Université Libre de Bruxelles

Faculté des Sciences Psychologiques et de l'Education

METHODE EN PSYCHOLOGIE : OUTILS ET RECOMMANDATIONS POUR COMPRENDRE ET GERER LES HYPOTHESES, CONDITIONS D'APPLICATION DES TESTS ET LES MESURES DE TAILLE D'EFFETS

Par

MARIE DELACRE

En vue de l'obtention du grade de docteur

Septembre 2021

Abstract

Abstract (voir nombre de mots requis)

Table des matières

Remerciements

Thank you for following this tutorial!

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

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 quantitatives[1] (Counsell & Harlow, 2017), les connaissances statistiques s'avèrent fondamentales pour comprendre, planifier et analyser une recherche (Everitt, 2001 ; Howitt & Cramer, 2017). C'est la raison pour laquelle les statistiques font 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 ; Judd, McClelland, & Ryan, 2011). Cela pourrait vraisemblablement expliquer pourquoi ces tests sont si persistants dans la recherche en psychologie (Counsell & Harlow, 2017). Ils 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 psychologie[2], 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 un test de 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 de ces tests de comparaison de moyennes (Aiken, West, & Millsap, 2008 ; Hoekstra, Kiers, & Johnson, 2012). Pourtant, certains indices semblent contredire cette conviction.

Bien qu'il existe plusieurs types de tests t et d'ANOVA, les chercheurs en psychologie privilégient souvent par défaut le test t de Student et l'ANOVA de Fisher[3]. Ceux-ci consistent à comparer les scores moyens de deux (ou plusieurs) groupes indépendants de sujets et reposent sur les hypothèses que les résidus, indépendants et identiquement distribués soient extraits d'une population qui se distribue normalement, et que tous les groupes soient extraits de populations

ayant la même variance (c'est ce qu'on appelle la condition d'homogénéité des variances). 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, Kiers, & Johnson (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 Harvey J. Keselman et al. (1998), seulement 5% 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.

Bien entendu, 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 généralement 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, Kiers, & Johnson, 2012)[4]. Dans un tel contexte, il n'y a que peu d'intérêt pour les chercheurs à en discuter, si ce n'est pour discuter des violations des conditions (et éventuellement, se servir de cette information pour justifier une décision qui en découle). Néanmoins, même en partant du postulat que les conditions ne seront mentionnées que lorsqu'elles sont violées, il est surprenant d'observer que

ces 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 Harvey J. 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, Kiers, & Johnson (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 dûe à leur manque de familiarité avec les conditions d'application des tests, plutôt qu'à 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 parait, et tout chercheur désireux d'améliorer la transparence dans la transmission des analyses de données resterait confronté à un problème majeur: les conditions d'homogénéité des variances et de normalité reposent sur les paramètres de *population* et non sur les paramètres des *échantillons*. Comme ces paramètres de population ne sont pas connus (Hoekstra, Kiers, & Johnson, 2012), on doit utiliser les paramètres des échantillons pour tenter d'inférer sur le respect des conditions d'application. Souvent, les chercheurs font cette inférence en utilisant des tests d'hypothèses, mais 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). La difficulté que représente la vérification des conditions d'application ne constituerait pas réellement un problème, en soi, si les tests t de Student et F de Fisher étaient susceptibles de fournir des conclusions non biaisées et fiables même en cas d'écart 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 (Harvey J. Keselman et al., 1998).

Compte tenu de tous les éléments précités, il semblerait donc qu'une solution 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 ; Harvey J. Keselman et al., 1998 ; Wilcox, 1998). Dans la mesure où une revue approfondie de la littérature démontre que les taux d'erreur de type I et II des tests t de Student et F de Fisher sont bien plus affectés par le non respect de la condition d'homogénéité des variances que par le non respect de la condition de normalité (Erceg-Hurn & Mirosevich, 2008 ; Grissom, 2000 ; Hoekstra, Kiers, & Johnson, 2012 ; Osborne & Waters, 2002), 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 semble avoir eu un impact limité sur les pratiques des chercheurs en psychologie. Pour s'assurer de faire passer notre message, nous nous appliquerons particulièrement, au sein des articles présentés dans les chapitres 2 à 3, à nous adresser directement à ce public de chercheurs. Pour ce faire, nous tenterons (1) d'expliquer concrètement pourquoi selon nous, 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 conclurons ces articles par des recommandations concrètes, afin d'aider les chercheurs à extraire le message clé de ces articles.

