normal from non-normal distributions, as which this is often framed, and which is what a normality test nominally addresses.

Given the difficulty to define a convincing formal approach, it is not surprising that informal approaches for model checking are often used. Many researchers do informal model checking (e.g., visual, such as looking at boxplots for diagnosing skewness and outliers, or using regression residual plots to diagnose heteroscedasticity or nonlinearity), and they may only decide how to proceed knowing the outcome of the model check (be it formal or informal), rather than using a combined procedure that was well defined in advance. In fact, searching the web for terms such as "checking model assumptions" finds far more recommendations to use graphical model assessment than formal MS tests. An obvious advantage of such an approach is that the researcher can see more specifically suspicious features of the data, often suggesting ways to deal with them such as transformations (by the way, running an MC test on transformed data conditionally on an MS test is also a combined procedure in our terminology). This may work well, however it depends on the researcher who may not necessarily be competent enough as a data analyst to do this better than a formal procedure, and it has the big disadvantage that it cannot be formally investigated, which would certainly be desirable. Obviously, if the way how the researcher makes a visual decision could be formalised, this could be analysed as another combined procedure.

In Section 2 we present our general perspective of model assumption checking. Section 3 formally introduces a combined procedure in which an MS test is used to decide between an MC and an AU main test. Section 4 reviews the controversial discussion of the role of model checking and testing in statistics. Section 5 runs through the literature that investigated the impact of misspecification testing and the performance of combined procedures in various scenarios. In Section 6 we present a new result that formalises a situation in which a combined procedure can be better than both the MC and the AU test. Section 7 provides the conclusion.

2 A general perspective on model assumption checking

Our view of model assumptions and assumption checking is based on the idea that models are thought constructs that necessarily deviate from reality but can be helpful devices to understand it (Hennig 2010, with elaboration for frequentist and Bayesian probability models in Section 5 of Gelman and Hennig 2017). Models for which we know the truth to be estimated or tested can be used to show that certain procedures are good or even optimal in a certain sense, such as the Neyman-Pearson Lemma on uniformly most powerful tests; the WMW-test is not normally justified by an optimality result but rather by results warranting the validity of the distribution of the test statistic, regardless of the specific form of the data distribution, and the unbiasedness against certain alternatives, see Fay & Proschan (2010). The term "model assumption" generally refers to the existence of such results, meaning that a method has a certain guaranteed quality if the model assumptions hold. But models are essentially different from reality, and therefore we do not think that it is ever appropriate to state that any model is "really true" or any model assumption "really holds". The best that can be said is that it may be appropriate and useful to treat reality as if a certain model were true, acknowledging that this is always an idealisation. A test generally checks whether observed data are compatible with a certain model in a certain respect, which is defined by the test statistic. All data is compatible with many models; there are always alternatives to any assumed model that cannot be ruled out by the data, such as non-identical distributions that allow a different parameter choice for each observation, or dependence structures that affect all data in the same way, so that they cannot be detected by looking at patterns in the data related to aspects such as time order, geographical distance, or different levels of a known but random factor. Starting from the classical work of Bahadur & Savage (1956), there are results on the impossibility to identify certain features of general families of distributions such as their means, or bounds on the density (Donoho (1988)). This means that it is ultimately impossible to make sure that model assumptions hold.

In order to increase our understanding of the performance of a statistical procedure, it is instructive to not only look at its results in situations in which the model assumptions are fulfilled, but also to explore it on models for which they are violated, but chosen so that if they were true, applying the procedure of interest still seems realistic. Such an approach is taken in the literature discussed in Section 5 as well as in much literature on robust statistics, the latter mostly interested in worst case considerations (e.g., Hampel et al. (1986)). The problem that a procedure is meant to solve is often defined in terms of the assumed model, so if other models are considered for data generation, an analogous problem has to be defined for those other models, which may not always be unique, as mentioned already in the introduction. A suitable way to think about this is that there is a scientific hypothesis and alternative of interest (such as "no difference between treatments" vs. "treatment A is better") that can be translated into various probability models, potentially in more than one way (e.g., "treatment A is better" may in a nonparametric setting translate into "treatment A's distribution is stochastically larger", or "treatment A's distribution is a positive shift of treatment B's distribution", or "the expected outcome value of treatment A is larger"). As already mentioned, in such situations it can sometimes be observed that the procedure's performance is still satisfactory, and in some other situations it may be bad, both in absolute terms or compared to available alternative procedures.

The implication is that the problem of checking the model assumptions is often wrongly framed as "checking whether the model assumptions hold", because in reality they will not hold precisely anyway, but a method may still perform well in that case, and the model assumption may not even be required to hold "approximately" (e.g., t-tests do very well on uniformly distributed samples). But there are certain violations of the model assumptions that have the potential to mislead the results in the sense of giving a wrong assessment of the underlying scientific hypothesis with high probability. "Checking the model assumptions" should rule such situations out as far as possible. This implies that model assumption checking needs to distinguish problematic violations from unproblematic ones, rather than distinguishing a true model from any wrong one. We think that some assumption checking does not work very well (see Section 5) because it tries to solve the latter problem but should actually solve the former. It is as misleading to claim that model assumptions are required to hold (which is an ultimately impossible demand) as it is to ignore them, or rather to ignore potential performance breakdown of the procedure to be applied on models other than the assumed one. In any case, knowledge of the context (such as sampling schemes and measurement procedures) should always be used to highlight potential issues on top of what can be diagnosed from the data.

Model assumption checking and choosing a subsequent method of inference conditionally on it, i.e., combined procedures, may help if done right, but may not help or even hurt if done wrong, and investigation of how well they work in all kinds of relevant situations is therefore of interest. Investigating them is however hard, because the performance depends on all kinds of details, including the choice of MS, MC, and AU test, and particularly the models under which the combined procedure is assessed. Unfortunately, assuming that data dependent decisions are not made before the combined procedure is applied, the user may have little information about what distribution to expect, so that a wide range of possibilities is conceivable, and different authors may well come to different conclusions regarding the same problem (see the example in the Introduction) based on different considered alternatives to the assumed model. This makes worst case considerations as in robust statistics attractive, but looking at a range of specific choices will give a more comprehensive picture. Here we will consider such investigations only as far as already covered in the literature (Section 5). Our own theoretical result regarding a more general setup in Section 6 will complement the overall rather critical assessment from the literature. The conditions stated in Section 7 may stimulate further research and development of model checking procedures that are better than existing ones at finding those issues with the model assumptions that matter.

3 Combined procedures

The general setup is as follows. Given is a statistical model defined by some model assumptions Θ,

$$M_\Theta = \{P_\theta, \theta \in \Theta\} \subset M,$$

where $P_\theta, \theta \in \Theta$ are distributions over a space of interest, indexed by a parameter θ. M_Θ is written here as a parametric model, but we are not restrictive about the nature of Θ. M_Θ may even be the set of all i.i.d. models for n observations, in which case Θ would be very large. However, in the literature, M_Θ is usually a standard parametric model with $\Theta \subseteq \mathbb{R}^m$ for some m. There is a model M containing distributions that do not require one or more assumptions made in M_Θ, but for data from the same space.

Given some data z, we want to test a parametric null hypothesis $\theta \in \Theta_0$, which has some suitably chosen "extension" $M^* \subset M$ so that $M^* \cap M_\Theta = M_{\Theta_0}$, against the alternative $\theta \notin \Theta_0$ corresponding to $M \setminus M^*$ in the bigger model. In some cases (for example when applying the original model to transformed variables) M may not contain M_Θ, and $M^* \subset M$ then needs to be some kind of "translation" of the research hypothesis M_{Θ_0} into M, the choice of which should be context guided and may or may not be trivial (e.g., equal group means for Gaussians will often correspond to the same research hypothesis as for logarithmised Gaussians).

In the simplest case, there are three tests involved, namely the MS test Φ_{MS}, the MC test Φ_{MC} and the AU test Φ_{AU}. Let α_{MS} be the level of Φ_{MS}, i.e., $Q(\Phi_{MS}(z) = 1) \leq \alpha_{MS}$ for all $Q \in M_\Theta$. Let α be the level of the two main tests, i.e., $P_\theta(\Phi_{MC}(z) = 1) \leq \alpha$ for all $P_\theta, \theta \in \Theta_0$ and $Q(\Phi_{AU}(z) = 1) \leq \alpha$ for all $Q \in M^*$. To keep things general, for now we do not assume that type 1 error probabilities are uniformly equal to α_{MS}, α, respectively, and neither do we assume tests to be unbiased (which may not be realistic considering a big nonparametric M).

The combined test is defined as

$$\Phi_C(z) = \begin{cases} \Phi_{MC}(z) & : \quad \Phi_{MS}(z) = 0, \\ \Phi_{AU}(z) & : \quad \Phi_{MS}(z) = 1. \end{cases}$$

This allows to analyse the characteristics of Φ_C, particularly its effective level (which is not guaranteed to be $\leq \alpha$) and power under P_θ with $\theta \in \Theta_0$ or not, or under distributions from M^* or $M \setminus M^*$. General results are often hard to obtain without making restrictive assumptions, although some exist, see Sections 5.1 and 5.4. At the very least, simulations are possible picking specific P_θ or $Q \in M$, and in many cases results may generalise to some extent because of invariance properties of model and test.

