



Université Libre de Bruxelles

Faculté des Sciences Psychologiques et de l'Éducation

**TITRE,
TITRE SUITE**

Par

MARIE DELACRE

En vue de l'obtention du grade de docteur

Septembre 2021

Abstract

Abstract (voir nombre de mots requis)

Table des matières

1	Chapitre 1: Introduction	iv
1.0.1	Limite 2: hypothèse nulle	viii
1.0.2	Pourquoi jusque là la sauce n'a pas pris?	ix
2	Chapitre 2: le test t de Welch (article 1)	1
3	Chapitre 3: le test W de Welch (article 2)	11
4	Chapitre 4: le d de Cohen avec variances non poolées (article 3)	12
5	Chapitre 5: les tests d'équivalence (article 4)	13
6	Conclusion	14
6.0.1	Usage des articles méthodo	14
6.0.2	Importance des simulations et des logiciels modernes pour enseigner les statis- tiques fréquentistes	14
6.0.3	Comment écrire/transmettre l'info aux psys	15
6.0.4	Recommandations générales	15
7	Bibliographie	16
8	Annexe(s)	19
8.1	Annexe A: erratum	19
8.1.1	Why psychologists Should by Default Use Welch's t -test Instead of Student's t -test	19
8.1.2	Taking parametric assumptions very seriously	20
8.1.3	Effect sizes	20
8.1.4	Equivalence tests	20
8.2	Annexe B	21

Remerciements

Thank you for following this tutorial!

I hope you'll find it useful to write a very professional dissertation.

1 Chapitre 1: Introduction

On attend des chercheurs en psychologie, et des psychologues en général, qu'ils soient capables de produire des connaissances fondées sur des preuves scientifiques (et non sur des croyances et opinions), et également de comprendre et évaluer les recherches menées par d'autres (Haslam & McGarty, 2014). Or, dans un domaine dominé par les analyses quantitatives¹ (Counsell & Harlow, 2017), les connaissances statistiques s'avèrent fondamentales pour comprendre, planifier et analyser une recherche (Everitt, 2001 ; Howitt & Cramer, 2017). Les statistiques font dès lors partie intégrante du cursus de formation des psychologues et jouent un rôle très important dans leur parcours (Hoekstra, Kiers, & Johnson, 2012).

Traditionnellement, depuis plus de 50 ans, les tests-*t* et les *ANOVA* se trouvent au coeur de la grande majorité des programmes dans les domaines des Sciences Psychologiques et de l'Education (Aiken, West, & Millsap, 2008 ; Curtis & Harwell, 1998 ; Golinski & Cribbie, 2009) et des livres d'introduction aux statistiques pour psychologues [Field (2013); autres exemples?]. Cela pourrait vraisemblablement expliquer pourquoi ils sont si persistants dans la recherche en psychologie (Counsell & Harlow, 2017). Ces tests sont les plus fréquemment cités dans la littérature scientifique depuis plus de 60 ans (Byrne, 1996 ; Golinski & Cribbie, 2009 ; Nunnally, 1960). Dans une revue de 486 articles publiés en 2000 dans des journaux populaires en psychologie², Golinski & Cribbie (2009) avaient relevé 140 articles ($\approx 29\%$) au sein desquels les auteurs avaient mené au moins une *ANOVA* à un ou plusieurs facteurs. Plus récemment, Counsell & Harlow (2017) mentionnaient que parmi un ensemble de 151 études soumises dans 4 revues canadiennes en 2013, environ 40% incluaient une comparaison de moyennes. Peut-être est-ce en raison de leur grande fréquence d'usage, ajoutée à leur apparente simplicité, qu'on tend à croire que la plupart des chercheurs, si pas tous, ont une bonne maîtrise des tests de comparaisons de moyennes (Aiken et al., 2008 ; Hoekstra et al., 2012). Pourtant, certains indices semblent contredire cette conviction.

Bien qu'il existe plusieurs types de tests *t* et d'*ANOVA*, les chercheurs en psychologie tendent à privilégier par défaut le test *t* de Student et l'*ANOVA* de Fisher. La statistique *t* de Student se calcule

1. parmi 68 articles analysés en 2013 par Counsell et ses collaborateurs (2017) dans 4 revues canadiennes, 92.7% incluaient au moins une analyse quantitative (contre 7.3% incluant une analyse qualitative)

2. Les revues analysées étaient les suivantes: "Child Development", "Journal of Abnormal Psychology", "Journal of Consulting and Clinical Psychology", "Journal of Experimental Psychology: General", "Journal of Personality" et "Social Psychology"

comme suit (Student, 1908):

$$t_{Student} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{N-2}\right) \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad (1)$$

où N = le nombre total de sujets, et n_j et \bar{X}_j sont respectivement la taille et la moyenne du $j^{ème}$ échantillon ($j = 1, 2$). Sous l'hypothèse de normalité, la statistique t de Student suit une distribution t avec $n_1 + n_2 - 2$ degrés de liberté. La statistique F de Fisher se calcule comme suit:

$$F = \frac{\frac{1}{k-1} \sum_{j=1}^k \left[n_j (\bar{x}_j - \bar{x}_{..})^2 \right]}{\frac{1}{N-k} \sum_{j=1}^k \left[(n_j - 1) S_j^2 \right]} \quad (2)$$

où k est le nombre d'échantillons indépendants et S_j^2 est la variance du $j^{ème}$ échantillon ($1 \leq j \leq k$).

Sous l'hypothèse de normalité, la statistique F suit loi de Fisher caractérisée par 2 paramètres:

$$df_1 = k - 1$$

$$df_2 = \sum_{j=1}^k n_j - k$$

Le test t de Student et l'ANOVA consistent à comparer les scores moyens de deux (ou plusieurs) groupes indépendants de sujets. Les deux tests reposent sur les hypothèses que les résidus, indépendants et identiquement distribués soient extraits d'une population qui se distribue normalement et qui a la même variance au sein de chaque groupe (c'est ce qu'on appelle la condition d'homogénéité des variances, requise pour pouvoir calculer le terme de variance poolée qui apparaît au dénominateur des équations (1) et (2)). Pourtant, on constate que dans les articles publiés, il n'est que rarement fait mention de ces conditions. Osborne & Christianson (2001), par exemple, avaient trouvé que seulement 8% des auteurs reportaient des informations sur les conditions d'application des tests, soit à peine 1% de plus qu'en 1969. Plus récemment, Hoekstra et al. (2012) ont montré que sur 50 articles publiés en 2011 dans *Psychological Science* utilisant au moins une ANOVA, test- t ou régression, seulement trois discutaient des questions de normalité et d'homogénéité des variances. Par ailleurs, les informations reportées sont souvent non exhaustives (Counsell & Harlow, 2017), et la condition d'homogénéité des variances est encore moins fréquemment citée que celle de normalité. Parmi les 61 articles analysés par Keselman et al. (1998), seulement 5% des articles mentionnaient simultanément les conditions de normalité et d'homogénéité des variances (et en tout, la condition de normalité était mentionnée dans 11% des cas, contre seulement 8% pour la condition d'homogénéité des variances). Golinski & Cribbie (2009) ont fait un constat similaire: parmi les 140 articles qu'ils ont analysé, seulement 11 mentionnaient

explicitement la condition de normalité, contre 3 qui mentionnaient celle d'homogénéité des variances.

Notons que la non mention des conditions d'application dans les articles ne veut pas forcément dire qu'elles n'ont pas été prises en compte dans les analyses. On pourrait imaginer que les auteurs vérifient les conditions d'application des tests mais ne le mentionnent la plupart du temps que lorsqu'elles sont violées (Counsell & Harlow, 2017). Golinski & Cribbie (2009), par exemple, ont constaté à travers leurs revue de littérature que parmi les 11 articles qui mentionnaient la condition de normalité, 10 montraient une violation de cette dernière. Il est possible que motivés par le désir de rentabiliser l'espace disponible dans les manuscrits (Counsell & Harlow, 2017), les auteurs soient tentés de se limiter aux informations explicitement demandées par les éditeurs et les reviewers des journaux (Counsell & Harlow, 2017). Or, les informations relatives aux conditions d'application des tests en font rarement partie. Par exemple, leur report n'est pas explicitement demandé dans le manuel des normes APA (Hoekstra et al., 2012)³. Dans un tel contexte, il n'y a que peu d'intérêt pour les chercheurs à en discuter, si ce n'est pour justifier une décision inhérente à leur violation. Il est néanmoins surprenant de constater que de telles discussions apparaissent dans un pourcentage si faible d'articles, puisqu'il a été argumenté à de nombreuses reprises que le respect des conditions de normalité et d'homogénéité des variances est plus l'exception que la norme dans de nombreux domaines de la psychologie (Cain, Zhang, & Yuan, 2017 ; Erceg-Hurn & Mirosevich, 2008 ; Grissom, 2000 ; Micceri, 1989 ; Yuan, Bentler, & Chan, 2004). Bien que l'on ne puisse totalement écarter la possibilité que certains chercheurs prennent des décisions inhérentes aux violations des conditions d'application sans le mentionner dans leur article, l'hypothèse privilégiée par Keselman et al. (1998) est que la majorité des chercheurs applique des tests paramétriques indépendamment du fait que leurs conditions soient ou non respectées. Cette hypothèse semble confirmée par une expérience de Hoekstra et al. (2012): afin d'étudier les pratiques des chercheurs lorsqu'ils étaient confrontés à un scénario qui impliquait la réalisation d'un test *t*, d'une *ANOVA* ou d'une régression linéaire, ces chercheurs ont observé 30 doctorants qui travaillaient depuis au moins deux ans dans des départements de psychologie aux Pays-Bas et qui avaient dû pratiquer tous ces tests au moins une fois. Alors que *tous* ont opté pour un test paramétrique, les conditions d'application de ces tests n'ont été testées que dans un faible pourcentage de cas. Après l'expérience, les 30 doctorants ont été soumis à un questionnaire. Celui-ci a révélé que la non vérification des conditions d'application des tests était due à leur manque de familiarité avec les conditions d'application des tests, plutôt que par un choix délibéré de leur part. Il est à noter qu'en réalité, vérifier les conditions d'application des tests est bien plus complexe qu'il n'y paraît, et tout chercheur désireux d'améliorer la transparence dans la transmission des analyses de données resterait confronté à

3. Depuis l'article de Hoekstra et al. (2012), la septième édition du manuel des normes APA est parue. La mention explicite des conditions d'application ne fait pas partie des mises à jours proposées dans cette nouvelle édition.

un problème majeur: les conditions d'homogénéité des variances et de normalité reposent sur les paramètres de *population* et non sur les paramètres d'échantillon. Or, ces paramètres de population ne sont pas connus (Hoekstra et al., 2012). De plus, tenter d'inférer sur le respect des conditions d'application en utilisant les paramètres de l'échantillon à travers des tests d'hypothèse ne semble pas approprié, puisqu'il a été démontré que l'application d'un test conditionnellement aux résultats d'un test statistique préliminaire a pour effet d'augmenter l'erreur de type I (Schucany & Tony Ng, 2006). Tout ceci ne constituerait pas réellement un problème, en soi, si les test t de Student et F de Fisher étaient susceptibles de fournir des conclusions non biaisées et fiables même en cas d'écarts à ces conditions, or ce n'est malheureusement pas toujours le cas. Ces tests sont particulièrement sensibles aux violations de la condition d'homogénéité des variances, et cette sensibilité est accentuée lorsque les échantillons n'ont pas tous la même taille (Keselman et al., 1998).