Au delà des tests d'hypothèse, de nombreux journaux de psychologie encouragent (voire même requièrent) de quantifier la taille des effets étudiés et de fournir un intervalle de confiance autour des estimations de taille d'effet (Cumming, Fidler, Kalinowski, & Lai, 2012). L'année 1999 a joué un rôle clé dans la mise en oeuvre de ces recommendations, puisque l'*APA Task Force* a publié un rapport dans lequel elle soulignait l'importance de reporter des mesures de taille d'effet.

Ce rapport a été suivi de recommandations précises de la part de l'American Psychological Association (APA) et de l'American Educational Research Association (AERA) quant à la manière de reporter ces mesures (Peng, Chen, Chiang, & Chiang, 2013). Or, il semblerait que ces diverses recommandations aient été associées à des modifications dans les pratiques des chercheurs. Peng, Chen, Chiang, & Chiang (2013) ont étudié l'évolution du taux moyen de report des mesures de taille d'effet en comparant ce taux moyen avant et après 1999, distinctement dans 19 revues consacrées à la recherche dans les domaines de la Psychologie et de l'Education. Ils ont noté une augmentation du taux variant de 5.2 % à 96.3 % dans chacun de ces journaux. Ils ont cependant également noté la persistance de pratiques inadéquates, telles que la dominance de la mesure du traditionnel d de Cohen. Le d de Cohen est une mesure de taille d'effet standardisée qui appartient à la famille d et qui entretient une relation mathématique directe avec le t de Student. Par conséquent, il dépend des mêmes conditions d'application que le test t de Student, c'est donc sans surprise qu'en cas de violation de ces conditions, son usage peut amener à une sous-représentation (ou au contraire à une sur-représentation) de la taille d'effet (Grissom & Kim, 2001). De même que pour le test t, il semblerait que ce soit essentiellement la violation de la condition d'homogénéité des variances qui soit problématique. De nombreux auteurs se sont demandés si le d de Cohen pourrait être remplacé par une autre mesure de la même famille lorsque les variances diffèrent d'une population à l'autre, mais ils n'ont pas trouvé de consensus quant à la solution la plus appropriée (Shieh, 2013). Pour répondre à cette problématique, nous présenterons, dans le chapitre 4, des simulations Monte Carlo pour comparer le traditionnel d de cohen aux mesures de la famille d les plus communément proposées pour le remplacer en cas d'hétéroscédasticité. Dans ce chapitre, nous nous focaliserons exclusivement sur les mesures de la famille d parce que les chercheurs utilisent très fréquemment le d de Cohen lorsqu'ils réalisent un test t. Nous tenons cependant à rappeler que la notion de taille d'effet est très vaste. Elle englobe toute mesure susceptible de fournir une information relative à l'ampleur d'un effet étudié, que ce soit à travers une mesure *non standardisée* (moyenne, médiane, coefficient de régression non standardisé...) ou à travers une mesure *standardisées* (R^2, coefficient de régression standardisé, différence de moyennes standardisée..., Counsell & Harlow, 2017). Les mesures non standardisées peuvent également s'avérer très utiles, et ce même si une emphase sur les tailles d'effet standardisées a pû donner l'impression que seules ces dernières étaient dignes d'intérêt (pour une discussion intéressante sur l'intérêt des mesures non standardisées,