Also of potential interest are $P_\theta\left(\Phi_C(z) = 1 | \Phi_{MS}(z) = 0\right)$, i.e., the type 1 error probability under M_{Θ_0} or the power under M_Θ in case the model was in fact passed by the MS test, $Q\left(\Phi_C(z) = 1 | \Phi_{MS}(z) = 0\right)$ for $Q \in M \setminus M_\Theta$, i.e., the situation that the model M_Θ is in fact violated but was passed by the MS test, and whether Φ_C can compete with Φ_{AU} in case that $\Phi_{MS}(z) = 1$ (M_Θ rejected). These are investigated in some of the literature, see below.

4 Controversial views of model checking

The necessity of model checking has been stressed by many statisticians for a long time, and this is what students of statistics are often taught. Fisher (1922) stated:

> For empirical as the specification of the hypothetical population may be, this empiricism is cleared of its dangers if we can apply a rigorous and objective test of the adequacy with which the proposed population represents the whole of the available facts. Once a statistic, suitable for applying such a test, has been chosen, the exact form of its distribution in random samples must be investigated, in order that we may evaluate the probability that a worse fit should be obtained from a random sample of a population of the type considered.

Neyman (1952) outlined the construction of a mathematical model in which he emphasised testing the assumptions of the model by observation and if the assumptions are satisfied, then the model "*may be used for deductions concerning phenomena to be observed in the future*". Pearson (1900) introduced the goodness of fit chi-square test, which was used by Fisher to test model assumptions. The term "misspecification test" was only coined as late as Fisher (1961) for the selection of exogenous variables in economic models. Spanos (1999) used the term extensively. See Spanos (2018) for the history and exhaustive discussion of the use of MS tests.

At first sight, model checking seems essential for two reasons. Firstly, statistical methods that a practitioner may want to use are often justified by theoretical results that require model assumptions, and secondly it is easy to construct examples for the breakdown of methods in case that model assumptions are violated in critical ways (e.g., inference based on the arithmetic mean, optimal under the assumption of normality, applied to data generated from a Cauchy distribution will not improve in performance for any number of observations compared with only having a single observation, because the distribution of the mean of $n > 1$ observations is still the same Cauchy distribution).

Regarding the foundations of statistics, checking of the model assumptions plays a crucial role in Mayo (2018)'s philosophy of "severe testing", in which frequentist significance tests are portrayed as major tools for subjecting scientific hypotheses to tests that they could be expected to fail in case they were wrong; and evidence in favour of such hypotheses can only be claimed in case that they survive such severe probing. Mayo acknowledges that significance tests can be misleading in case that the model assumptions are violated, but this does not undermine her philosophy in her view, because the model assumptions themselves can be tested. A problem with this is that to our knowledge there are no results regarding the severity of MS tests, meaning that it is unclear to what extent a non-rejection of model assumptions implies that they are indeed not violated in ways that endanger the validity of the main test.

A problem with preliminary model checking is that the theory of the model-based methods usually relies on the implicit assumption that there is no data-dependent pre-selection or pre-processing. A check of the model assumptions is a form of pre-selection. This is largely ignored but occasionally mentioned in the literature. Bancroft (1944) was probably the first to show how this can bias a model-based method after model checking. Chatfield (1995) gives a more comprehensive discussion of the issue. Hennig (2010) coined the term "goodness-of-fit paradox" (from now on called "misspecification paradox" here) to emphasise that in case that model assumptions hold, checking them in fact actively invalidates them. Assume that the original distribution of the data fulfills a certain model assumption. Given a probability $\alpha > 0$ that the MS test rejects the model assumption if it holds, the conditional probability for rejection under passing the MS test is obviously $0 < \alpha$, and therefore the conditional distribution must be different from the one originally assumed. It is this conditional distribution that eventually feeds the model-based method that a user wants to apply.

How big a problem is the misspecification paradox, and more generally the fact that MS tests cannot technically ensure the validity of the model assumptions? Spanos (2010) argues that it is not a problem at all, because the MS test and the main test "*pose very different questions to data*". The MS test tests whether the data "*constitute a truly typical realisation of the stochastic mechanism described by the model*". He argues that therefore model checking and the model-based testing can be considered separately; model checking is about making sure that the model is "*valid for the data*" (Spanos (2018)), and if it is, it is appropriate to go on with the model-based analysis.

The point of view taken here, as in Chatfield (1995), Hennig (2010), and elsewhere in the literature reviewed below, is different: We should analyse the characteristics of what is actually done. In case the model-based (MC) test is only applied if the model is not rejected, the behaviour of the MC test should be analysed conditionally on data not being rejected by the MS test, and this differs from the behaviour under the nominal model assumption. We do not think that the misspecification paradox automatically implies that combined procedures are invalid; as argued in Section 2 we do not believe that the model assumptions are true in reality anyway, and a combined procedure is worthwhile if it has good performance characteristics regarding the underlying scientific hypothesis, which may have formalisations regarding both the assumed model and the usually more general model employed by the AU test.

If the distribution of the test statistic is independent of the outcome of the MS test, formally the misspecification paradox still holds, but it is statistically irrelevant. Conditioning on the result of the MS test will not affect the statistical characteristics of the MC test. An example for this is a MS test based on studentised residuals and a main test based on the minimal sufficient statistic of a Gaussian distribution (Spanos (2010)). More generally it can be expected that if what the MS test does is at most very weakly stochastically connected to the main test (i.e., if in Spanos's terms they indeed "pose very different questions to the data"), differences between the conditional and the unconditional behaviour of the MC test should be small. This can be investigated individually for every combination of MS test and main test, and there is no guarantee that the result will always be that the difference is negligible, but in many cases this will be the case.

Even in situations in which inference is only very weakly affected by preliminary model checking in case the assumed model holds indeed, the practice of model checking may still be criticised on the grounds that it may not help in case that the model assumption is violated, i.e., if data is generated by a model that deviates from the assumed one, the conditional distribution of the MC test statistic, given that the model assumption is not rejected, may not have characteristics that are any better than if applying the MC test to data with violated model-assumptions in all cases, see Easterling & Anderson (1978).

Some kinds of visual informal model checking can be thought of as useful in a relatively safe manner if they lead to model rejections only in case of strikingly obvious assumption violations that are known to have an impact (which can be more precisely assessed looking at the data in a more holistic manner than a formal test can). In this case the probability to reject a true model can be suspected to be very close to zero, in turn not incurring much "pretest bias". But this relies on the individual researcher and their competence to recognise a violation of the model assumptions that matters. Furthermore, some results in the literature presented in Section 5 suggest that it can be advantageous to reject the model behind the MC test rather more easily than an MS test with the usual levels of 0.01 or 0.05 would do.

A view opposite to Spanos's one, namely that model checking and inference given a parametric model should not be separated, but rather that the problems of finding an appropriate distributional "shape" and parameter values compatible with the data should be treated in a fully integrated fashion, can also be found in the literature (Easterling (1976), Draper (1995), Davies (2014)). Davies (2014) argues that there is no essential difference between fitting a distributional shape, an (in)dependence structure, and estimating a location (which is usually formalised as parameter of a parametric model, but could as well be defined as a nonparametric functional).

Bayesian statistics allows for an integrated treatment by putting prior probabilities on different candidate models, and averaging their contributions. Robust and nonparametric procedures may be seen as alternatives not only in case that model assumptions of model-based procedures are violated; they have also been recommended for unconditional use (Hampel et al. (1986), Hollander & Sethuraman (2001)), making prior model checking supposedly superfluous. All these approaches still make assumptions; the Bayesian approach assumes that prior distribution and likelihood are correctly specified, robust and nonparametric methods still assume data to be i.i.d., or make other structural assumptions violation of which may mislead the inference. So the checking of assumptions issue does not easily go away, unless it is claimed (as some subjectivist Bayesians do) that such assumptions are subjective assessments and cannot be checked against data; for a contrary point of view see Gelman & Shalizi (2013). To our knowledge, however, there is hardly any literature assessing the performance of model checking combined in which the "MC role" is taken by robust, nonparametric or Bayesian inference, but see Bickel (2015) for a combined procedure that involves model checking and robust Bayesian inference.

Some authors in the econometric literature (Discovery & in Econometrics (2014), Spanos (2018)) prefer "respecification" of parametric models to robust or nonparametric approaches in the case that model assumptions are rejected. In some situations the advantage of respecification is obvious, particularly where a specific parametric form of a model is required, for example for prediction and simulation. More generally, Spanos (2018) argues that the less restrictive assumptions of nonparametric or robust approaches such as moment conditions or smooth densities are often untestable, as opposed to the more specific assumptions of parametric models. But this seems unfair, because to the extent that violations from such assumptions cannot be detected for more general models, it cannot be detected that any parametric model holds either. Impossibility results such as in Bahadur & Savage (1956) or Donoho (1988) imply that distributions violating conditions such as bounded means, higher order moments, or existing densities are undistiguishably close to any parametric distribution. Ultimately Spanos is right that nonparametric and robust methods are not 100% safe either, but they will often work under a wider range of distributions than a parametric model; e.g., classical robust estimation does safeguard against mixture distributions of the type $(1-\varepsilon)\mathcal{N}+\varepsilon Q$, where $\mathcal{N}$ refers to a normal distribution, Q to any distribution, $0<\varepsilon$ small enough, which can have arbitrary or non-existing means and cannot be distinguished from a normal distribution with large probabilty for a given fixed sample size and ε small enough. Ultimately parametric respecification can be useful and can be successful in some cases such as sufficiently regular violations of independence where robust and nonparametric tools are lacking. Regarding the setup of interest here, the AU test can legitimately be derived from a parametric respecification of the model. When it comes to general applicability, in our view the cited authors seem too optimistic to regarding whether a respecified model that can be confirmed by MS testing of all assumptions (as required by Spanos) to be reasonably valid can always or often be found. Cited results in Section 5 suggest in particular that situations in which a violated model assumption is not detected by the MS test for testing that very assumption can harm the performance of the MC test in a combined procedure. Furthermore, a respecification procedure as implied by Spanos including testing all relevant assumptions is to our knowledge not yet fully formalised and will be hard to formalise given the complexity of the problem, so that currently its performance characteristics in various possible situations cannot be investigated systematically.