Compte tenu de tous les éléments précités, il semblerait donc que la solution la plus viable serait d'utiliser des tests qui ne reposent pas sur les conditions de normalité et d'homogénéité des variances. Il existe, par exemple, des tests qui reposent sur la comparaison d'autres indicateurs de tendance centrale que la moyenne (comme la moyenne trimmée), mais ces derniers font très souvent face à une forte résistance de la part des chercheurs, qui persistent à vouloir comparer les moyennes (Erceg-Hurn & Mirosevich, 2008 ; Keselman et al., 1998 ; Wilcox, 1998). Dans la mesure où une revue approfondie de la littérature démontre que le non respect de la condition d'homogénéité des variances affecte bien plus le taux d'erreur de type I ainsi que la puissance de tests t de Student et F de Fisher (Erceg-Hurn & Mirosevich, 2008 ; Grissom, 2000 ; Hoekstra et al., 2012 ; Osborne & Waters, 2002) que le non respect de la condition de normalité, nous recommandons aux psychologues de remplacer les tests t de Student et F de Fisher par le test de Welch, un test de comparaison de moyennes qui ne requiert pas la condition d'homogénéité des variances. Cette solution a été suggérée par de nombreux auteurs avant nous (voir, par exemple Rasch, Kubinger, & Moder, 2011 ; Ruxton, 2006 ; Zimmerman, 2004), pourtant, cela ne semble pas avoir eu d'impact sur les pratiques des chercheurs en psychologie. Afin de tenter de changer leurs pratiques, nous nous sommes particulièrement appliqués, au sein des articles présentés dans les chapitres 2 à 3, à nous adresser directement à ce public de chercheurs. Pour ce faire, nous avons tenté (1) d'expliquer concrètement pourquoi la condition d'homoscédasticité n'est pas réaliste, en nous appuyant sur des exemples directement issus de la recherche en psychologie, (2) de définir certaines notions statistiques de la manière la plus simple possible, en limitant les explications mathématiques et (3) d'illustrer graphiquement l'impact des violations de la condition d'homoscédasticité, plutôt que de fournir des tableaux de chiffres lourds et complexes. De plus, nous avons conclu ces articles par des recommandations concrètes, afin d'aider les chercheurs à extraire le

message clé de ces articles. Ajoutons que les deux articles ont été soumis et publiés dans une revue Open Access (l'*International Review of Social Psychology*) afin d'assurer la diffusion la plus large possible de notre message.

1.0.1 Limite 2: hypothèse nulle

Effect sizes are an important outcome of empirical research. Moving beyond decisions about statistical significance, there is a strong call for researchers to report and interpret effect sizes and associated confidence intervals. This practice is highly endorsed by the American Psychological Association (APA) and the American Educational Research Association (American Educational Research Association, 2006; American Psychological Association, 2010).

En parlant des tailles d'effets, on commence de plus en plus à les utiliser (j'ai une réf qui le dit) mais: - on les calcule sans vraiment les comprendre/interpréter - comme pour le test t de Student et l'ANOVA, on utilise un test qui dépend des mm conditions d'application. Utiliser des tests plus adéquats permettrait d'améliorer les pratiques et à termes, de déterminer des mesures de taille d'effets qui pourront être utilisées a priori dans des tests plus informatifs que ceux visant à détecter l'absence d'effet (cf. tests d'équivalence).

Un paragraphe relatif à la taille d'effet. EN EXPLORATOIRE, ce qui à termes pourrait servir à définir des hypothèses plus informatives pour d'autres chercheurs, qui pourraient être utilisées, soit dans des tests d'effets minimaux, soit pour des tests d'équivalence. Et that's it.

Rem.: "une violation des conditions d'application peut amener à une sous- ou sur-estimation des mesure de taille d'effet (Osborne & Waters, 2002, cités par Hoekstra!)

Le NHST fait l'objet d'énormément de critiques, si bien que certains recommandent de le remplacer par une mesure de taille d'effet accompagnée d'un intervalle de confiance autour de la taille d'effet. Le raisonnement est que si l'IC contient la valeur 0, on ne peut conclure à une différence significative (Counsell & Harlow, 2017).

Une des principales critiques des tests d'hypothèse est le fait que l'on compare la différence observée à l'absence totale de différence (= un effet de 0). C'est une question qui est peu intéressante, car peu surprenante. Mais pourquoi comparer à 0 et pas à une autre valeur?

D'après Lakens (2021), un test d'hypothèse (selon l'approche de Neyman-Pearson) vaut la peine à 2 conditions:

1) que l'hypothèse nulle soit assez plausible pour que son rejet puisse surprendre au moins certains;

2) le chercheur veut appliquer une procédure méthodol qui l'autorise à prendre des décisions quant à la manière d'agir, tout en contrôlant le taux d'erreur. Agir peut vouloir dire: adopter un traitement, une politique, une intervention, ou abandonner un domaine de recherche, modifier une manipulation, ou de faire un certain type de déclaration ou revendication.

Counsell & Harlow (2017): *the constant calls for reporting effect sizes appears to have had an effect on the Canadian psychology articles as just over 90% of the analyses that used a significance test also included a standardized or unstandardized effect size. Few articles presented an effect size without hypothesis testing, and few of the analyses' results included a CI.*

Ca se fait apparemment de plus en plus de reporter la taille d'effet (dans leur analyse de 151 études, 90% des analyses incluait une mesure de taille d'effet, standardisée ou non... mais très peu incluait les IC et de plus, ils les donnaient mais sans vraiment en discuter... @Counsell & Harlow (2017) dans la discussion).

Comme déjà mentionné, l'hypothèse nulle est l'absence d'effet. On en reste sur la nil-hypothesis. Du coup, un effet significatif n'a pas vraiment de valeur. En réponse à ce problème, on a écrit deux articles:

- On peut commencer par ajouter une information sur les tailles d'effets (mais du coup ça n'oblige pas à réfléchir à l'avance à l'effet qui nous intéresse)

Dans la revue de Keselman et al. (1998), ils mentionnent que les tailles d'effet ne sont pratiquement jamais reportées malgré les recommandations du manuel de l'APA (1994) (et qu'elles ne sont fournies qu'en cas d'effet significatif).

- On peut aussi faire des tests plus informatifs (tests d'équivalence et/ou tests d'effets minimaux).

*One of the most widely suggested improvements of the use of p values is to replace null-hypothesis tests (where the goal is to reject an effect of exactly 0) with tests of range predictions (where the goal is to reject effects that fall outside of the range of effects that is predicted or considered practically important) (Lakens, 2021).

1.0.2 Pourquoi jusque là la sauce n'a pas pris?

Je suis loin d'être la première à signaler tt ça. Ce qui manque encore dans mon plan d'introduction, c'est que je dois encore trouver le moyen de montrer en quoi mes articles sont une plus-value, ce qu'ils apportent. 2) Parler des packages, des applications Shiny, etc.

PARLER DES DIFFERENTES REVUES DE LITTERATURE QUI LE DISENT.

Qu'est-ce qui pourrait expliquer cela? 1) Sharpe (2013): lack of awareness (p.573) Manque de conscience des développements dans le domaine?

2) Sharpe (2013): journal editors (p.573) Les éditeurs ne poussent pas assez? -> Pas convaincue que ça m'intéresse

3) Sharpe (2013): Publish or perish? (p.574) je ne comprends mm pas en quoi c'est un argument

4) Sharpe (2013): Software (p.574) -> aaahh! Certaines pratiques comme les équations structurelles et les analyses de puissance ont été facilitées par des logiciels comme gpower. Cela explique leur popularité. En ce qui concerne les statistiques plus robustes, par contre, elles ont moins de succès car non disponibles dans les logiciels disponibles. Les gens veulent juste qu'on leur dise où cliquer pour avoir le test qu'ils veulent! C'est triste mais faut faire avec (à mon avis).

5) Sharpe (2013): inadequate education (p.574)

6) Sharpe (2013): mindset: facteurs psychologiques t.q. la peur de dévier des pratiques courantes (comme si on n'allait pas être publié si on ne faisait pas comme tout le monde).

Anecdote: les chercheurs font souvent l'erreur de croire qu'il faut vérifier la normalité de la VD en faisant une régression. Dans SPSS, il est assez complexe de le faire car il faut d'abord calculer les résidus, ce qui implique de comprendre que les tests t et ANOVA sont des cas particuliers de régression, puis ensuite a posteriori représenter graphiquement les résidus. C'est chronophage et complexe. Dans Jamovi, par contre, la vérification de la normalité des résidus est automatiquement réalisée lorsqu'on fait un test t. Le rôle des méthodologistes, à mon sens, est de prémâcher le travail, pour permettre à d'autres de créer des outils conçus pour améliorer les pratiques de recherche. À partir du moment où c'est automatiquement fait correctement, il devient moins problématique que les psychologues maîtrisent le détail. Débarassés de ces questions, ils pourront peut-être alors plus se focaliser sur l'important pour mieux comprendre et interpréter les résultats de leur tests: c'est-à-dire comprendre la distribution d'échantillonnage, dont pratiquement tout découle.