nous recommandons l'article de Pek & Flora, 2018). Au sein de ce chapitre, nous tenterons de comparer l'efficacité de différents estimateurs sous des déviations réalistes de la condition de normalité, en nous appuyant sur l'investigation de Cain, Zhang, & Yuan (2017). Ces auteurs avaient en effet calculé les indicateurs d'asymétrie et d'aplatissement[5] de 1567 distributions univariées provenant de 194 articles publiés dans *Psychological Science* (de Janvier 2013 à juin 2014) et *American Education Research Journal* (de janvier 2010 à juin 2014). De plus, au delà des simulations, nous proposerons des outils (package *R* et applications Shiny pour ceux qui ne sont pas familiers avec *R*) afin d'aider les chercheurs à calculer différents estimateurs de taille d'effet ainsi que les bornes de l'intervalle de confiance autour de ces estimateurs. Cela nous semble être une nécessité puisque malgré les recommandations, il semblerait que les mesures de taille d'effet soient rarement accompagnées d'un intervalle de confiance dans la littérature (Counsell & Harlow, 2017 ; Peng, Chen, Chiang, & Chiang, 2013). C'est vrai même lorsque ces mesures sont utilisées seules (indépendamment d'un test d'hypothèse, Counsell & Harlow, 2017). Dans le cas des mesures standardisées, le calcul des intervalles de confiance est particulièrement complexe puisqu'il requiert l'usage des distributions non centrales (Balluerka, Gómez, & Hidalgo, 2005).

Les mesures de taille d'effet, standardisées ou non, et leur intervalles de confiance sont parfois vus comme des outils qui permettent de combler certaines limites des tests d'hypothèse. Une critique fréquemment avancée à l'égard des tests d'hypothèse est le fait qu'un rejet de l'hypothèse nulle ne fournit qu'une idée de la direction de l'effet, sans information relative à son ampleur. Cette critique repose implicitement sur la conception d'après laquelle l'hypothèse nulle doit être définie comme l'absence d'effet (ou l'absence de différence entre les groupes). Il est vrai que c'est l'hypothèse nulle la plus couramment définie par les chercheurs (Lakens, Scheel, & Isager, 2018 ; Nickerson, 2000). Pourtant, lorsque c'est pertinent, il est possible d'incorporer la significativité pratique dans les tests d'hypothèse. Cela implique de réfléchir *a priori* aux effets qui présentent un intérêt pratique aux yeux des chercheurs et des praticiens (Fraas & Newman, 2000), ce qui peut se faire sur base de diverses considérations, telles que des comparaisons coûts/bénéfices, par exemple (Fraas & Newman, 2000). Dans ce contexte, l'hypothèse nulle n'est plus que l'effet soit nul, mais qu'il ne dépasse pas une certaine valeur ou autrement dit, dans le cadre d'un test de comparaison de moyenne, que la différence de moyennes entre les groupes ne dépasse pas une certaine valeur (Newman, Fraas, & Herbert, 2001). Un rejet de l'hypothèse nulle ne constituera alors plus un soutien en faveur de n'importe quel effet non nul, mais plutôt un soutien en faveur

d'un effet jugé pertinent. Il est également possible de montrer un soutien en faveur de l'absence d'un effet jugé pertinent, en définissant comme hypothèse nulle que l'effet dépasse une certaine valeur (Lakens, Scheel, & Isager, 2018). C'est le principe des tests d'équivalence, qui feront l'objet du cinquième et dernier chapitre de cette thèse.

Notes de fin de chapitre

[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.

[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".

[3]Parfois, ils le font de manière implicite, en indiquant réaliser un test t (ou une ANOVA) mais sans préciser duquel (ou de laquelle) il s'agit. Cela arrive même avec des méthodologistes! Dans l'article de Tomczak & Tomczak (2014), par exemple, ils parlent de l'ANOVA et du test t, sans précision, et ce n'est qu'en lisant l'ensemble de l'article qu'on comprend qu'en réalité, ils font allusion exclusivement au test t de Student et à l'ANOVA de Fisher, entre autres, parce qu'ils proposent d'associer ces tests à des mesures de taille d'effet qui impliquent l'usage du terme de variance poolée, qui sera décrit juste après.