Another potential objection to model assumption checking is that, in the famous words of George Box, "all models are wrong but some are useful". It may be argued that model assumption checking is pointless, because we know anyway that model assumptions will be violated in reality in one way or another (e.g., it makes some sense to hold that in the real world no two events can ever be truly independent, and continuous distributions are obviously not "true" as models for data that are discrete because of the limited precision of all human measurement). This has been used as argument against any form of model-based frequentist inference, particularly by subjectivist Bayesians (e.g., de Finetti (1974)'s famous "probability does not exist"). Mayo (2018) however argues that "all models are wrong" on its own is a triviality that does not preclude a successful use of models, and that it is still important and meaningful to test whether models are adequately capturing the aspect of reality of interest in the inquiry. According to Section 2, it is at least worthwhile to check whether the data are incompatible with the model in ways that will mislead the desired model-based inference, which can happen in a Bayesian setting just as well. This does not require models to be "true".

5 Results for some specific test problems

In this Section we will review and bring together results from the literature investigating the performance characteristics of combined procedures. Our focus is not on the detailed recommendations, but on general conditions under which combined procedures have been compared to unconditional use of MC or AU test, and have been found superior or inferior.

5.1 The problem of whether to pool variances, and related work

Historically the first problem for which preliminary MS testing and combined procedures were investigated was whether to test the equal variances assumption for comparing the means of two samples. Until now this is the problem for which most work investigating combined procedures exists. Let $X_1, X_2, ..., X_n$ be distributed i.i.d. according to P_{μ_1,σ_1^2} and $Y_1, Y_2, ..., Y_n$ be distributed i.i.d. according to P_{μ_2,σ_2^2}, where P_{μ,σ^2} denotes the normal distribution with mean μ and variance σ^2. If $\sigma_1^2 = \sigma_2^2$, the standard two-sample t-test using a pooled variance estimator from both samples (MC test) is optimal.

For $\sigma_1^2 \neq \sigma_2^2$ Welch's approximate t-test with adjusted degrees of freedom depending on the two individual variances (AU test) is often recommended, see Welch (1938), Satterthwaite (1946), Welch (1947).

The normal distribution assumption will be discussed below, but normality has often been seen as not problematic due to the Central Limit Theorem, and the historical starting point is the equal variances assumption. Early authors beginning from Bancroft (1944) did not frame the problem in terms of "making sure that model assumptions are fulfilled", but rather asked, in a pragmatic manner, under what circumstances pooling variances is advantageous. If the two variances are in fact equal or very similar, it is better to use all observations for estimating a single variance hopefully precisely, whereas if the two variances are very different, the use of a pooled variance will give a biased assessment of the variation of the means and their difference.

It has been demonstrated that the two sample t-test is very robust against violations of equality of variances when sample sizes are equal as shown by Hsu (1938), Scheffé (1970), Posten et al. (1982), Zimmerman (2006). When both variances and sample sizes are unequal, the probability of the Type-I error exceeds the nominal significance level if the larger variance is associated with the smaller sample size and vice versa (Zimmerman (2006), Wiedermann & Alexandrowicz (2007), Moder (2010)), which is amended by Welch's t-test. Bancroft & Han (1977) published a bibliography of the considerable amount of literature on that problem available already at that time. One reason for the popularity of the variance pooling problem in early work is that, as long as normality is assumed, only the ratio of the variances needs to be varied to cover the case of violated model assumptions, which makes it easier to achieve theoretical results without computer-intensive simulations.

Work that investigated sizes and/or power of combined procedures involving an MS test for variance equality for a main test of the equality of means, theoretically or by simulation, comprises Gurland & McCullough (1962), Bancroft (1964), Gans (1981), Moser et al. (1989), Gupta & Srivastava (1993), Moser & Stevens (1992), Albers et al. (2000a), Zimmerman (2014). General findings are that the combined procedure can achieve a competitive performance regarding power and size beating Welch's t-test, which is usually recommended as the AU test, only in small subspaces of the parameter space with specific sample sizes, and none of these authors recommends it for default use; Moser & Stevens (1992) recommended to never test the equal variances assumption. Often the unconditional Welch's t-test is recommended, which is only ever beaten by a very small margin where the MC test or the combined procedure are better; occasionally recommendations of using either the MC test or the AU test unconditionally depend on sample sizes.

Markowski & Markowski (1990) hinted at what the problem with the combined procedure is. They evaluated the F-test as MS test of homogeneity of variances for detecting deviations from variance equality that are known to matter for the standard t-test by simulations, and showed that the F-test is ineffective at finding these. Like Gans (1981), they also involved non-normal distributions in their comparisons, but this did not lead to substantially different recommendations.

Albers et al. (2000a) presented a second order asymptotic analysis of the combined procedure for pooling variances with the F-test as MS-test. They argue that this procedure can only achieve a better power than unconditional testing under the unconstrained model if the test size is also increased. This means that there are only two possibilities for the combined procedure to improve upon the MC test. Either the combined procedure is anti-conservative, i.e., violates the desired test level, which would be deemed unacceptable in most applications, or the size of the MC test is smaller than the nominal level, which if its assumptions are not fulfilled is sometimes the case. Albers et al. (2000b) extend these results to the analysis of a more general problem for distributions $P_{\theta,\tau}$ from a parametric family with two parameters θ and τ, where $\theta = 0$ is the main null hypothesis of interest and the decision between an MC test assuming $\tau = 0$ and an AU test without that assumption is made based on an MS test testing $\tau = 0$. In the two-sample variance pooling problem, τ could be the logarithm of the ratio between the variances; a simpler example would be the choice between Gauss- and t-test in the one-sample problem, where the MS test tests whether the variance is equal to a given fixed value. Once more, the combined procedure can only achieve better power at the price of a larger size, potentially being anti-conservative. Another key aspect is that the authors introduced a correlation parameter ρ formalising the dependence between the MS-test and the main tests. In line with the discussion in Section 4, they state that for strong dependence preliminary testing is not sensible, and their results consider the case $\rho \to 0$.

Arnold (1970) considered a different problem, namely whether to pool observations of two groups if the mean of the first group is the main target for testing. Pooling assumes that the two means are equal, so a test for equality of means here is the MS test. In line with the general experiences regarding MS testing for equality of variances, Arnold observed that in vast regions of the parameter space a better power can be achieved without pooling.

5.2 Tests of normality in the one-sample problem

The simplest problem in which preliminary misspecification testing has been investigated is the problem of testing a hypothesis about the location of a sample. The standard model-based procedure for this is the one-sample Student's t-test. It assumes the observations $X_1, X_2, ..., X_n$ to be i.i.d.

normal. For non-normal distributions with existing variance the t-test is asymptotically equivalent to the Gauss-test, which is asymptotically correct due to the Central Limit Theorem. The t-test is therefore often branded robust against non-normality if the sample is not too small, see, e.g., Bartlett (1935), Lehmann & Romano (2005). An issue is that the quality of the asymptotic approximation does not only depend on n, but also on the underlying distributional shape, as the speed of improvement of the normal approximation is not uniform. Very skew distributions or extreme outliers can affect the power of the t-test for large n, see Cressie (1980). Cressie mentions that the biggest problems occur for violations of independence, however we are not aware of any literature examining of independence testing combined with the t-test. Instead, a number of publications examine preliminary normality testing for the t-test.

Some work focuses just on the quality of the MS tests without specific reference to its effect on subsequent inference and combined procedures, see Razali & Wah (2011), Mendes & Pala (2003), Farrell & Rogers-Stewart (2006), Keskin (2006).

Schoder et al. (2006*a*) and Keselman et al. (2013) investigated normality tests regarding its use for subsequent inference without explicitly involving the later main test. Both advise against the Kolmogorov-Smirnov test. Keselman et al. (2013) concluded that the Anderson-Darling test is the most effective one at detecting non-normality relevant to subsequent t-testing, and they suggested that for deciding whether the MC test should be used, the MS test be carried out at a significance level larger than 0.05, for example 0.15 or 0.20, in order to increase the power, as all these tests may have difficulties to detect deviations that are problematic for the t-test.