2 Chapitre 2: le test t de Welch (article 1)

RESEARCH ARTICLE

Why Psychologists Should by Default Use Welch's t -test Instead of Student's t -test

Marie Delacre*, Daniël Lakens† and Christophe Leys*

When comparing two independent groups, psychology researchers commonly use Student's t -tests. Assumptions of normality and homogeneity of variance underlie this test. More often than not, when these conditions are not met, Student's t -test can be severely biased and lead to invalid statistical inferences. Moreover, we argue that the assumption of equal variances will seldom hold in psychological research, and choosing between Student's t -test and Welch's t -test based on the outcomes of a test of the equality of variances often fails to provide an appropriate answer. We show that the Welch's t -test provides a better control of Type 1 error rates when the assumption of homogeneity of variance is not met, and it loses little robustness compared to Student's t -test when the assumptions are met. We argue that Welch's t -test should be used as a default strategy.

Keywords: Welch's t -test; Student's t -test; homogeneity of variance; Levene's test; Homoscedasticity; statistical power; type 1 error; type 2 error

Independent sample t -tests are commonly used in the psychological literature to statistically test differences between means. There are different types of t -tests, such as Student's t -test, Welch's t -test, Yuen's t -test, and a bootstrapped t -test. These variations differ in the underlying assumptions about whether data is normally distributed and whether variances in both groups are equal (see, e.g., Rasch, Kubinger, & Moder, 2011; Yuen, 1974). Student's t -test is the default method to compare two groups in psychology. The alternatives that are available are considerably less often reported. This is surprising, since Welch's t -test is often the preferred choice and is available in practically all statistical software packages.

In this article, we will review the differences between Welch's t -test, Student's t -test, and Yuen's t -test, and we suggest that Welch's t -test is a better default for the social sciences than Student's and Yuen's t -tests. We do not include the bootstrapped t -test because it is known to fail in specific situations, such as when there are unequal sample sizes and standard deviations differ moderately (Hayes & Cai, 2007).

When performing a t -test, several software packages (i.e., R and Minitab) present Welch's t -test by default. Users can request Student's t -test, but only after explicitly stating that the assumption of equal variances is

met. Student's t -test is a parametric test, which means it relies on assumptions about the data that are analyzed. Parametric tests are believed to be more powerful than non-parametric tests (i.e., tests that do not require assumptions about the population parameters; Sheskin, 2003). However, Student's t -test is generally only more powerful when the data are normally distributed (the assumption of normality) and the variances are equal in both groups (homoscedasticity; the assumption of homogeneity of variance; Carroll & Schneider, 1985; Erceg-Hurn & Mirosevich, 2008).

When sample sizes are equal between groups, Student's t -test is robust to violations of the assumption of equal variances as long as sample sizes are big enough to allow correct estimates of both means and standard deviations (i.e., $n \geq 5$),¹ except when distributions underlying the data have very high skewness and kurtosis, such as a chi-square distribution with 2 degrees of freedom. However, if variances are *not* equal across groups and the sample sizes differ across independent groups, Student's t -test can be severely biased and lead to invalid statistical inferences (Erceg-Hurn & Mirosevich, 2008).^{2,3} Here, we argue that there are no strong reasons to assume equal variances in the psychological literature by default nor substantial costs in abandoning this assumption.

In this article, we will first discuss why we need a default test and why a two-step procedure where researchers decide whether or not to use Welch's t -test based on a check of the assumption of normality and equal variances is undesirable. Then, we will discuss whether the assumption of equal variances is plausible in psychology and point out research areas where this assumption is implausible.

* Université Libre de Bruxelles, Service of Analysis of the Data (SAD), Bruxelles, BE

† Eindhoven University of Technology, Human Technology Interaction Group, Eindhoven, NL

Corresponding author: Marie Delacre (marie.delacre@ulb.ac.be)

We will then review differences between Student's *t*-test, Welch's *t*-test, and Yuen's *t*-test and show through simulations that the bias in Type 1 error rates when Yuen's *t*-test is used is often severely inflated (above 0.075, which is "critical inflation", following Bradley, 1978) and that the bias in Type 1 error rates when Student's *t*-test is used has a larger impact on statistical inferences than the rather modest impact on the Type 2 error rate of always using Welch's *t*-test by default. Given our analysis and the availability of Welch's *t*-test in all statistical software, we recommend a procedure where Welch's *t*-test is used by default when sample sizes are unequal.

Limitations of Two-Step Procedures

Readers may have learned that the assumptions of normality and of equal variances (or the homoscedasticity assumption) must be examined using assumption checks prior to performing any *t*-test. When data are not normally distributed, with small sample sizes, alternatives should be used. Classic nonparametric statistics are well-known, such as the Mann-Whitney U-test and Kruskal-Wallis. However, unlike a *t*-test, tests based on rank assume that the distributions are the same between groups. Any departure to this assumption, such as unequal variances, will therefore lead to the rejection of the assumption of equal distributions (Zimmerman, 2000). Alternatives exist, known as the "modern robust statistics" (Wilcox, Granger, & Clark, 2013). For example, data sets with low kurtosis (i.e., a distribution flatter than the normal distribution) should be analyzed with the two-sample trimmed *t*-test for unequal population variances, also called Yuen's *t*-test (Luh & Guo, 2007; Yuen, 1974). However, analyses in a later section will show that the normality assumption is not very important for Welch's *t*-test and that there are good reasons to, in general, prefer Welch's *t*-test over Yuen's *t*-test.

With respect to the assumption of homogeneity of variance, if the test of the equality of variance is non-significant and the assumption of equal variances cannot be rejected, homoscedastic methods such as the Student's *t*-test should be used (Wilcox et al., 2013). If the test of the equality of variances is significant, Welch's *t*-test should be used instead of Student's *t*-test because the assumption of equal variances is violated. However, testing the equality of variances before deciding which *t*-test is performed is problematic for several reasons, which will be explained after having described some of the most widely used tests of equality of variances.

Different Ways to Test for Equal Variances

Researchers have proposed several tests for the assumption of equal variances. Levene's test and the F-ratio test are the most likely to be used by researchers because they are available in popular statistical software (Hayes & Cai, 2007). Levene's test is the default option in SPSS. Levene's test is the One-Way ANOVA computed on the terms $|x_{ij} - \hat{\theta}_j|$, where x_{ij} is the *i*th observation in the *j*th group, and $\hat{\theta}_j$ is the "center" of the distribution for the *j*th group (Carroll & Schneider, 1985). In R, the "center" is by default the median, which is also called "Brown Forsythe test for equal variances". In SPSS, the "center" is by default the mean

(which is the most powerful choice when the underlying data are symmetrical).⁴ The F-ratio statistic is obtained by computing $SD2/SD1$ (standard deviation ratio, SDR). A generalization of the F-ratio test, to be used when there are more than two groups to compare, is known as the Bartlett's test.

The F-ratio test and the Bartlett test are powerful, but they are only valid under the assumption of normality and collapse as soon as one deviates even slightly from the normal distribution. They are therefore not recommended (Rakotomalala, 2008).

Levene's test is more robust than Bartlett's test and the F-ratio test, but there are three arguments against the use of Levene's test. First, there are several ways to compute Levene's test (i.e., using the median or mean as center), and the best version of the test for equal variances depends on how symmetrically the data is distributed, which is itself difficult to statistically quantify.

Second, performing two tests (Levene's test followed by a *t*-test) on the same data makes the alpha level and power of the *t*-test dependent upon the outcome of Levene's test. When we perform Student's or Welch's *t*-test conditionally on a significant Levene's test, the long-run Type 1 and Type 2 error rates will depend on the power of Levene's test. When the power of Levene's test is low, the error rates of the conditional choice will be very close to Student's error rates (because the probability of choosing Student's *t*-test is very high). On the other hand, when the power of Levene's test is very high, the error rates of the conditional choice will be very close to Welch's error rate (because the probability of choosing Welch's *t*-test is very high; see Rasch, Kubinger, & Moder, 2011). When the power of Levene's test is medium, the error rates of the conditional choice will be somewhere between Student's and Welch's error rates (see, e.g., Zimmerman, 2004). This is problematic when the test most often performed actually has incorrect error rates.

Third, and relatedly, Levene's test can have very low power, which leads to Type 2 errors when sample sizes are small and unequal (Nordstokke & Zumbo, 2007). As an illustration, to estimate the power of Levene's test, we simulated 1,000,000 simulations with balanced designs of different sample sizes (ranging from 10 to 80 in each condition, with a step of 5) under three SDR where the true variances are unequal, respectively, 1.1, 1.5, and 2, yielding 45,000,000 simulations in total. When $SDR = 1$, the equal variances assumption is true when $SDR > 1$ the standard deviation of the second sample is bigger than the standard deviation of the first sample and when $SDR < 1$ the standard deviation of the second sample is smaller than the standard deviation of the first sample. We ran Levene's test centered around the mean and Levene's test centered around the median and estimated the power (in %) to detect unequal variances with equal sample sizes (giving the best achievable power for a given total *N*; see Figure 1).⁵

As we can see in the graph, the further SDR is from 1, the smaller the sample size needed to detect a statistically significant difference in the SDR. Furthermore, for each SDR, power curves of the Levene's test based on the mean

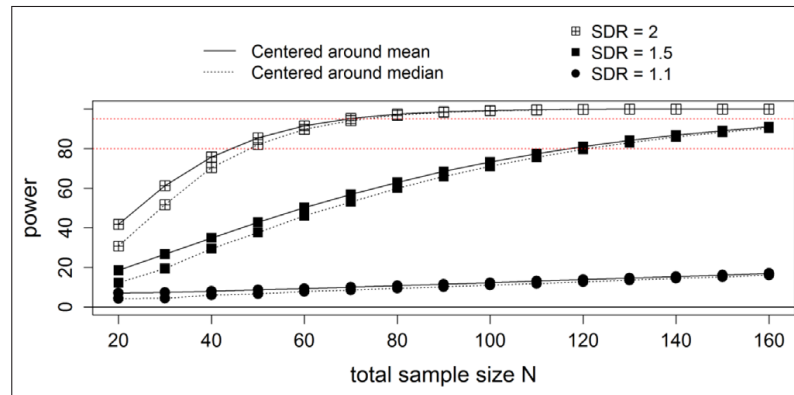


Figure 1: Estimated power of Levene's test as a function of sample size, SDR and centering parameter.

are slightly above power curves of the Levene's test based on the median, meaning that it leads to slightly higher power than Levene's test based on the median. This can be due to the fact that data is extracted from normal distributions. With asymmetric data, the median would perform better. When $SDR = 2$, approximately 50 subjects are needed to have 80 percent power to detect differences, while approximately 70 subjects are needed to have 95 percent power to detect differences (for both versions of Levene's test). To detect an SDR of 1.5 with Levene's test, approximately 120 subjects are needed to reach a power of 0.80 and about 160 to reach a power of 0.95. Since such an SDR is already very problematic in terms of the type 1 error rate for the Student's *t*-test (Bradley, 1978), needing such a large sample size to detect it is a serious hurdle. This issue becomes even worse for lower SDR , since an SDR as small as 1.1 already calls for the use of Welch's *t*-test (See table A3.1 to A3.9 in the additional file). Detecting such a small SDR calls for a huge sample size (a sample size of 160 provides a power rate of 0.16).