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

[5]Nous utilisons ce terme "aplatissement" parce que c'est ainsi que l'on traduit communément le terme anglais "kurtosis". Comme nous le préciserons cependant dans le chapitre 2, cette traduction commune ne représente que mal cette mesure, qui nous informe plus sur la densité des distributions, au niveau des extrêmités, que sur leur aplatissement

Chapitre 2: utiliser le test t de Welch par défaut[6]

INTERNATIONAL REVIEW OF SOCIAL PSYCHOLOGY

Delacre, M., et al (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

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 ith observation in the jth group, and $\hat{\theta}_j$ is the "center" of the distribution for the jth 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

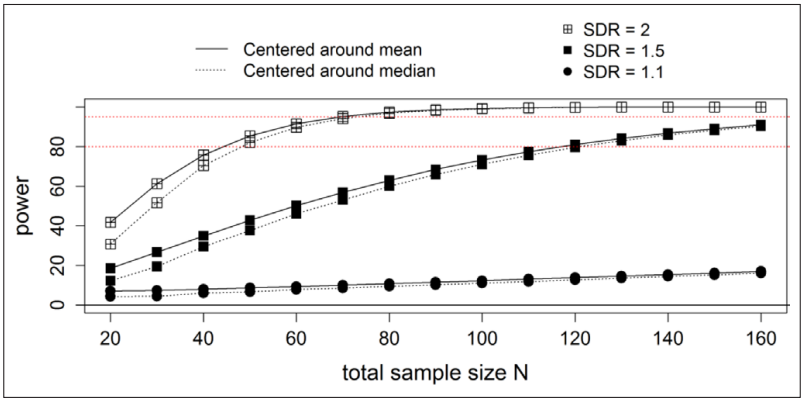


Figure 1: Estimated power of Levene's test as a function of sample size, SDR and centering parameter.

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 *SPPS* (*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 & Mirosevich, 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 & Mirosevich, 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 & Mirosevich, 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 than 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 ≈ 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]

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

The degrees of freedom are computed as follows:

$$\mathrm{df} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}} \tag{4}$$

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	$t_{Welch} = t_{Student}$ $df_{Welch} = df_{Student}$ $p_{Welch} = p_{Student}$	$t_{Welch} = t_{Student}$ $df_{Welch} \neq df_{Student}$ $p_{Welch} \neq p_{Student}$
Unbalanced design	$t_{Welch} = t_{Student}$ $df_{Welch} \neq df_{Student}$ $p_{Welch} \neq p_{Student}$	$t_{Welch} \neq t_{Student}$ $df_{Welch} \neq df_{Student}$ $p_{Welch} \neq p_{Student}$