Another group of work examines running a t-test conditionally on passing normality by a preliminary normality test. Most of these do not consider what happens if normality is rejected. Easterling & Anderson (1978) considered various distributions such as normal, uniform, exponential, two central and two non-central t-distributions. They generated 1000 samples each for which normality was passed and rejected, respectively, at 10% significance level, using both the Anderson-Darling and the Shapiro-Wilk normality tests. In the case that normality was passed, they compared the empirical distribution of the resulting t-values to Student's t-distribution. This worked reasonably well when the samples were drawn from the normal distribution. For symmetric non-normal distributions, the results were mixed, and for situations where the distributions were asymmetric, the distribution of the t-values did not resemble a Student's t-distribution, which they take as an argument against the practice of preliminary normality testing, because in case that the underlying distribution is not normal, normality testing does not help. As a result they favoured a nonparametric approach.

In a similar manner Schoder et al. (2006*b*) investigated the conditional type 1 error rate of the one sample t-test given that the sample has passed a test for normality for data from normal, uniform, exponential, and Cauchy populations. They conclude that the MS test makes matters worse in the sense that the Type I error rate is further away from the nominal 5% (lower for the uniform and Cauchy, higher for the exponential) for data that pass the normality test than when the t-test is used unconditionally (which works rather well for the uniform and exponential distribution, but not for the Cauchy), and this becomes worse for larger sample sizes. For the Cauchy distribution they also investigated running a Wilcoxon signed rank test as AU test conditionally on rejecting normality, which works worse than using the AU test unconditionally. Rochon & Kieser (2011) come to similar conclusions using a somewhat different collection of MS tests and underlying distributions. The problem with the results of the latter papers is that their setups to investigate he workings of a combined procedure implying that the underlying true distribution is fixed and given. This ignores the capability of a combined procedure to distinguish between underlying distributions for which the MC test works better or worse, like here the normal, uniform, and exponential distributions on one hand, and the Cauchy distribution on the other. Section 6 suggests a setup that can take this into account.

5.3 Tests of normality in the two-sample problem

For the two-sample problem, the Wilcoxon-Mann-Whitney (WMW) rank test is a popular alternative to the two-sample t-test with (in the context of preliminary normality testing) mostly assumed equal variances. In principle most arguments and results from the one-sample problem apply here as well, with the additional complication that normality is assumed for both samples, and can be tested either by testing both samples separately, or by pooling residuals from the mean. As for the one-sample problem, there are also claims and results that the two-sample t-test is rather robust to violations of the normality assumption (Hsu & Feldt (1969), Rasch & Guiard (2004)), but also some evidence that this is sometimes not the case, and that the WMW rank test can be superior and does not lose much power even if normality is fulfilled (Neave & Granger (1968)). Fay & Proschan (2010) presented a survey on comparing the two-sample t-test with the WMW test (involving further options such as Welch's t-test and a permutation t-test for exploring its distribution under H_0), concluding that the WMW test is superior where underlying distributions are heavy tailed or contain a certain amount of outliers; it is well known that the power of the t-test can break down under addition of a single outlier in the worst case, see He et al. (1990). Although Fay and Proschan did not explicitly investigate decision between t- and WMW-test by normality testing, they advise against it, stating that normality tests tend to have little power for detecting distributions that cause problems for the t-test.

Rochon et al. (2012) investigated by simulation combined procedures based on preliminary normality testing both for both samples separately, and pooled residuals using a Shapiro-Wilk test of normality. The MC test was the two sample t-test, the AU test was the WMW test. Data were simulated from normal, exponential, and uniform distributions. In fact, for these distributions, the MC test was always better than the AU test, which makes a combined procedure superfluous; it reached acceptable performance characteristics, but inferior to the MC test. A truly heavy tailed distribution to challenge the MC test was not involved.

Zimmerman (2011) achieved good simulation results with an alternative approach, namely running both the two-sample t-test and the WMW test, choosing the two-sample t-test in case the suitably standardised values of the test statistics are similar and the WMW test in case the p-values are very different. This seems to address the problem of detecting violations of normality better where it really matters. The tuning of this approach is somewhat less intuitive than for using a standard MS test.

5.4 Regression

In standard linear regression,

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + e_i,\ i = 1,\ldots,n,$$

with response $Y = (y_1,\ldots,y_n)$ and explanatory variables $X_j = (x_{j1},\ldots,x_{jn}),\ j = 1,\ldots,p$. $e_1,\ldots,e_n$ are in the simplest case assumed i.i.d. normally distributed with mean 0 and equal variances.

The regression model selection problem is the problem to select a subset of a given set of explanatory variables $\{X_1,\dots,X_p\}$. This can be framed as a model misspecification test problem, because a standard regression assumes that all variables that systematically influence the response variable are in the model. If it is of interest, as main test problem, to test $\beta_j = 0$ for a specific j, the MS test would be a test of null hypotheses $\beta_k = 0$ for one or more of the explanatory variables with $k \neq j$. The MC test would test $\beta_j = 0$ in a model with X_k removed, and the AU test would test $\beta_j = 0$ in a model including X_k. This problem was mentioned as second example in Bancroft (1944)'s seminal paper on preliminary assumption testing. Spanos (2018) however argued that this is very different from MS testing in the earlier discussed settings, because if a model including β_k is chosen based on a rejection of $\beta_k = 0$ by what is interpreted as MS test, the conditionally estimated β_k will be systematically large in absolute value, and can through dependence on the estimated β_j also be strongly dependent on the MC test.

Traditional model selection approaches such as forward selection and backward elimination are often based on such tests and have been analysed (and criticised) a lot in the literature. We will not review this literature here. There is sophisticated and innovative literature on post-selection inference in this problem. Berk et al. (2013) propose a procedure in which main inference is adjusted for simultaneous testing taking into account all possible sub-models that could have been selected. Efron (2014) uses bootstrap methods to do inference that takes the model selection process into account. Both approaches could also involve other MS testing such as of normality, homoscedasticity, or linearity assumptions, as long as combined procedures are fully specified. For specific model selection methods there now exists work allowing for exact post-selection inference, see Lee et al. (2016). For a critical perspective on these issues see Leeb & Pötscher (2005), Leeb et al. (2015), noting particularly that asymptotic results regarding the distribution of post-selection statistics (i.e., results of combined procedures) will not be uniformly valid for finite samples. In econometrics, David Hendry and co-workers developed an automatic modeling system that involves MS testing and conditional subsequent testing with adjustments for decisions in the modeling process, see, e.g., Discovery & in Econometrics (2014). They mentioned that their experience from experiments is that involving MS tests does not affect the final results much in case the model assumptions for the final procedure are fulfilled, however to our knowledge these experiments are nowhere published. Earlier, some authors such as Saleh & Sen (1983) analysed the effect of preliminary variable selection testing on later conditional main testing.

Godfrey (1988) listed a plethora of MS tests to test the various assumptions of linear regression. However, no systematic way to apply these tests was discussed. In fact, Godfrey noted that the literature left more questions open rather than answered. Some of these questions are: (i) the choice among different MS tests, (ii) whether to use nonparametric or parametric tests, (iii) what to do when any of the model assumptions are invalid as well as (iv) some potential problems with MS testing such as repeated use of data, multiple testing and pre-test bias. Godfrey (1996) concluded that efforts should be made to develop 'attractive', useful and simple combined procedures as these were lacking at the time; to a large extent this still is the case. One suggestion was to use the Bonferroni correction for each test as "*the asymptotic dependence of test statistics is likely to be the rule, rather than the exception, and this will reduce the constructive value of individual checks for misspecification*".

Giles & Giles (1993) reviewed the substantial amount of work done in econometrics regarding preliminary testing in regression up to that time, a limited amount of which is about MC and/or AU tests conditionally on MS tests. This involves pre-testing of a known fixed variance value, homoscedasticity, and independence against an auto-correlation alternatives. The cited results are mixed. King & Giles (1984) comment positively on a combined procedure in which absence of auto-correlation is tested first by a Durbin-Watson or t-test. Conditionally on the result of that MS test, either a standard t-test of a regression parameter was run (MC test), or a test based on an empirically generalised least squares estimator taking auto-correlation into account (AU test). In simulations the combined procedure performs similar to the MC test and better than the AU test in absence of auto-correlation, and similar to the AU test and better than the MC test in presence of auto-correlation. Also here it is recommended to run the MS test at a level higher than the usual 5%. Most related post-1993 work in econometrics seems to be on estimation after pre-testing, and regression model selection. Ohtani & Toyoda (1985) proposed a combined procedure for testing linear hypotheses in regression conditionally on testing for known variance. Toyoda & Ohtani (1986) tested the equality of different regressions conditionally on testing for equal variances. In both papers power gains for the combined procedure are reported, which are sometimes but not always accompanied with an increased type 1 error probability.

5.5 Cross-over trials

Cross-over trials are an example for a specific problem-adapted combined procedure discussed in the literature. In a two-treatment, two-period cross-over trial, patients are randomly allocated either to one group that receives treatment A followed by treatment B, or to another group that receives the treatments in the reverse order. The straightforward analysis of such data could analyse within-patients differences between the effects of the two treatments by a paired test (MC test). This requires the assumption that there is no "carry-over", i.e., no influence of the earlier treatment on the effect of the later treatment. In case that there is carry-over, the somewhat wasteful analysis of the effect of the first treatment only for each patient is safer (AU test). Grizzle (1967) proposed a combined procedure that became well established for some time. It consists of computing a score for each patient that contrasts the two treatment effects with the baseline values, and tests, e.g., using a two-sample t-test, whether this is the same on average in both groups, corresponding to the absence of carry-over on average (MS test). Freeman (1989) analysed this combined procedure analytically under a Gaussian assumption and potential existence of carry-over, comparing it to both the MC test and the AU test run unconditionally. He observed that due to strong dependence between the MS test and both the MC- and the AU-test, the combined procedure has more or less strongly inflated type 1 errors whether there is carry-over or not. Its power behaves typically for combined procedures, being better than the AU test but worse than the MC test in absence of carry-over and the other way round in its presence. Overall Freeman advises against the use of this procedure.