Since Welch's *t*-test has practically the same power as Student's *t*-test, even when $SDR = 1$, as explained using simulations later, we should seriously consider using Welch's *t*-test by default.

The problems in using a two-step procedure (first testing for equality of variances, then deciding upon which test to use) have already been discussed in the field of statistics (see e.g., Rasch, Kubinger, & Moder, 2011; Ruxton, 2006; Wilcox, Granger, & Clark, 2013; Zimmerman, 2004), but these insights have not changed the current practices in psychology, as of yet. More importantly, researchers do not even seem to take the assumptions of Student's *t*-test into consideration before performing the test, or at least rarely discuss assumption checks.

We surveyed statistical tests reported in the journal *SPSS* (*Social Psychological and Personality Science*) between April 2015 and April 2016. From the total of 282 studies, 97 used a *t*-test (34.4%), and the homogeneity of variance was explicitly discussed in only 2 of them. Moreover, based on the reported degrees of freedom in the results section, it seems that Student's *t*-test is used most often and that alternatives are considerably less popular. For 7 studies,

there were decimals in the values of the degrees of freedom, which suggests Welch's *t*-test might have been used, although the use of Welch's *t*-test might be higher but not identifiable because some statisticians recommend rounding the degrees of freedom to round numbers.

To explain this lack of attention to assumption checks, some authors have argued that researchers might have a lack of knowledge (or a misunderstanding) of the parametric assumptions and consequences of their violations or that they might not know how to check assumptions or what to do when assumptions are violated (Hoekstra, Kiers, & Johnson, 2012).⁶ Finally, many researchers don't even know there are options other than the Student's *t*-test for comparing two groups (Erceg-Hurn & Miroseovich, 2008). How problematic this is depends on how plausible the assumption of equal variances is in psychological research. We will discuss circumstances under which the equality of variances assumption is especially improbable and provide real-life examples where the assumption of equal variances is violated.

Homogeneity of Variance Assumptions

The homogeneity of variances assumption is rarely true in real life and cannot be taken for granted when performing a statistical test (Erceg-Hurn & Miroseovich, 2008; Zumbo & Coulombe, 1997). Many authors have examined real data and noted that SDR is often different from the 1:1 ratio (see, e.g., Grissom, 2000; Erceg-Hurn & Miroseovich, 2008). This shows that the presence of unequal variances is a realistic assumption in psychological research.⁷ We will discuss three different origins of unequal standard deviations across two groups of observations.

A first reason for unequal variances across groups is that psychologists often use *measured variables* (such as age, gender, educational level, ethnic origin, depression level, etc.) instead of random assignment to condition. In their review of comparing psychological findings from all fields of the behavioral sciences across cultures, Henrich, Heine, and Norenzayan (2010) suggest that parameters vary largely from one population to another. In other words, variance is not systematically the same in every pre-existing group. For example, Feingold (1992) has

shown that intellectual abilities of males were more variable than intellectual abilities of females when looking at several standardized test batteries measuring general knowledge, mechanical reasoning, spatial visualization, quantitative ability, and spelling. Indeed, the variability hypothesis (that men demonstrate greater variability than women) is more than a century old (for a review, see Shields, 1975). In many research domains, such as mathematics performance, there are strong indicators that variances ratios differ between 1.1 and 1.2, although variances ratios do not differ in all countries, and the causes for these differences are not yet clear. Nevertheless, it is an empirical fact that variances ratios can differ among pre-existing groups.

Furthermore, some pre-existing groups have different variability by definition. An example from the field of education is the comparison of selective school systems (where students are accepted on the basis of selection criterions) versus comprehensive school systems (where all students are accepted, whatever their aptitudes; see, e.g., Hanushek & Wößmann, 2006). At the moment that a school accepts its students, variability in terms of aptitude will be greater in a comprehensive school than in a selective school, by definition.

Finally, a quasi-experimental treatment can have a different impact on variances between groups. Hanushek and Wößmann (2006) suggest that there is an impact of the educational system on variability in achievement. Even if variability, in terms of aptitude, is greater in a comprehensive school than in a selective school at first, a selective school system at primary school increases inequality (and then variability) in achievement in secondary school. Another example is variability in moods. Cowdry, Gardner, O'Leary, Leibenluft, & Rubinow (1991) noted that intra-individual variability is larger in patients suffering from premenstrual syndrome (PMS) than in normal patients and larger in normal patients than in depressive patients. Researchers studying the impact of an experimental treatment on mood changes can expect a bigger variability of mood changes in patients with PMS than in normal or depressive patients and thus a higher standard deviation in mood measurements.

A second reason for unequal variances across groups is that while variances of two groups are the same when group assignment is completely randomized, deviation from equality of variances can occur later, as a *consequence of an experimental treatment* (Cumming, 2013; Erceg-Hurn & Mirosevich, 2008; Keppel, 1991). For example, psychotherapy for depression can increase the variability in depressive symptoms, in comparison with a control group, because the effectiveness of the therapy will depend on individual differences (Bryk & Raudenbush, 1988; Erceg-Hurn & Mirosevich, 2008). Similarly, Kester (1969) compared the IQs of students from a control group with the IQs of students when high expectancies about students were induced in the teacher. While no effect of teacher expectancy on IQ was found, the variance was bigger in the treatment group than in the control group (56.52 vs. 32.59, that is, $SDR \approx 1.32$). As proposed by Bryk

and Raudenbush (1988), this can result from the interaction between the treatment and the students' reactions: students can react differently to the induced expectations. More generally, whenever a manipulation has individual moderators, variability should increase compared to a control condition.

Knowing whether standard deviations differ across conditions is important information, but in many fields, we have no accurate estimates of the standard deviation in the population. Whereas we collect population effect sizes in meta-analyses, these meta-analyses often do not include the standard deviations from the literature. As a consequence, we regrettably do not have easy access to aggregated information about standard deviations across research areas, despite the importance of this information. It would be useful if meta-analysts start to code information about standard deviations when performing meta-analyses (Lakens, Hilgard, & Staaks, 2016), such that we can accurately quantify whether standard deviations differ between groups, and how large the SDR is.

The Mathematical Differences Between Student's *t*-test, Welch's *t*-test, and Yuen's *t*-test

So far, we have simply mentioned that Welch's *t*-test differs from Student's *t*-test in that it does not rely on the equality of variances assumption. In this section, we will explain why this is the case. The Student's *t* statistic is calculated by dividing the mean difference between group $\bar{x}_1 - \bar{x}_2$ by a pooled error term, where s_1^2 and s_2^2 are variance estimates from each independent group, and where n_1 and n_2 are the respective sample sizes for each independent group (Student, 1908):

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \right) * \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \quad (1)$$

The degrees of freedom are computed as follows (Student, 1908):

$$df = n_1 + n_2 - 2 \quad (2)$$

Student's *t*-test is calculated based on a *pooled* error term, which implies that both samples' variances are estimates of a common population variance. Whenever the variances of the two normal distributions are not similar and the sample sizes in each group are not equal, Student's *t*-test results are biased (Zimmerman, 1996). The more unbalanced the distribution of participants across both independent groups, the more Student's *t*-test is based on the incorrect standard error (Wilcox et al., 2013) and, consequently, the less accurate the computation of the *p*-value will be.

When the larger variance is associated with the *larger* sample size, there is a decrease in the nominal Type 1 error rate (Nimon, 2012; Overall, Atlas, & Gibson, 1995). The reason for this is that the error term increases, and,

as a consequence, the Student's *t*-value decreases, leading to fewer significant findings than expected with a specific alpha level. When the larger variance is associated with the *smaller* sample size, the Type 1 error rate is inflated (Nimon, 2012; Overall, Atlas, & Gibson, 1995). This inflation is caused by the under-evaluation of the error term, which increases Student's *t* value and thus leads to more significant results than are expected based on the alpha level.

As discussed earlier, Student's *t*-test is robust to unequal variances as long as the sample sizes of each group are similar (Nimon, 2012; Ruxton, 2006; Wallenstein, Zucker, & Fleiss, 1980), but, in practice, researchers often have different sample sizes in each of the independent groups (Ruxton, 2006). Unequal sample sizes are particularly common when examining measured variables, where it is not always possible to determine *a priori* how many of the collected subjects will fall in each category (e.g., sex, nationality, or marital status). However, even with complete randomized assignment to conditions, where the same number of subjects are assigned to each condition, unequal sample sizes can emerge when participants have to be removed from the data analysis due to being outliers because the experimental protocol was not followed when collecting the data (Shaw & Mitchell-Olds, 1993) or due to missing values (Wang et al., 2012).

Previous work by many researchers has shown that Student's *t*-test performs surprisingly poorly when variances are unequal and sample sizes are unequal (Glass, Peckham, & Sanders, 1972; Overall, Atlas, & Gibson, 1995; Zimmerman, 1996), especially with small sample sizes and low alpha levels (e.g., alpha = 1%; Zimmerman, 1996). The poor performance of Student's *t*-test when variances are unequal becomes visible when we look at the error rates of the test and the influence of both Type 1 errors and Type 2 errors. An increase in the Type 1 error rate leads to an inflation of the number of false positives in the literature, while an increase in the Type 2 error rate leads to a loss of statistical power (Banerjee et al., 2009).