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**),

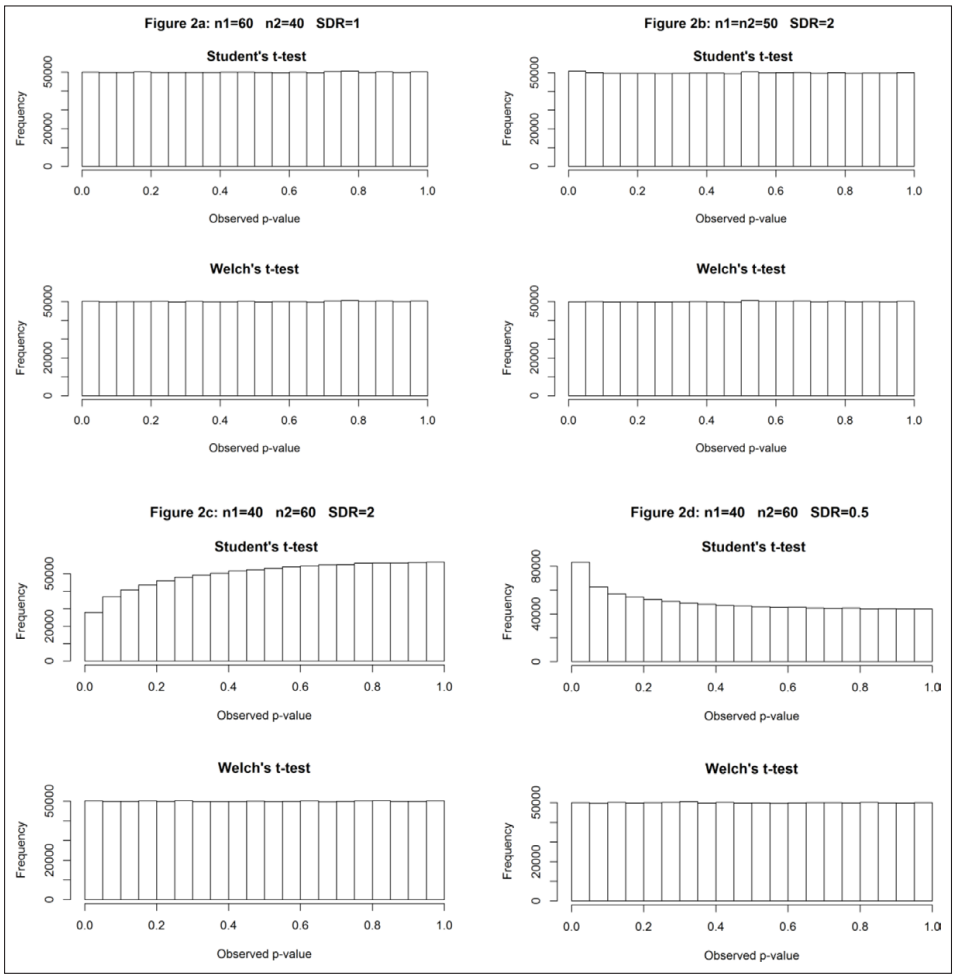


Figure 2: *P*-value distributions for Student's and Welch's *t*-test under the null as a function of SDR, and sample size.

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

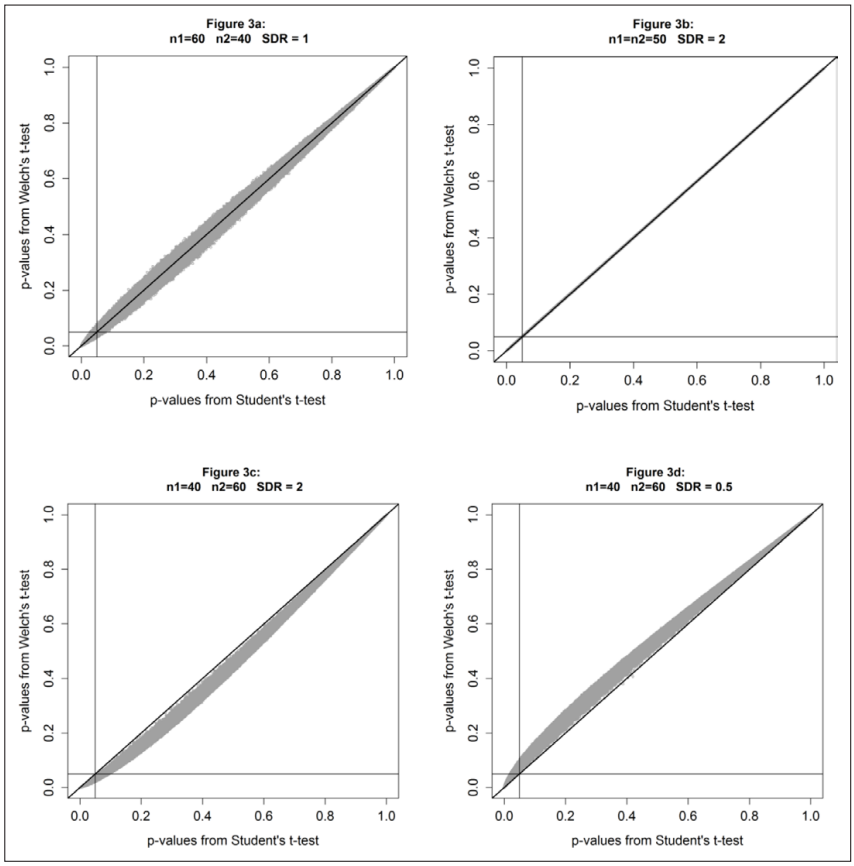


Figure 3: *P*-values from Student's *t*-test against *p*-values from Welch's *t*-test under the null.

variance differs between groups (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 (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 (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 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, SDR = X will return the same answer in terms of percent power of Levene's test as SDR = 1/X. For example, SDR = 2 will return the same answer as SDR = ½ = 0.5.