5.6 More than one misspecification test

Rasch et al. (2011) assessed the statistical properties of a three-stage procedure including testing for normality and for homogeneity of the variances taking into account a number of different distributions, and ratios of the standard deviation. They considered three main statistical tests, the Student's t-test, the Welch's t-test and the WMW test. For the MS testing, they used the Kolmogorov-Smirnov test for testing normality and Levene's test for testing the homogeneity of the variances of the two generated samples (Levene (1960)). If normality was rejected by the Kolmogorov-

Smirnov test, the WMW test was used. If normality was not rejected, the Levene's test was run and if homogeneity was rejected, the Welch's t-test was used and if homogeneity was not rejected, the standard t-test was used. The authors presented the rejection rates and the power of the procedure and compared it with the tests when the model assumption were not checked. Welch's t-test performed so well overall that the authors recommended its unconditional use, which is in line with recommendations by Rasch & Guiard (2004) from investigations of the robustness of various tests against non-normality. All of the investigated distributions had existing kurtosis, meaning that the tails were not really heavy. Furthermore some of the literature cited in Section 5.2 advised against using the Kolmogorov-Smirnov test, so that it is conceivable that more positive results for the combined procedure could have been achieved with a different setup. To our knowledge this is the only investigation of a combined procedure involving more than one MS test apart from the work on regression model selection cited in Section 5.4.

5.7 Discussion

Although many authors have, in one way or another, investigated the effects of preliminary MS testing or later application of model-based procedures, there are some limitations in the existing literature. Only very few papers have compared the performance of a fully specified combined procedure with unconditional uses of both the MC and the AU test. Some of these have only looked at type 1 error probabilities but not power, some have only looked at the situation in which the model assumption is in fact fulfilled, and some have studied setups in which either the unconditional MC or the AU test works well across the board, making a combined procedure superfluous, although it is widely acknowledged that situations in which either unconditional test can perform badly depending on the unknown data generating process do exist.

Reasons why authors advised against model checking in specific situations were:

(a) The MC test was better or at least not clearly worse than the AU test for all considered distributions in which the model assumptions of the MC test were not fulfilled (in which case the MC test can be used unconditionally),

(b) The AU test was not clearly worse than the MC test where model assumptions of the MC test were fulfilled (in which case the AU test can be used unconditionally),

(c) The MS test did not work well distinguishing situations in which the MC test was better from situations in which the AU test was better, possibly despite being good at testing just the formal model assumption.

(d) Due to dependence the application of the MS test distorted the performance of the conditionally performed tests.

For model checking to be worthwhile, these situations need to be avoided.

Comparing a full combined procedure with unconditional use of the MC test or the AU test, a typical pattern should be that under the model assumption for the MC test, the MC test is best regarding power, and the combined procedure performs between the unconditional MC test and AU test, and if that model assumption is violated, the AU test is best, and the combined procedure is once more between the MC test and the AU test. King & Giles (1984), Toyoda & Ohtani (1986) are examples for this. Results on test size are consistent with this (i.e., in cases where the combined procedure violates the nominal test level, at least one of the unconditional procedures does that as well). Such results can be interpreted charitably for the combined procedure, which allows for some kind of maximin performance. It seems to us that part of the criticism of the combined procedure is motivated by the fact that it does not do what some seem to expect or hope it to do, namely to help making sure that model assumptions are fulfilled, and to otherwise leave performance characteristics untouched, which is destroyed by the misspecification paradox. This however requires both the MC test and the AU test to be superior in some situations.

A sober look at the results reveals that the combined procedures are almost always competitive with at least one of the unconditional tests, and often with them both. It is clear, though, that recommendations need to depend on the specific problem, the specific tests involved. Results often also depend on in what way exactly model assumptions of the MC test are violated, which is hard to know without some kind of data dependent reasoning.

6 A positive result for combined procedures

The overall message from the literature does not seem very satisfactory. On the one hand, model assumptions are important and their violation can severely damage results. On the other hand, most comments on testing the model assumptions and conditionally choosing a main test are rather critical.

In this section we present a setup and a result that makes us assess the impact of preliminary model testing somewhat more positively. A characteristic of the literature analysing combined procedures is that they compare the combined procedure with unconditional MC or AU tests both in situations where the model assumption of the MC test is fulfilled, or not fulfilled. However, they do not investigate a situation in which the MS test can do what it is supposed to do, namely to *distinguish* between these situations. This can be modelled in the simplest case as follows, using the notation from Section 3. Let P_θ be a distribution that fulfills the model assumptions of the MC test, and $Q \in M \setminus M_\Theta$ a distribution that violates these assumptions. For considerations of power, let the null hypothesis of the main test be violated, i.e., $\theta \notin \Theta_0$ and $Q \notin M^*$ (an analogous setup is possible for considerations of size). We may observe data from P_θ or from Q. Assume that a dataset is with probability $\lambda \in [0,1]$ generated from P_θ and with probability $1-\lambda$ from Q (we stress that as opposed to standard mixture models, λ governs the distribution of the whole dataset, not every single observation independently). The cases $\lambda = 0$ and $\lambda = 1$ are those that have been treated in the literature, but only if $\lambda \in (0,1)$ the ability of the MS test to inform the researcher whether the data are more likely from P_θ or from Q is actually required.

We ran several simulations of such a setup (looking for example at normality in the two-sample problem), which will in detail be published elsewhere. Figure 1 shows a typical pattern of results. In this situation, for $\lambda = 0$ (model assumption violated), the AU test is best and the MC test is worst. For $\lambda = 1$, the MC test is best and the AU test is worst. The combined procedure is in between, which was mostly the case in our simulations. Here, the combined procedure is for both of these situations close to the better one of the unconditional tests (to what extent this holds depends on details of the setup). The powers of all three tests are linear functions of λ (linearity in the plot is distorted by random variation only), and the consequence is that the combined procedure performs clearly better than both unconditional tests over the best part of the range of λ. In our simulations it was mostly the case that for a good range of λ-values the combined procedure was the best. To brand the combined procedure "winner" would require the nominal level to be respected under H_0 (i.e., for both $P_\theta,\ \theta \in \Theta_0$ and $Q \in M^*$), which was very often though not always the case.

I such a setup relevant? Obviously it is not realistic that only two distributions are possible, one of which fulfills the model assumptions of the MC test. We wanted to keep the setup simple, but of course one could look at mixtures of a wider range of distributions, even a continuous range (for example for ratios between group-wise variances). In any case, the setup is more flexible than looking at $\lambda = 0$ and $\lambda = 1$ only, which is what has been done in the literature up to now. Of course model assumptions will never hold precisely, but the idea seems appealing to us that a researcher in a certain field who very often applies certain tests comes across a certain percentage different from 0 or 1 of cases which are well-behaved in the sense that a certain model assumption is a good if not perfect description of what is going on (the setup has a certain Bayesian flavor, but the researcher may not be interested in priors or posteriors for λ because the proportion λ under such an interpretation is pieced together from situations concerning different research topics).

We use the notation from Section 3 with the following additions. P_λ stands for distribution of the overall two step experiment, i.e., first selecting either $\tilde{P} = P_\theta$ or $\tilde{P} = Q$ with probabilities $\lambda,\ 1-\lambda$ respectively, and then generating a dataset z from $\tilde{P}$. The events of rejection of the respective H_0 are denoted $R_{MS} = \{\Phi_{MS}(z) = 1\}$, $R_{MC} = \{\Phi_{MC}(z) = 1\}$, $R_{AU} = \{\Phi_{AU}(z) = 1\}$, $R_C = \{\Phi_C(z) = 1\}$. Here are some assumptions:

(I) $\Delta_\theta = P_\theta(R_{MC}) - P_\theta(R_{AU}) > 0,$

(II) $\Delta_Q = Q(R_{AU}) - Q(R_{MC}) > 0,$

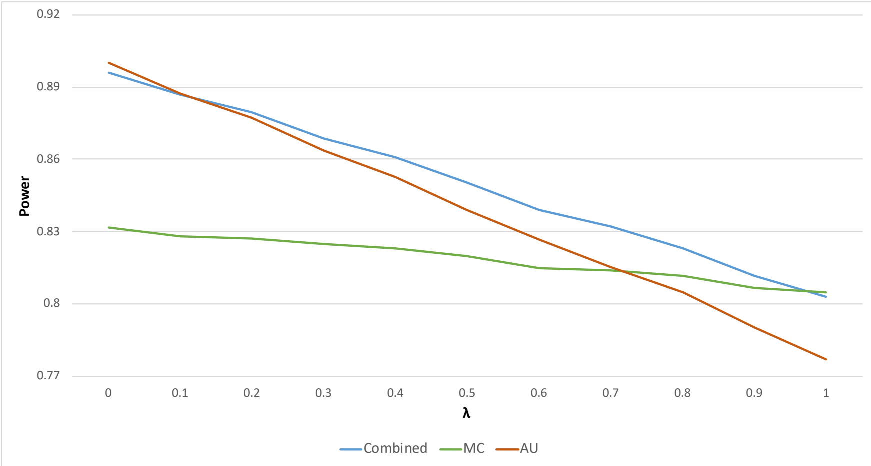
(III) $\alpha^*_{MS} = Q(R_{MS}) > \alpha_{MS} = P_\theta(R_{MS}),$

(IV) Both R_{MC} and R_{AU} are independent of R_{MS} under both P_θ and Q.