To address these limitations of Student's *t*-test, Welch (1947) proposed a separate-variances *t*-test computed by dividing the mean difference between group $\bar{x}_1 - \bar{x}_2$ by an unpooled error term, where s_1^2 and s_2^2 are variance estimates from each independent group, and where n_1 and n_2 are the respective sample sizes for each independent group:⁸

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad (3)$$

The degrees of freedom are computed as follows:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^2}{n_1}\right)^2 + \left(\frac{s_2^2}{n_2}\right)^2} \quad (4)$$

When both variances and sample sizes are the same in each independent group, the *t*-values, degrees of freedom, and the *p*-values in Student's *t*-test and Welch's *t*-test are the same (see **Table 1**). When the variance is the same in both independent groups but the sample sizes differ, the *t*-value remains identical, but the degrees of freedom differ (and, as a consequence, the *p*-value differs). Similarly, when the variances differ between independent groups but the sample sizes in each group are the same, the *t*-value is identical in both tests, but the degrees of freedom differ (and, thus, the *p*-value differs). The most important difference between Student's *t*-test and Welch's *t*-test, and indeed the main reason Welch's *t*-test was developed, is when both the variances and the sample sizes differ between groups, the *t*-value, degrees of freedom, and *p*-value all differ between Student's *t*-test and Welch's *t*-test. Note that, in practice, samples practically never show exactly the same pattern of variance as populations, especially with small sample sizes (Baguley, 2012; also see table A2 in the additional file).

Yuen's *t*-test, also called "20 percent trimmed means test", is an extension of Welch's *t*-test and is allegedly more robust in case of non-normal distributions (Wilcox & Keselman, 2003). Yuen's *t*-test consists of removing the lowest and highest 20 percent of the data and applying Welch's *t*-test on the remaining values. The procedure is explained and well-illustrated in a paper by Erceg-Hurn and Mirosevich (2008).

Simulations: Error Rates for Student's *t*-test versus Welch's *t*-test

When we are working with a balanced design, the statistical power (the probability of finding a significant effect, when there is a true effect in the population, or 1 minus the Type 2 error rate) is very similar for Student's *t*-test

	Equal variances	Unequal variances
Balanced design	$t_{\text{Welch}} = t_{\text{Student}}$ $df_{\text{Welch}} = df_{\text{Student}}$ $p_{\text{Welch}} = p_{\text{Student}}$	$t_{\text{Welch}} = t_{\text{Student}}$ $df_{\text{Welch}} \neq df_{\text{Student}}$ $p_{\text{Welch}} \neq p_{\text{Student}}$
Unbalanced design	$t_{\text{Welch}} = t_{\text{Student}}$ $df_{\text{Welch}} \neq df_{\text{Student}}$ $p_{\text{Welch}} \neq p_{\text{Student}}$	$t_{\text{Welch}} \neq t_{\text{Student}}$ $df_{\text{Welch}} \neq df_{\text{Student}}$ $p_{\text{Welch}} \neq p_{\text{Student}}$

Table 1: Comparison of *t*-value and Degrees of Freedom of Welch's and Student's *t*-test.

and Welch's *t*-test. Even with extremely large SDR (respectively, 0.01, 0.1, 10, and 100) and small sample sizes (10 subjects per group), the biggest increase in power of Student's *t*-test compared to Welch's *t*-test is approximately 5 percent when the test is applied on two normal skewed distributions with unequal shapes. In all other cases, the difference in power between both tests is smaller (See table A1.1 to A1.9 in the additional file).

Considering the cases where sample sizes are unequal and SDR = 1, Student's *t*-test is sometimes better than Welch's *t*-test, and sometimes the reverse is true. The difference is small, except in three scenarios (See table A5.2, A5.5, and A5.6 in the additional file). However, because there is no correct test to perform that assures SDR = 1, and because variances are likely not to be equal in certain research areas, our recommendation is to always use Welch's *t*-test instead of Student's *t*-test.

To illustrate the differences in Type 1 error rates between Student's *t*-test and Welch's *t*-test, we simulated 1,000,000 studies under the null hypothesis (no difference between

the means in each group) under four scenarios. We chose a small sample ratio ($n_1 = 40$ vs. $n_2 = 60$) to show that when the equal variances assumption was not met and SDR = 2, biased error rates are observed in Student's *t*-test. We compared Scenario 1, where the variance is the same in each group (SDR = 1; homoscedasticity assumption met) and sample sizes are unequal (See **v 2a**), with Scenario 2, where the variance differs between groups (SDR = 2) but sample sizes are equal ($n_1 = n_2 = 50$; see **Figure 2b**). Furthermore, we simulated Scenario 3, where both sample sizes and variances were unequal between groups and the larger variance is associated with the larger sample size (SDR = 2; see **Figure 2c**), and a similar Scenario 4, where the larger variance is associated with the smaller sample size (SDR = 0.5; see **Figure 2d**). *P*-value distributions for both Student's and Welch's *t*-tests were then plotted. When there is no true effect, *p*-values are distributed uniformly.

As long as the variances are equal between groups or sample sizes are equal, the distribution of Student's *p*-values is uniform, as expected (see **Figures 2a** and **2b**),

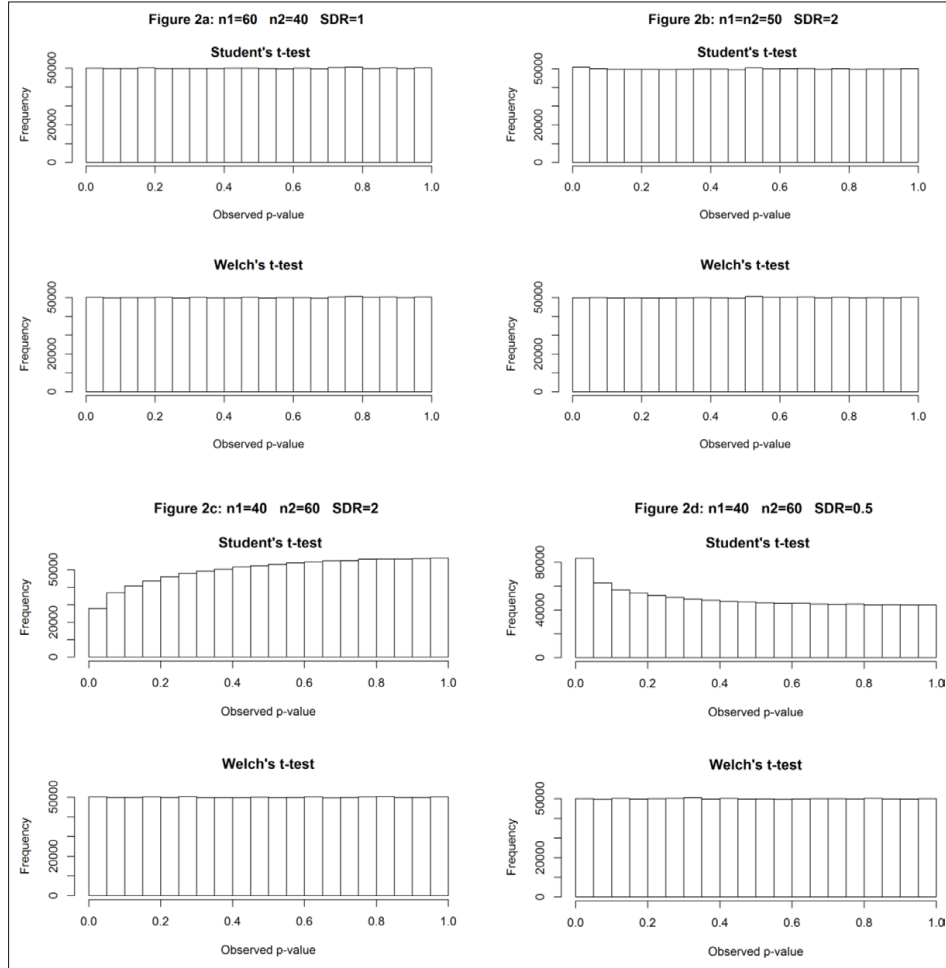


Figure 2: *P*-value distributions for Student's and Welch's *t*-test under the null as a function of SDR, and sample size.

which implies that the probability of rejecting a true null hypothesis equals the alpha level for any value of alpha. On the other hand, when the larger variance is associated with the larger sample size, the frequency of *p*-values less than 5 percent decreases to 0.028 (see **Figure 2c**), and when the larger variance is associated with the smaller sample size, the frequency of *p*-values less than 5 percent increases to 0.083 (see **Figure 2d**). Welch's *t*-test has a more stable Type 1 error rate (see Keselman et al., 1998; Keselman, Othman, Wilcox, & Fradette, 2004; Moser & Stevens, 1992; Zimmerman, 2004). Additional simulations, presented in the additional file, show that these scenarios are similar for several shapes of distributions (see tables A3.1 to A3.9 and table A4 in the additional file).

Moreover, as discussed previously, with very small SDRs, Welch's *t*-test still has a better control of Type 1 error rates than Student's *t*-test, even if neither of them give critical values (i.e., values under 0.025 or above 0.075, according to the definition of Bradley, 1975). With SDR = 1.1, when the larger variance is associated with the larger sample size, the frequency of Student's *p*-value being less than 5 percent decreases to 0.046, and when the larger variance is associated with the smaller sample size, the frequency of Student's *p*-value being less than 5 percent increases to

0.054. On the other side, the frequency of Welch's *p*-values being below 0.05 is exactly 5 percent in both cases.

Yuen's *t*-test is not a good unconditional alternative because we observe an unacceptable departure from the nominal alpha risk of 5 percent for several shapes of distributions (see tables A3.1, A3.4, A3.7, A3.8, and A3.9 in the additional file), particularly when we are studying asymmetric distributions of unequal shapes (see tables A3.8 and A3.9 in the additional file). Moreover, even when Yuen's Type 1 error does not show a critical departure from the nominal alpha risk (i.e., values above 0.075), Welch's *t*-test more accurately controls the Type 1 error rate (see tables A3.2, A3.3, A3.5, and A3.6 in the additional file). The Type 1 error rate of Welch's *t*-test remains closer to the nominal size (i.e., 5%) in all the previously discussed cases and also performs better with very extreme SDRs and unbalanced designs, as long as there are at least 10 subjects per groups (See table A4 in the additional file).