Keep in mind that this is about power, i.e., we take the H_0 of the main test as violated for both P_θ and Q. Assumption (I) means that the MC test has the better power under P_θ, (II) means that the AU test has the better power under Q. Assumption (III) means that the MS test has some use, i.e., it has a certain (possibly weak) ability to distinguish between P_θ and Q. All these are essential requirements for preliminary model assumption testing to make sense. Assumption (IV) though is very restrictive. It asks that rejection of the main null hypothesis by both main tests is independent of the decision made by the MS test. This is unrealistic in most situations. However, it can be relaxed (at the price of a more tedious proof that we do not present here) to demanding that there is a small enough $\delta > 0$ (dependent on the involved probabilities) so that $|P_\theta(R_{MC}|R_{MS}) - P_\theta(R_{MC}|R^c_{MS})|,\ |P_\theta(R_{AU}|R_{MS}) - P_\theta(R_{AU}|R^c_{MS})|,\ |Q(R_{MC}|R_{MS}) - Q(R_{MC}|R^c_{MS})|$, and $|Q(R_{AU}|R_{MS}) - Q(R_{AU}|R^c_{MS})|$ are all smaller than δ, which can be fulfilled in many cases of interest. As emphasised earlier, approximate independence of the MS test and the main tests has also been found in other literature to be an important desirable feature of a combined test, and it should not surprise that a condition of this kind is required.

The following Lemma states that the combined procedure has a better power than both the MC test and the AU test for at least some λ. Although this in itself is not a particularly strong result, in many situations, according to our simulations, the range of λ for which this holds is quite large. Furthermore the result concerns general models and choices of tests, whereas to our knowledge everything that already exists in the literature is for specific choices.

Figure 1: Power of combined procedure, MC, and AU test across different λs from an exemplary simulation. The MC test here is Welch's two-sample t-test, the AU test the WMW-test, the MS test Shapiro-Wilks, for $\lambda = 1$ corresponds to normal distributions with mean difference 1, $\lambda = 0$ corresponds to t_3-distributions with mean difference 1.



Despite the somewhat restrictive set of assumptions, none of the involved tests and distributions is actually specified, so that the Lemma (at least with a relaxed version of (IV)) applies to a very wide range of problems.

Lemma 1. *Assuming (I)-(IV),* $\exists\lambda \in (0,1)$ *such that both* $P_\lambda(R_C) > P_\lambda(R_{MC})$ *and* $P_\lambda(R_C) > P_\lambda(R_{AU})$.

Proof. Obviously,

$$P_\lambda(R_{MC}) = \lambda P_\theta(R_{MC}) + (1-\lambda)Q(R_{MC}),$$
$$P_\lambda(R_{AU}) = \lambda P_\theta(R_{AU}) + (1-\lambda)Q(R_{AU}).$$

By (I), for $\lambda = 1:\ P_\lambda(R_{MC}) > P_\lambda(R_{AU})$ and, by (II), for $\lambda = 0:\ P_\lambda(R_{AU}) > P_\lambda(R_{MC})$. As $P_\lambda(R_{MC})$ and $P_\lambda(R_{AU})$ are linear functions of λ, there must be $\lambda^* \in (0,1)$ so that $P_{\lambda^*}(R_{AU}) = P_{\lambda^*}(R_{MC})$. Obtain

$$P_{\lambda^*}(R_{MC}) = P_{\lambda^*}(R_{AU}) \Leftrightarrow$$
$$\lambda^* P_\theta(R_{MC}) + (1-\lambda^*)Q(R_{MC}) = \lambda^* P_\theta(R_{AU}) + (1-\lambda^*)Q(R_{AU}) \Leftrightarrow$$
$$\lambda^*(\Delta_\theta + \Delta_Q) = \Delta_Q \Leftrightarrow$$
$$\lambda^* = \frac{\Delta_Q}{\Delta_\theta + \Delta_Q}.$$

This yields, by help of (IV),

$$
\begin{aligned}
P_{\lambda^*}(R_C) &= \lambda^* P_\theta(R_C) + (1-\lambda^*)Q(R_C) \\
&= \lambda^* \left[\alpha_{MS} P_\theta(R_{AU}|R_{MS}) + (1-\alpha_{MS}) P_\theta(R_{MC}|R_{MS}^c)\right] \\
&+ (1-\lambda^*) \left[\alpha_{MS}^* Q(R_{AU}|R_{MS}) + (1-\alpha_{MS}^*) Q(R_{MC}|R_{MS}^c)\right] \\
&= \lambda^* \left[\alpha_{MS} P_\theta(R_{AU}) + (1-\alpha_{MS}) P_\theta(R_{MC})\right] \\
&+ (1-\lambda^*) \left[\alpha_{MS}^* Q(R_{AU}) + (1-\alpha_{MS}^*) Q(R_{MC})\right] \\
&= \frac{\Delta_Q}{\Delta_\theta + \Delta_Q}\left[-\alpha_{MS}\Delta_\theta - \alpha_{MS}^*\Delta_Q\right] + \alpha_{MS}^*\Delta_Q + P_{\lambda^*}(R_{MC}) \\
&= \Delta_Q \left[\frac{-\alpha_{MS}\Delta_\theta - \alpha_{MS}^*\Delta_Q + \alpha_{MS}^*\Delta_\theta + \alpha_{MS}^*\Delta_Q}{\Delta_\theta + \Delta_Q}\right] + P_{\lambda^*}(R_{MC}) \\
&= \frac{\Delta_\theta \Delta_Q}{\Delta_\theta + \Delta_Q}\left[\alpha_{MS}^* - \alpha_{MS}\right] + P_{\lambda^*}(R_{MC}) \\
&= \frac{\Delta_\theta \Delta_Q}{\Delta_\theta + \Delta_Q}\left[\alpha_{MS}^* - \alpha_{MS}\right] + P_{\lambda^*}(R_{AU}).
\end{aligned}
$$

$\frac{\Delta_\theta \Delta_Q}{\Delta_\theta + \Delta_Q}\left[\alpha_{MS}^* - \alpha_{MS}\right]$ is larger than zero by (I)-(III), so $P_{\lambda^*}(R_C)$ is larger than both $P_{\lambda^*}(R_{MC})$ and $P_{\lambda^*}(R_{AU})$. □

7 Conclusion

Given that statisticians often emphasise that statistical inference relies on model assumptions, and that these need to be checked, the literature investigating this practice is surprisingly critical. Preliminary tests of model assumptions have in many situations been found to affect the characteristics of subsequent inference and to invalidate the theory based on the very model assumptions the approach was meant to secure. In some setups either running a less constrained test or running the model-based test without preliminary testing have been found superior to the combined procedure involving preliminary MS testing. This is in contrast to a fairly general view among statisticians that model assumptions should be checked. The existence of situations in which performance characteristics rely strongly on whether model assumptions are fulfilled or not has been acknowledged also by authors that were more critical of preliminary testing, and therefore there is certainly a role for model checking. There is however little elaboration of its benefits in the literature. A key contribution of the present work is the investigation of general combined procedures in a setup in which both distributions fulfilling and violating model assumptions can occur. This is more favourable for combined procedures than just looking at either fulfilled or violated model assumptions in isolation.

We believe that overall the literature gives a somewhat too pessimistic assessment of combined procedures involving MS testing, and that model checking (and drawing consequences from the result) is more useful than the literature suggests. The fact that preliminary assumption checking technically violates the assumptions it is meant to secure is probably assessed more negatively from the position that models can and should be "true", whereas it may be a rather mild problem if it is acknowledged that model assumptions, while providing ideal and potentially optimal conditions for the application of model-based procedures, are not necessary conditions for their use.

Lemma 1 also serves to give an idea of the required ingredients for successful model checking, i.e., what is important for the combined procedure to be superior to both the MC and the AU test. In order to put this into practice, the researcher should have at least a rough idea about what kinds of deviations from the model assumptions of the MC test may happen, although one may also use "worst cases" (such as distributions with non-existing variances for t-tests) as a starting point. Call $\{P_\theta\}$ the family of distributions that fulfill the model assumptions of the MC test, and Q a possible distribution that violates these assumptions; one can also involve different options for Q.

(a) The MC test should be clearly better than the AU test if its model assumptions are fulfilled (otherwise the unconditional AU test can be used without much performance loss).

(b) The AU test should be clearly better than the MC test for Q (otherwise the unconditional MC test can be used without much performance loss).

(c) The MS test should be good at distinguishing $\{P_\theta\}$ from Q.

(d) The MS test Φ_{MS} should be approximately independent of both Φ_{MC} and Φ_{AU} under $\{P_\theta\}$ and Q.