In **Figure 3**, *p*-values from Welch's *t*-test and Student's *t*-test tests, shown separately in **Figure 2** (through histograms), are now plotted against each other. **Figure 3a** shows Student's *p*-values plotted against Welch's *p*-values of Scenario 1, where the variance is the same in each group (SDR = 1) and sample sizes are unequal. **Figure 3b** displays Student's *p*-values plotted against Welch's *p*-values of Scenario 2, where the

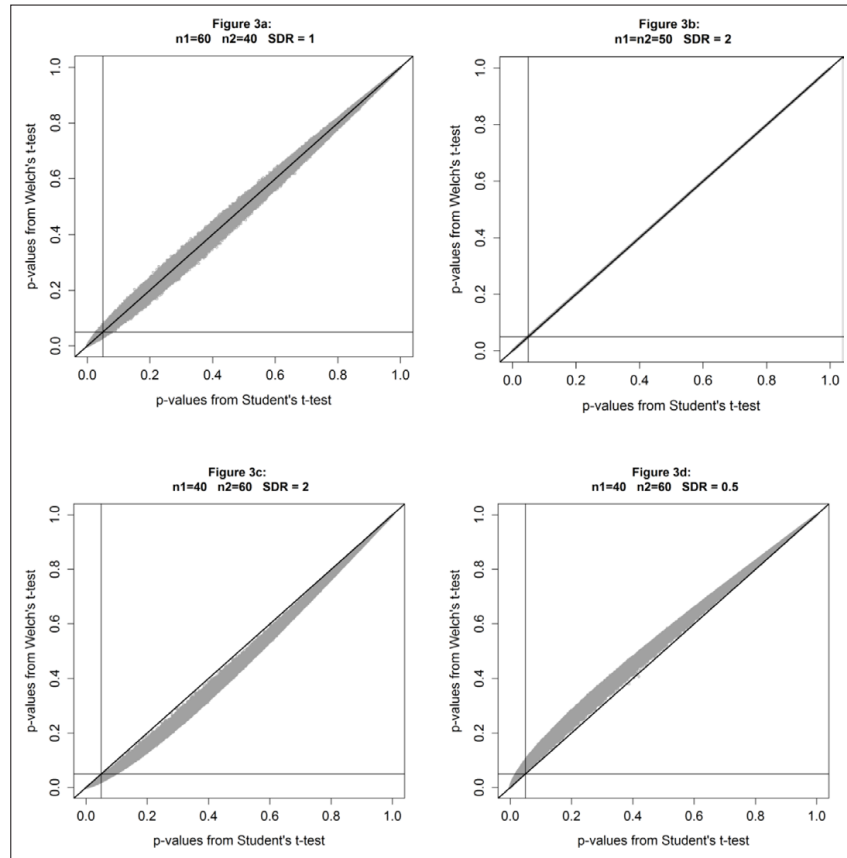


Figure 3: *P*-values from Student's *t*-test against *p*-values from Welch's *t*-test under the null.

variance differs between groups ($\text{SDR} = 2$) but sample sizes are equal ($n_1 = n_2 = 50$). **Figure 3c** shows Student's *p*-values plotted against Welch's *p*-values of Scenario 3, where both sample sizes and variances are unequal between groups and the larger variance is associated with the larger sample size ($\text{SDR} = 2$). And, finally, **figure 3d** plots Student's *p*-values against Welch's *p*-values of Scenario 4, where the greater variance is associated with the smaller sample size ($\text{SDR} = 0.5$).

Dots are marked on the black diagonal line when both tests return the same *p*-value. The top left quadrant contains all *p*-values less than 0.05 according to a Student's *t*-test, but greater than 0.05 according to Welch's *t*-test. The bottom right quadrant reports all *p*-values less than 0.05 according to Welch's *t*-test, but greater than 0.05 according to Student's *t*-test. The larger the standard deviations ratio and the greater the sample sizes ratio, the larger the difference between *p*-values from Welch's *t*-test and Student's *t*-test.

Conclusion

When the assumption of equal variances is not met, Student's *t*-test yields unreliable results, while Welch's *t*-test controls Type 1 error rates as expected. The widely recommended two-step approach, where the assumption of equal variances is tested using Levene's test and, based on the outcome of this test, a choice of Student's *t*-test or Welch's *t*-test is made, should not be used. Because the statistical power for this test is often low, researchers will inappropriately choose Student's *t*-test instead of more robust alternatives. Furthermore, as we have argued, it is reasonable to assume that variances are unequal in many studies in psychology, either because measured variables are used (e.g., age, culture, gender) or because, after random assignment to conditions, variance is increased in the experimental condition compared to the control condition due to the experimental manipulation. As it is explained in the additional file, Yuen's *t*-test is not a better test than Welch's *t*-test, since it often suffers high departure from the alpha risk of 5 percent. Therefore, we argue that Welch's *t*-test should always be used instead of Student's *t*-test.

When using Welch's *t*-test, a very small loss in statistical power can occur, depending on the shape of the distributions. However, the Type 1 error rate is more stable when using Welch's *t*-test compared to Student's *t*-test, and Welch's *t*-test is less dependent on assumptions that cannot be easily tested. Welch's *t*-test is available in practically all statistical software packages (and already the default in R and Minitab) and is easy to use and report. We recommend that researchers make clear which test they use by specifying the analysis approach in the result section.

Convention is a weak justification for the current practice of using Student's *t*-test by default. Psychologists should pay more attention to the assumptions underlying the tests they perform. The default use of Welch's *t*-test is a straightforward way to improve statistical practice.

Notes

- ¹ There is a Type 1 error rate inflation in a few cases where sample sizes are extremely small and SDR is big (e.g., when $n_1 = n_2 = 3$ are sampled from uniform distributions and $\text{SDR} = 2$, the Type 1 error rate = 0.083;

or when $n_1 = 3$ is sampled from a uniform distribution and $n_2 = 3$ is sampled from a double exponential distribution). However, with extremely small sample sizes ($N \leq 5$), the estimate of means and standard deviations is extremely inaccurate anyway. As we mentioned in table A2 (see the additional file), the smaller the sample size, the further the average standard deviation is from the population standard deviation, and the larger the dispersion around this average.

- ² This is called the Behren-Fisher problem (Hayes & Cai, 2007).

- ³ In a simulation that explored Type 1 error rates, we varied the size of the first sample from 10 to 40 in steps of 10 and the sample sizes ratio and the standard deviation ratio from 0.5 to 2 in steps of 0.5, resulting in 64 simulations designs. Each design was tested 1,000,000 times. Considering these parameter values, we found that the alpha level can be inflated up to 0.11 or deflated down to 0.02 (see the additional file).

- ⁴ Other variants have been proposed, such as the percent trimmed mean (Lim & Loh, 1996).

- ⁵ Because sample sizes are equal for each pair of samples, which sample has the bigger standard deviation is not applicable. In this way, $\text{SDR} = X$ will return the same answer in terms of percent power of Levene's test as $\text{SDR} = 1/X$. For example, $\text{SDR} = 2$ will return the same answer as $\text{SDR} = 1/2 = 0.5$.

- ⁶ For example, many statistical users believe that the Mann-Whitney non-parametric test can cope with both normality and homoscedasticity issues (Ruxton, 2006). This assumption is false, since the Mann-Whitney test remains sensitive to heteroscedasticity (Grissom, 2000; Nachar, 2008; Neuhauser & Ruxton, 2009).

- ⁷ Like Bryk and Raudenbush (1988), we note that unequal variances between groups does not systematically mean that population variances are different: standard deviation ratios are more or less biased estimates of population variance (see table A2 in the additional file). Differences can be a consequence of bias in measurement, such as response styles (Baumgartner & Steenkamp, 2001). However, there is no way to determine what part of the variability is due to error rather than the true population value.

- ⁸ Also known as the Satterwaite's test, the Smith/Welch/Satterwaite test, the Aspin-Welch test, or the unequal variances *t*-test.

Competing Interests

The authors have no competing interests to declare.

Additional File

The additional file for this article can be found as follows:

- DOI: <https://doi.org/10.5334/irsp.82s1>

Author's Note

All code needed to recreate the simulations resulting in the figures and appendices is available at <https://osf.io/bver8/files/>, as are as the .txt files containing the results of all simulations.

References

- Baguley, T.** (2012). *Serious stats: A guide to advanced statistics for the behavioral sciences*. Palgrave Macmillan. Retrieved from <https://books.google.fr/books?hl=fr&lr=&id=ObUcBQAAQBAJ&oi=fnd&pg=PP1&dq=baguley+2012&ots=-eiUlHiCYs&sig=YUUKZ7jiGF33wdo3WVO-8l-OUu8>.
- Banerjee, A., Chitnis, U., Jadhav, S., Bhawalkar, J. & Chaudhury, S.** (2009). Hypothesis testing, type I and type II errors. *Industrial Psychiatry Journal*, 18(2), 127. DOI: <https://doi.org/10.4103/0972-6748.62274>
- Baumgartner, H. & Steenkamp, J.-B. E.** (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, 38(2), 143–156. DOI: <https://doi.org/10.1509/jmkr.38.2.143.18840>
- Bryk, A. S. & Raudenbush, S. W.** (1988). Heterogeneity of variance in experimental studies: A challenge to conventional interpretations. *Psychological Bulletin*, 104(3), 396. DOI: <https://doi.org/10.1037/0033-2909.104.3.396>
- Carroll, R. J. & Schneider, H.** (1985). A note on Levene's tests for equality of variances. *Statistics & Probability Letters*, 3(4), 191–194. DOI: [https://doi.org/10.1016/0167-7152\(85\)90016-1](https://doi.org/10.1016/0167-7152(85)90016-1)
- Cowdry, R. W., Gardner, D. L., O'Leary, K. M., Leibenluft, E. & Rubinow, D. R.** (1991). Mood variability: A study of four groups. *American Journal of Psychiatry*, 148(11), 1505–1511. DOI: <https://doi.org/10.1176/ajp.148.11.1505>
- Cumming, G.** (2013). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge. Retrieved from https://books.google.fr/books?hl=fr&lr=&id=1W6laNc7Xt8C&oi=fnd&pg=PR1&dq=understanding+the+new+statistics:+effect+sizes,+confidence+intervals,+and+meta-analysis&ots=PujZVHb03Q&sig=IhSjkfzp4o5OXAKhZ_zYzP9nsr8.
- Erceg-Hurn, D. M. & Miroseovich, V. M.** (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist*, 63(7), 591. DOI: <https://doi.org/10.1037/0003-066X.63.7.591>
- Feingold, A.** (1992). Sex differences in variability in intellectual abilities: A new look at an old controversy. *Review of Educational Research*, 62(1), 61–84. DOI: <https://doi.org/10.3102/00346543062001061>
- Glass, G. V., Peckham, P. D. & Sanders, J. R.** (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, 42(3), 237–288. DOI: <https://doi.org/10.3102/00346543042003237>
- Grisson, R. J.** (2000). Heterogeneity of variance in clinical data. *Journal of Consulting and Clinical Psychology*, 68(1), 155. DOI: <https://doi.org/10.1037/0022-006X.68.1.155>
- Hanushek, E. A. & Wößmann, L.** (2006). Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries*. *Economic Journal*, 116(510), C63–C76. DOI: <https://doi.org/10.1111/j.1468-0297.2006.01076.x>
- Hayes, A. F. & Cai, L.** (2007). Further evaluating the conditional decision rule for comparing two independent means. *British Journal of Mathematical and Statistical Psychology*, 60(2), 217–244. DOI: <https://doi.org/10.1348/000711005X62576>
- Henrich, J., Heine, S. J. & Norenzayan, A.** (2010). Most people are not WEIRD. *Nature*, 466(7302), 29–29. DOI: <https://doi.org/10.1038/466029a>
- Hoekstra, R., Kiers, H. & Johnson, A.** (2012). Are assumptions of well-known statistical techniques checked, and why (not)? *Frontiers in Psychology*, 3, 137. DOI: <https://doi.org/10.3389/fpsyg.2012.00137>
- Keppel, G.** (1991). *Design and analysis: A researcher's handbook*. Prentice-Hall, Inc. Retrieved from <http://psycnet.apa.org/psycinfo/1991-98751-000>.
- Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., Levin, J. R., et al.** (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research*, 68(3), 350–386. DOI: <https://doi.org/10.3102/00346543068003350>
- Keselman, H. J., Othman, A. R., Wilcox, R. R. & Fradette, K.** (2004). The new and improved two-sample *t* test. *Psychological Science*, 15(1), 47–51. DOI: <https://doi.org/10.1111/j.0963-7214.2004.01501008.x>
- Kester, S. W.** (1969). The communication of teacher expectations and their effects on the achievement and attitudes of secondary school pupils. University of Oklahoma. Retrieved from <https://shareok.org/handle/11244/2570>.
- Lakens, D., Hilgard, J. & Staaks, J.** (2016). On the reproducibility of meta-analyses: Six practical recommendations. *BMC Psychology*, 4(1), 1. DOI: <https://doi.org/10.1186/s40359-016-0126-3>
- Lim, T.-S. & Loh, W.-Y.** (1996). A comparison of tests of equality of variances. *Computational Statistics & Data Analysis*, 22(3), 287–301. DOI: [https://doi.org/10.1016/0167-9473\(95\)00054-2](https://doi.org/10.1016/0167-9473(95)00054-2)
- Luh, W.-M. & Guo, J.-H.** (2007). Approximate sample size formulas for the two-sample trimmed mean test with unequal variances. *British Journal of Mathematical and Statistical Psychology*, 60(1), 137–146. DOI: <https://doi.org/10.1348/000711006X100491>
- Moser, B. K. & Stevens, G. R.** (1992). Homogeneity of variance in the two-sample means test. *American Statistician*, 46(1), 19–21. DOI: <https://doi.org/10.1080/00031305.1992.10475839>
- Nachar, N.** (2008). The Mann-Whitney U: A test for assessing whether two independent samples come from the same distribution. *Tutorials in Quantitative Methods for Psychology*, 4(1), 13–20. DOI: <https://doi.org/10.20982/tqmp.04.1.p013>
- Neuhäuser, M. & Ruxton, G. D.** (2009). Distribution-free two-sample comparisons in the case of heterogeneous variances. *Behavioral Ecology and Sociobiology*, 63(4), 617–623. DOI: <https://doi.org/10.1007/s00265-008-0683-4>

- Nimon, K. F.** (2012). Statistical assumptions of substantive analyses across the general linear model: A mini-review. *Frontiers in Psychology*, 3, 322. DOI: <https://doi.org/10.3389/fpsyg.2012.00322>
- Nordstokke, D. W. & Zumbo, B. D.** (2007). A Cautionary Tale about Levene's Tests for Equal Variances. *Journal of Educational Research & Policy Studies*, 7(1), 1–14.
- Overall, J. E., Atlas, R. S. & Gibson, J. M.** (1995). Tests that are robust against variance heterogeneity in $k \times 2$ designs with unequal cell frequencies. *Psychological Reports*, 76(3), 1011–1017. DOI: <https://doi.org/10.2466/pr0.1995.76.3.1011>
- Rakotomalala, R.** (2008). Comparaison de populations. *Tests Non Paramétriques*, Université Lumière Lyon, 2. Retrieved from http://www.academia.edu/download/44989200/Comp_Pop_Tests_Nonparametriques.pdf.
- Rasch, D., Kubinger, K. D. & Moder, K.** (2011). The two-sample *t* test: Pre-testing its assumptions does not pay off. *Statistical Papers*, 52(1), 219–231. DOI: <https://doi.org/10.1007/s00362-009-0224-x>
- Ruxton, G. D.** (2006). The unequal variance *t*-test is an underused alternative to Student's *t*-test and the Mann–Whitney U test. *Behavioral Ecology*, 17(4), 688–690. DOI: <https://doi.org/10.1093/beheco/ark016>
- Shaw, R. G. & Mitchell-Olds, T.** (1993). ANOVA for unbalanced data: An overview. *Ecology*, 74(6), 1638–1645. DOI: <https://doi.org/10.2307/1939922>
- Sheskin, D. J.** (2003). *Handbook of parametric and nonparametric statistical procedures* (3rd ed.). Boca Raton, Florida: CRC Press. DOI: <https://doi.org/10.1201/9781420036268>
- Shields, S.** (1975). Functionalism, Darwinism, and the psychology of women. *American Psychologist*, 30(7), 739. DOI: <https://doi.org/10.1037/h0076948>
- Student.** (1908). The probable error of a mean. *Biometrika*, 1–25. DOI: <https://doi.org/10.1093/biomet/6.1.1>
- Wallenstein, S., Zucker, C. L. & Fleiss, J. L.** (1980). Some statistical methods useful in circulation research. *Circulation Research*, 47(1), 1–9. DOI: <https://doi.org/10.1161/01.RES.47.1.1>
- Wang, H., Smith, K. P., Combs, E., Blake, T., Horsley, R. D. & Muehlbauer, G. J.** (2012). Effect of population size and unbalanced data sets on QTL detection using genome-wide association mapping in barley breeding germplasm. *Theoretical and Applied Genetics*, 124(1), 111–124. DOI: <https://doi.org/10.1007/s00122-011-1691-8>
- Welch, B. L.** (1947). The generalization of Student's problem when several different population variances are involved. *Biometrika*, 34(1/2), 28–35. DOI: <https://doi.org/10.2307/2332510>
- Wilcox, R. R., Granger, D. A. & Clark, F.** (2013). Modern robust statistical methods: Basics with illustrations using psychobiological data. *Universal Journal of Psychology*, 1(2), 21–31.
- Wilcox, R. R. & Keselman, H. J.** (2003). Modern robust data analysis methods: Measures of central tendency. *Psychological Methods*, 8(3), 254. DOI: <https://doi.org/10.1037/1082-989X.8.3.254>
- Yuen, K. K.** (1974). The two-sample trimmed *t* for unequal population variances. *Biometrika*, 61(1), 165–170. DOI: <https://doi.org/10.1093/biomet/61.1.165>
- Zimmerman, D. W.** (1996). Some properties of preliminary tests of equality of variances in the two-sample location problem. *Journal of General Psychology*, 123(3), 217–231. DOI: <https://doi.org/10.1080/00221309.1996.9921274>
- Zimmerman, D. W.** (2000). Statistical significance levels of nonparametric tests biased by heterogeneous variances of treatment groups. *Journal of General Psychology*, 127(4), 354–364. DOI: <https://doi.org/10.1080/00221300009598589>
- Zimmerman, D. W.** (2004). A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology*, 57(1), 173–181. DOI: <https://doi.org/10.1348/000711004849222>
- Zumbo, B. D. & Coulombe, D.** (1997). Investigation of the robust rank-order test for non-normal populations with unequal variances: The case of reaction time. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, 51(2), 139. DOI: <https://doi.org/10.1037/1196-1961.51.2.139>

How to cite this article: Delacre, M., Lakens, D. and Leys, C. (2017). Why Psychologists Should by Default Use Welch's *t*-test Instead of Student's *t*-test. *International Review of Social Psychology*, 30(1), 92–101, DOI: <https://doi.org/10.5334/irsp.82>

Published: 05 April 2017

Copyright: © 2017 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

 *International Review of Social Psychology* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS 

3 Chapitre 3: le test W de Welch (article 2)

4 Chapitre 4: le d de Cohen avec variances non poolées (article 3)

5 Chapitre 5: les tests d'équivalence (article 4)

6 Conclusion

The conclusion section should specify the key findings of your study, explain their wider significance in the context of the research field and explain how you have filled the knowledge gap that you have identified in the introduction. This is your chance to present to your reader the major take-home messages of your dissertation research. It should be similar in content to the last sentence of your summary abstract. It should not be a repetition of the first paragraph of the discussion. They can be distinguished in their connection to broader issues. The first paragraph of the discussion will tend to focus on the direct scientific implications of your work (i.e. basic science, fundamental knowledge) while the conclusion will tend to focus more on the implications of the results for society, conservation, etc.

6.0.1 Usage des articles méthodo

Ces derniers semblent assez peu utilisés par les chercheurs. En tout cas s'ils les utilisent, ils les citent très peu dans leurs références pour justifier leurs choix (cf. article de Mills, Abdulla, & Cribbie (2010): le mode du nombre de citation méthodo dans les articles de recherche appliqués est 0, et la médiane vaut 1...). Dans l'autre sens, on constate que les articles méthodologiques sont généralement peu cités, et ils le sont encore 3 fois moins par les chercheurs appliqués que par les autres méthodologistes (Mills et al., 2010, p. 56).

On est en droit de questionner l'impact réel des publications méthodologiques, pour 2 raisons, d'après Mills et al. (2010):

- (1) Les chercheurs appliqués sont noyés sous les articles dans leur domaine d'expertise si bien que cela limite le temps dont ils disposent pour se consacrer aux articles méthodologiques.
- (2) malgré que des nouvelles méthodes sont disponibles, les chercheurs continuent à opter pour des tests traditionnels et familiaux (mais souvent inappropriés).