In practice it is of course not known what Q will be encountered, but given the unsatisfactory state of the art, developing combined procedures fulfilling (a)-(d) based on choices of Q seems a promising approach to improve matters.

Considering informal (visual) model checking, issues (a) and (b) are not different from formal combined procedures, although the visual display may help to pick a suitable AU test (be it implicitly by formulating a model that does not require a rejected assumption). An expert data analyst may do better based on suitable graphs than existing formal procedures regarding (c); many users will probably do worse (see Hoekstra et al. (2012) for a study investigating misconceptions and lack of knowledge about model checking among empirical researchers). (d) may be plausible if displays are used in which the parameters tested by the MC and AU test such as location or regression parameters do not have a visible impact, such as residual plots, although there is a danger of this being critically violated in case that the AU test is chosen based on what is seen in the graphs.

We believe that the focus of model checking is too much on the formal assumptions and not enough on deriving tests that can find the particular violations of model assumptions that are most problematic in terms of level and power (issue (c) above in case Q is chosen accordingly).

The development of MS tests that are better suited for this task and the investigation of the resulting combined procedures is a promising research area. We believe that the approach of Lemma 1 considering a random draw of either fulfilled and violated model assumptions could also help in more complex situations, for example concerning different assumption violations, more than one MS test, and more than two main tests.

References

Abdulhafedh, A. (2017), 'How to detect and remove temporal autocorrelation in vehicular crash data', *Journal of Transportation Technologies* **7**, 133–147.

Albers, W., Boon, P. C. & Kallenberg, W. C. (2000a), 'The asymptotic behavior of tests for normal means based on a variance pre-test', *Journal of Statistical Planning and Inference* **88**, 47–57.

Albers, W., Boon, P. C. & Kallenberg, W. C. (2000b), 'Size and power of pretest procedures', *Annals of Statistics* **28**, 195–214.

Arnold, B. C. (1970), 'Hypothesis testing incorporating a preliminary test of significance', *Journal of the American Statistical Association* **65**, 1590–1596.

Bahadur, R. & Savage, L. (1956), 'The nonexistence of certain statistical procedures in nonparametric problems', *Annals of Mathematical Statistics* **27**, 1115–1122.

Bancroft, T. A. (1944), 'On biases in estimation due to the use of preliminary tests of significance', *Annals of Mathematical Statistics* **15**, 190–204.

Bancroft, T. A. (1964), 'Analysis and inference for incompletely specified models involving the use of preliminary test(s) of significance', *Biometrics* **20**, 427–442.

Bancroft, T. A. & Han, C. (1977), 'Inference based on conditional specification: A note and a bibliography', *International Statistical Review* **45**, 117–127.

Bartlett, M. S. (1935), 'The effect of non-normality on the *t* distribution', *Mathematical Proceedings of the Cambridge Philosophical Society* **31**, 223–231.

Berk, R., Brown, L., Buja, A., Zhang, K. & Zhao, L. (2013), 'Berk, r., brown, l., buja, a., zhang, k., zhao, l.', *Annals of Statistics* **41**, 802–837.

Bickel, D. R. (2015), 'Inference after checking multiple bayesian models for data conflict and applications to mitigating the influence of rejected priors', *International Journal of Approximate Reasoning* **66**, 53–72.

Chatfield, C. (1995), 'Model uncertainty, data mining and statistical inference (with discussion)', *Journal of the Royal Statistical Society, Series B* **158**, 419–466.

Cox, D. R. (2006), *Principles of Statistical Inference*, Cambridge University Press, Cambridge.

Cressie, N. (1980), 'Relaxing assumptions in the one-sample *t*-test', *Australian Journal of Statistics* **22**, 143–153.

Davies, P. L. (2014), *Data Analysis and Approximate Models*, Chapman & Hall/CRC, Boca Raton FL.

de Finetti, B. (1974), *Theory of Probability*, Wiley, New York.

Discovery, E. M. & in Econometrics, T. E. A. S. M. (2014), *Hendry, D. and Doornik, J.*, MIT Press, Cambridge MA.

Donoho, D. (1988), 'One-sided inference about functionals of a density', *Annals of Statistics* **16**, 1390–1420.

Dowdy, S., Wearden, S. & Chilko, D. (2004), *Statistics for Research*, Wiley, New York.

Draper, D. (1995), 'Assessment and propagation of model uncertainty (with discussion)', *Journal of the Royal Statistical Society, Series B* **57**, 45–97.

Easterling, R. G. (1976), 'Goodness of fit and parameter estimation', *Technometrics* **18**, 1–9.

Easterling, R. G. & Anderson, H. E. (1978), 'The effect of preliminary normality goodness of fit tests on subsequent inference', *Journal of Statistical Computation and Simulation* **8**, 1–11.

Efron, B. (2014), 'Estimation and accuracy after model selection', *Journal of the American Statistical Association* **109**, 991–1007.

Farrell, P. J. & Rogers-Stewart, K. (2006), 'Comprehensive study of tests for normality and symmetry: extending the spiegelhalter test', *Journal of Statistical Computation and Simulation* **76**, 803–816.

Fay, M. P. & Proschan, M. A. (2010), 'Wilcoxon-mann-whitney or *t*-test? on assumptions for hypothesis tests and multiple interpretations of decision rules', *Statistics Surveys* **4**, 1–39.

Fisher, F. M. (1961), 'On the cost of approximate specification in simultaneous equation estimation', *Econometrica: Journal of the Econometric Society* **29**, 139–170.

Fisher, R. A. (1922), 'On the mathematical foundation of theoretical statistics', *Philosophical Transactions of the Royal Society of London A* **22**, 309–368.

Freeman, P. (1989), 'The performance of the two-stage analysis of two-treatment, two-period cross-over trials', *Statistics in Medicine* **8**, 1421–1432.

Gambichler, T., Bader, A., Vojvodic, M., Bechara, F. G., Sauermann, K., Altmeyer, P. & Hoffmann, K. (2002), 'Impact of uva exposure on psychological parameters and circulating serotonin and melatonin', *BMC Dermatology* **2**, 163–174.

Gans, D. J. (1981), 'Use of a preliminary test in comparing two sample means', *Communications in Statistics - Simulation and Computation* **10**, 163–174.

Gelman, A. & Loken, E. (2014), 'The statistical crisis in science', *The American Statistician* **102**, 460–465.

Gelman, A. & Shalizi, C. R. (2013), 'Philosophy and the practice of bayesian statistics', *British Journal of Mathematical and Statistical Psychology* **66**, 8–38.

Giles, D. E. A. & Giles, J. A. (1993), 'Pre-test estimation and testing in econometrics: Recent developments', *Journal of Economic Surveys* **7**, 145–197.

Godfrey, L. G. (1988), *Misspecification tests in econometrics. The Lagrange Multiplier principle and other applications*, Cambridge University Press, Cambridge.

Godfrey, L. G. (1996), 'Misspecification tests and their uses in econometrics', *Journal of Statistical Planning and Inference* **49**, 241–260.

Grizzle, J. E. (1967), 'The two-period change-over design and its use in clinical trials', *Biometrics* **21**, 469–480 (Corrigendum in Biometrics, 30, 727, 1974).

Gupta, V. P. & Srivastava, V. K. (1993), 'Upper bound for the size of a test procedure using preliminary tests of significance', *Journal of the Indian Statistical Association* **7**, 26–29.

Gurland, J. & McCullough, R. (1962), 'Testing equality of means after a preliminary test of equality of variances', *Biometrika* **49**, 403–417.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. & Stahel, W. A. (1986), *Robust Statistics*, Wiley, New York.

Hasler, G., Suker, S., Schoretsanitis, G. & Mihov, Y. (2020), 'Sustained improvement of negative self-schema after a single ketamine infusion: An open-label study', *Frontiers in Neuroscience* **14**, 687.

He, X., Simpson, D. G. & Portnoy, S. L. (1990), 'Breakdown robustness of tests', *Journal of the American Statistical Association* **85**, 446–452.

Hennig, C. (2010), 'Falsification of propensity models by statistical tests and the goodness-of-fit paradox', *Philosophia Mathematica* **15**, 166–192.

Hoekstra, R., Kiers, H. & Johnson, A. (2012), 'Are assumptions of well-known statistical techniques checked, and why (not)?', *Frontiers in Psychology* **3**, 137.

Hollander, M. & Sethuraman, J. (2001), Nonparametric statistics: Rank-based methods, *in* N. J. Smelser & P. B. Baltes, eds, 'International Encyclopedia of the Social and Behavioral Sciences', Pergamon, Oxford, pp. 10673–10680.

Holman, A. J. & Myers, R. R. (2005), 'A randomized, double-blind, placebo-controlled trial of pramipexole, a dopamine agonist, in patients with fibromyalgia receiving concomitant medications', *Arthritis & Rheumatism* **52**, 2495–2505.

Hsu, P. L. (1938), 'Contribution to the theory of "student's" t-test as applied to the problem of two samples', *Statistical Research Memoirs* **2**, 1–24.

Hsu, T. C. & Feldt, L. S. (1969), 'The effect of limitations on the number of criterion score values on the significance level of the F-test', *American Educational Research Journal* **6**, 515–527.

Kass, R. E., Caffo, B. S., Davidian, M., Meng, X. L., Yu, B. & Reid, N. (2016), 'Ten simple rules for effective statistical practice', *PLoS Computational Biology* **12**, e1004961.