-> Qu'est-ce qui va pousser les chercheurs à lire des articles méthodologiques? -> Si je trouve la réponse à ceci, j'ai mon intro.

6.0.2 Importance des simulations et des logiciels modernes pour enseigner les statistiques fréquentistes

On sait que les chercheurs tendent à privilégier les méthodes qui sont proposées par défaut dans des logiciels de clique bouton (comme SPSS). C'est en tout cas ce que dit Counsell & Harlow (2017) dans le contexte de la gestion des données manquantes (mais je crois que c'est vrai pour tout). Une manière d'améliorer les pratiques serait d'améliorer les options proposées par défaut dans les logiciels de

clic-bouton. C'est à ce genre de choses que j'aspire à travers mes articles.

Malgré tout, un logiciel ne fait pas tout et après avoir utilisé le test adéquat, il est important d'être capable de l'interpréter correctement. Les tests font appel à des notions faussement simples telles que les p-valeurs et les distributions d'échantillonnage. A mon sens, le seul moyen d'enseigner correctement ces notions, c'est à travers des simulations.

6.0.3 Comment écrire/transmettre l'info aux psys

Un consultant doit pouvoir parler de langage des psys, c'est-à-dire décrire et expliquer les méthodes requises d'une manière compréhensible pour les clients (Golinski & Cribbie, 2009). Est-ce bien de demander à des mathématiciens/Statisticiens d'enseigner les stat aux psy's? Par forcément, car un psychologue spécialisé en méthodo quanti sera plus à même de comprendre les procédures et méthodes requises par les psys (ex. de la question de la taille d'effet qui n'intéresse pas vraiment les statisticiens; Golinski & Cribbie (2009)).

6.0.4 Recommandations générales

Mills et al. (2010):

- au moins un reviewer compétant pour analyser le caractère approprié des méthodes stat
- que les éditeurs/reviewers encouragent l'usage d'article de méthodo dans leur recherche → interesting, mais réaliste? La proportion de méthodologistes parmi les psychologues n'est pas assez élevée... Ou alors il faut vraiment de l'interdisciplinarité!

- 2) Par manque de connaissances, les chercheurs se contentent souvent des informations fournies dans les logiciels clic/bouton. *for example, if software does not report a CI on Cohen's d, it is unlikely that a researcher will calculate one his or herself* (Counsell & Harlow (2017)). *Une chance qu'on a, c'est Jamovi* (regarder si Jamovi me cite)

7 Bibliographie

- Aiken, L. S., West, S. G., & Millsap, R. E. (2008). Doctoral training in statistics, measurement, and methodology in psychology: Replication and extension of Aiken, West, Sechrest, and Reno's (1990) survey of PhD programs in North America. *American Psychologist*, 63(1), 32.
- Byrne, B. M. (1996). The status and role of quantitative methods in psychology: Past, present, and future perspectives. *Canadian Psychology/Psychologie canadienne*, 37(2), 76.
- Cain, M. K., Zhang, Z., & Yuan, K.-H. (2017). Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation. *Behavior research methods*, 49(5), 1716-1735.
- Counsell, A., & Harlow, L. (2017). Reporting practices and use of quantitative methods in Canadian journal articles in psychology. *Canadian Psychology/psychologie canadienne*, 58(2), 140.
- Curtis, D. A., & Harwell, M. (1998). Training doctoral students in educational statistics in the United States: A national survey. *Journal of Statistics Education*, 6(1).
- Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods: an easy way to maximize the accuracy and power of your research. *American Psychologist*, 63(7), 591.
- Everitt, B. S. (2001). *Statistics for psychologists: An intermediate course*. (S.l.) : Psychology Press.
- Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. (S.l.) : sage.
- Golinski, C., & Cribbie, R. A. (2009). The expanding role of quantitative methodologists in advancing psychology. *Canadian Psychology/Psychologie canadienne*, 50(2), 83.
- Grissom, R. J. (2000). Heterogeneity of variance in clinical data. *Journal of consulting and clinical psychology*, 68(1), 155.
- Haslam, S. A., & McGarty, C. (2014). *Research methods and statistics in psychology*. (S.l.) : Sage.
- Hoekstra, R., Kiers, H., & Johnson, A. (2012). Are assumptions of well-known statistical techniques checked, and why (not)? *Frontiers in psychology*, 3, 137.
- Howitt, D., & Cramer, D. (2017). *Understanding statistics in psychology with SPSS*. (S.l.) : Pearson London, UK:

- Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., . . . Keselman, J. C. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of educational research*, 68(3), 350-386.
- Lakens, D. (2021). The practical alternative to the p value is the correctly used p value. *Perspectives on psychological science*, 16(3), 639-648.
- Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological bulletin*, 105(1), 156.
- Mills, L., Abdulla, E., & Cribbie, R. (2010). Quantitative methodology research: Is it on psychologists' reading lists?
- Nunnally, J. (1960). The place of statistics in psychology. *Educational and Psychological Measurement*, 20(4), 641-650.
- Osborne, J. W., & Christianson, W. R. (2001). Educational Psychology from a Statistician's Perspective: A Review of the Quantitative Quality of Our Field.
- Osborne, J. W., & Waters, E. (2002). Four assumptions of multiple regression that researchers should always test. *Practical assessment, research, and evaluation*, 8(1), 2.
- Rasch, D., Kubinger, K. D., & Moder, K. (2011). The two-sample t test: pre-testing its assumptions does not pay off. *Statistical papers*, 52(1), 219-231.
- Ruxton, G. D. (2006). The unequal variance t-test is an underused alternative to Student's t-test and the Mann-Whitney U test. *Behavioral Ecology*, 17(4), 688-690.
- Schucany, W. R., & Tony Ng, H. K. (2006). Preliminary goodness-of-fit tests for normality do not validate the one-sample Student t. *Communications in Statistics-Theory and Methods*, 35(12), 2275-2286.
- Sharpe, D. (2013). Why the resistance to statistical innovations? Bridging the communication gap. *Psychological methods*, 18(4), 572.
- Student. (1908). The probable error of a mean. *Biometrika*, 1-25.
- Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist*, 53(3), 300.
- Yuan, K.-H., Bentler, P. M., & Chan, W. (2004). Structural equation modeling with heavy tailed distributions. *Psychometrika*, 69(3), 421-436.

Zimmerman, D. W. (2004). A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology*, 57(1), 173-181.

8 Annexe(s)

8.1 Annexe A: erratum

8.1.1 Why psychologists Should by Default Use Welch's t -test Instead of Student's t -test

8.1.1.1 Erreurs conceptuelles

- Erreur (p. [99]): “As it is explained in the additional file, Yuen's t -test is not a better test than Welch's t -test, since it often suffers high departure from the alpha risk of 5 percent”.

Explication: d'un point de vue purement statistique, à travers le test de Yuen, on ne compare plus les moyennes de chaque groupe, mais les moyennes *trimmées* (soit les moyennes calculées sur les données après avoir écarté les 20% des scores les plus bas ainsi que les 20% des scores les plus élevés). Or, à travers nos simulations, les scénarios créés en vue de tester le taux d'erreur de type I (risque alpha) étaient systématiquement des scénarios dans lesquels les moyennes de chaque groupe étaient identiques. Lorsque les distributions sont symétriques, les moyennes de populations sont identiques aux moyennes trimmées des populations. Quand les distributions sont asymétriques, par contre, la moyenne trimmée sera plus proche du mode de la distribution que la moyenne. Cela explique pourquoi nous avons noté ceci: "Yuen's t -test is not a good unconditional alternative because we observe an unacceptable departure from the nominal alpha risk of 5 percent for several shapes of distributions [...] particularly when we are studying asymmetric distributions of unequal shapes.

A partir du moment où cela ne correspond *pas* à l'hypothèse nulle du test de Yuen, il est erroné de conclure à une inflation du risque alpha pour ce test sur base de nos scénarios. *D'un point de vue méthodologique, par contre, nous pensons qu'il n'est pas clair pour beaucoup de gens que l'hypothèse nulle du test de Yuen n'est pas la même que celle du test t de Student, ce qui pourrait être source d'ambiguïté. En conclusion, ce test ne devrait être utilisé que par des personnes ayant pleinement conscience du fait que le test de Student et celui de Yuen ne testent pas la même hypothèse. **Relire l'article de Wilcoxon et celui de Erceg-Hurn pour voir si c'est vraiment bien clair dedans, histoire de me défendre un peu*

When both the normality and equal variances assumptions are violated, we can use a combination of the Trimmed Means t -Test and Welch's t -Test, called the Yuen-Welch Test. Using the notation for the Trimmed Means t -test, the Yuen-Welch Test is

8.1.1.2 Mise en forme et Notations En termes de mises en forme, nous avons omis à plusieurs endroits d'italiser les notations mathématiques. Par exemple, à la page [93], nous avons indiqué “F-ratio

test” au lieu de “ F -ratio test” à plusieurs reprises. A la même page, nous avons également noté “ x_{ij} ” au lieu de “ x_{ij} ”, et $|x_{ij} - \hat{\theta}_j|$ au lieu de $|x_{ij} - \hat{\theta}_j|$.

En termes de notation, nous avons relevé des inconsistantes dans la notation de la variance de chaque groupes. Par exemple, dans l’équation 1, nous utilisons S_j^2 , alors que nous utilisons s_j^2 dans l’équation 4 ou encore SD_j lorsque nous définissons la statistique du F -ratio à la page [93].

8.1.2 Taking parametric assumptions very seriously

8.1.2.1 Erreurs conceptuelles

8.1.2.2 Mise en forme et Notations Dû à un manque de connaissance de Latex lors de mes premières tentatives d’écritures d’articles via Rmarkdown, certaines majuscules sont manquantes dans les références bibliographiques. S’assurer qu’une lettre apparaissent en majuscule, via latex, implique de l’entourer des symboles $\{\}$, ce qui n’a pas été fait. Par exemple, dans le titre de l’article de Tiku(1971), il aurait fallu indiquer “Power function of the $\{F\}$ -test..” au lieu de “Power function of the F -test...”. Cela ne serait pas arrivé, si j’avais utilisé un outil comme Zotero, afin d’exporter directement un fichier au format Bibtex (puisque via ces outils, ce genre de détail est automatiquement inclu), mais je n’ai découvert cette possibilité que très récemment.

8.1.3 Effect sizes

8.1.4 Equivalence tests

8.2 Annexe B

Insert code (if any) used during your dissertation work here.