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., Kovalchuk, R. K., Lowman, L. L., Petoskey, M. D., Keselman, J. C. & Levin, J. R. (1998), 'Statistical practices of educational researchers: An analysis of their anova, manova, and ancova analyses', *Review of Educational Research* **68**, 350–386.

Keselman, H. J., Othman, A. R. & Wilcox, R. R. (2013), 'Preliminary testing for normality: Is this a good practice?', *Journal of Modern Applied Statistical Methods* **12**, 2–19.

Keskin, S. (2006), 'Comparison of several univariate normality tests regarding type i error rate and power of the test in simulation based small samples', *Journal of Applied Science Research* **2**, 296–300.

King, M. L. & Giles, D. E. A. (1984), 'Autocorrelation pre-testing in the linear model: Estimation, testing and prediction', *Journal of Econometrics* **25**, 35–48.

Kokosinska, D., Gieraltowski, J. J., Zebrowski, J. J., Orlowska-Baranowska, E. & Baranowski, R. (2018), 'Heart rate variability, multifractal multiscale patterns and their assessment criteria', *Physiological Measurement* **39**, 114010.

Lee, J. D., Sun, D. L., Sun, Y. & Taylor, J. E. (2016), 'Exact post-selection inference, with application to the lasso', *Annals of Statistics* **44**, 907–927.

Leeb, H. & Pötscher, B. M. (2005), 'Model selection and inference: Fact and fiction', *Econometric Theory* **21**, 21–59.

Leeb, H., Pötscher, B. M. & Ewald, K. (2015), 'On various confidence intervals post-model-selection', *Statistical Science* **30**, 216–227.

Lehmann, E. L. & Romano, J. P. (2005), *Testing Statistical Hypotheses*, Springer, New York.

Levene, H. (1960), Robust tests for equality of variances, *in* I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow & H. B. Mann, eds, 'Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling', Stanford University Press, Redwood City CA, pp. 278–292.

Markowski, C. A. & Markowski, E. P. (1990), 'Conditions for the effectiveness of a preliminary test of variance', *The American Statistician* **44**, 131–136.

Maydeu-Olivares, A., Forero, C. A., Gallardo-Pujol, D. & Renom, J. (2009), 'Testing categorized bivariate normality with two-stage polychoric correlation estimates', *Methodology* **5**, 131–136.

Mayo, D. G. (2018), *Statistical Inference as Severe Testing*, Cambridge University Press, Cambridge.

Mendes, M. & Pala, A. (2003), 'Type i error rate and power of three normality tests', *Pakistan Journal of Information and Technology* **2**, 135–139.

Moder, K. (2010), 'Alternatives to F-test in one way *ANOVA* in case of heterogeneity of variances (a simulation study)', *Psychological Testing and Assessment Modeling* **52**, 343–353.

Moser, B. K. & Stevens, G. R. (1992), 'Homogeneity of variance in the two-sample means test', *The American Statistician* **46**, 19–21.

Moser, B. K., Stevens, G. R. & Watts, C. L. (1989), 'The two-sample t test versus satterthwaite's approximate f test', *Communications in Statistics-Theory and Methods* **18**, 3963–3975.

Neave, H. R. & Granger, C. W. J. (1968), 'A monte carlo study comparing various two sample tests for differences in means', *Technometrics* **10**, 509–522.

Neyman, J. (1952), *Lectures and Conferences on Mathematical Statistics and Probability (2nd ed.)*, U.S. Department of Agriculture, Washington DC.

Nour-Eldein, H. (2016), 'Statistical methods and errors in family medicine articles between 2010 and 2014-suez canal university, egypt: A cross-sectional study', *Journal of Family Medicine and Primary Care* **5**, 24–33.

Ohtani, K. & Toyoda, T. (1985), 'Testing linear hypothesis on regression coefficients after a pre-test for disturbance variance', *Economics Letters* **17**, 111–114.

Pearson, K. (1900), 'On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling', *Philosophical Magazine* **5**, 157–175.

Posten, H. O., Yeh, H. C. & Owen, D. B. (1982), 'Robustness of the two-sample *t*-test under violations of the homogeneity of variance assumptions', *Communications in Statistics: Theory and Methods* **11**, 109–126.

Rasch, D. & Guiard, V. (2004), 'The robustness of parametric statistical methods', *Psychology Science* **46**, 175–208.

Rasch, D., Kubinger, K. D. & Moder, K. (2011), 'The two-sample *t* test: pre-testing its assumptions does not pay off', *Statistical Papers* **52**, 219–231.

Ravichandran, J. (2012), 'A review of preliminary test-based statistical methods for the benefit of six sigma quality practitioners', *Statistical Papers* **53**, 531–547.

Razali, N. M. & Wah, Y. B. (2011), 'Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests', *Journal of Statistical Modeling and Analytics* **2**, 21–33.

Rochon, J., Gondan, M. & Kieser, M. (2012), 'To test or not to test: Preliminary assessment of normality when comparing two independent samples', *BMC Medical Research Methodology* **12**, 81–91.

Rochon, J. & Kieser, M. (2011), 'A closer look at the effect of preliminary goodness-of-fit testing for normality for the one-sample *t*-test', *British Journal of Mathematical and Statistical Psychology* **64**, 410–426.

Saleh, A. M. E. (2006), *Theory of preliminary test and Stein-type estimation with applications*, Wiley-Interscience, Hoboken, NJ.

Saleh, A. M. E. & Sen, P. K. (1983), 'Asymptotic properties of tests of hypothesis following a preliminary test', *Statistics and Risk Modeling* **1**, 455–478.

Satterthwaite, F. E. (1946), 'An approximate distribution of estimates of variance components', *Biometrics Bulletin* **2**, 110–114.

Scheffé, H. (1970), 'Practical solutions of the behrens-fisher problem', *Journal of the American Statistical Association* **65**, 1501–1508.

Schoder, V., Himmelmann, A. & Wilhelm, K. P. (2006*a*), 'Preliminary testing for normality: some statistical aspects of a common concept', *Clinical and Experimental Dermatology* **31**, 757–761.

Schoder, V., Himmelmann, A. & Wilhelm, K. P. (2006*b*), 'Preliminary testing for normality: some statistical aspects of a common concept', *Communications in Statistics - Theory and Methods* **35**, 2275–2286.

Spanos, A. (1999), *Probability Theory and Statistical Inference: Econometric Modeling with Observational Data*, Cambridge University Press, Cambridge.

Spanos, A. (2010), 'Akaike-type criteria and the reliability of inference: Model selection versus statistical model specification', *Journal of Econometrics* **158**, 204–220.

Spanos, A. (2018), 'Mis-specification testing in retrospect', *Journal of Economic Surveys* **32**, 541–577.

Sridharan, K. & Gowri, S. (2015), 'Reporting quality of statistics in indian journals: Analysis of articles over a period of two years', *Journal of Scientometric Research* **4**, 10–13.

Strasak, A. M., Zaman, Q., Marinell, G., Pfeiffer, K. P. & Ulmer, H. (2007a), 'The use of statistics in medical research: A comparison of the new england journal of medicine and nature medicine', *The American Statistician* **61**, 47–55.

Strasak, A. M., Zaman, Q., Marinell, G., Pfeiffer, K. P. & Ulmer, H. (2007b), 'The use of statistics in medical research: A comparison of wiener klinische wochenschrift and wiener medizinische wochenschrift', *Austrian Journal of Statistics* **36**, 141–152.

Toyoda, T. & Ohtani, K. (1986), 'Testing equality between sets of coefficients after a preliminary test for equality of disturbance variances in two linear regressions', *Journal of Econometrics* **31**, 67–80.

Welch, B. L. (1938), 'The significance of the difference between two means when the population variances are unequal', *Biometrika* **29**, 350–62.

Welch, B. L. (1947), 'The generalisation of student's problem when several different population variances are involved', *Biometrika* **34**, 28–35.

Wiedermann, W. & Alexandrowicz, R. (2007), 'A plea for more general tests than those for location only: Further considerations on rasch & guiard's 'the robustness of parametric statistical methods'', *Psychology Science* **49**, 2–12.

Wu, S., Jin, Z., Wei, X., Gao, Q., Lu, J., Ma, X., Wu, C., He, Q., Wu, M., Wang, R., Xu & He (2011), 'Misuse of statistical methods in 10 leading chinese medical journals in 1998 and 2008', *The Scientific World Journal* **11**, 2106–2114.

Wu, W., Hartless, J., Tesei, A., Gunji, V., Ayer, S. & London, J. (2019), 'Design assessment in virtual and mixed reality environments: Comparison of novices and experts', *Journal of Construction Engineering and Management* **145**, 04019049.

Zimmerman, D. W. (2006), 'Two separate effects of variance heterogeneity on the validity and power of significance tests of location', *Statistical Methodology* **3**, 351–374.

Zimmerman, D. W. (2011), 'A simple and effective decision rule for choosing a significance test to protect against non-normality', *British Journal of Mathematical and Statistical Psychology* **64**, 388–409.

Zimmerman, D. W. (2014), 'Consequences of choosing samples in hypothesis testing to ensure homogeneity of variance', *British Journal of Mathematical and Statistical Psychology* **67**, 1–29.