[6] For example, many statistical users believe that the Mann-Whitney non-parametric test can cope with both normality and homosedasticity issues (Ruxton, 2006). This assumption is false, since the Mann-Whitney test remains sensitive to heterosedasticity (Grissom, 2000; Nachar, 2008; Neuhäuser & 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.82.s1

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

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=lhSjkfzp4o5OXAKhZ_zYzP9nsr8.

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. 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

Grissom, 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

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 × 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 Lumiere 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. DO: 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

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

OPEN ACCESS

Note de fin de chapitre

[6]Un erratum de cet article figure dans l'annexe A.

Chapitre 3: utiliser l'ANOVA F de Welch par défaut[7]

INTERNATIONAL REVIEW OF SOCIAL PSYCHOLOGY

Delacre, M., et al. (2019). Taking Parametric Assumptions Seriously: Arguments for the Use of Welch's *F*-test instead of the Classical *F*-test in One-Way ANOVA. *International Review of Social Psychology*, 32(1): 13, 1–12. DOI: https://doi.org/10.5334/irsp.198

RESEARCH ARTICLE

Taking Parametric Assumptions Seriously: Arguments for the Use of Welch's *F*-test instead of the Classical *F*-test in One-Way ANOVA

Marie Delacre*, Christophe Leys*, Youri L. Mora* and Daniël Lakens†

Student's *t*-test and classical *F*-test ANOVA rely on the assumptions that two or more samples are independent, and that independent and identically distributed residuals are normal and have equal variances between groups. We focus on the assumptions of normality and equality of variances, and argue that these assumptions are often unrealistic in the field of psychology. We underline the current lack of attention to these assumptions through an analysis of researchers' practices. Through Monte Carlo simulations, we illustrate the consequences of performing the classic parametric *F*-test for ANOVA when the test assumptions are not met on the Type I error rate and statistical power. Under realistic deviations from the assumption of equal variances, the classic *F*-test can yield severely biased results and lead to invalid statistical inferences. We examine two common alternatives to the *F*-test, namely the Welch's ANOVA (*W*-test) and the Brown-Forsythe test (*F**-test). Our simulations show that under a range of realistic scenarios, the *W*-test is a better alternative and we therefore recommend using the *W*-test by default when comparing means. We provide a detailed example explaining how to perform the *W*-test in SPSS and R. We summarize our conclusions in practical recommendations that researchers can use to improve their statistical practices.

Keywords: ANOVA; Welch test; parametric test; parametric assumptions; replicability crisis

When comparing independent groups researchers often analyze the means by performing a Student's *t*-test or classical Analysis of Variance (ANOVA) *F*-test (Erceg-Hurn & Mirosevich, 2008; Keselman et al., 1998; Tomarken & Serlin, 1986). Both tests rely on the assumptions that independent and identically distributed residuals (1) are sampled from a normal distribution and (2) have equal variances between groups (or homoscedasticity; see Lix, Keselman, & Keselman, 1996). While a deviation from the normality assumption generally does not strongly affect either the Type I error rates (Glass, Peckham, & Sanders, 1972; Harwell, Rubinstein, Hayes, & Olds, 1992; Tiku, 1971) or the power of the *F*-test (David & Johnson, 1951; Harwell et al., 1992; Srivastava, 1959; Tiku, 1971), the *F*-test is not robust against unequal variances (Grissom, 2000). Unequal variances can alter both the Type I error rate (David & Johnson, 1951; Harwell et al., 1992) and statistical power (Nimon, 2012; Overall, Atlas, & Gibson, 1995) of the *F*-test.

* 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)

Although it is important to make sure test assumptions are met before a statistical test is performed, researchers rarely provide information about test assumptions when they report an *F*-test. We examined statistical tests reported in 116 articles in the *Journal of Personality and Social Psychology* published in 2016. Fourteen percent of these articles reported a one-way *F*-test, but only one article indicated that the homogeneity of variances assumption was taken into account. They reported corrected degrees of freedom for unequal variances, which could signal the use of the *W*-test instead of the classical *F*-test. A similar investigation (Hoekstra, Kiers & Johnson, 2012) yielded conclusions about the lack of attention to both the homoscedasticity and the normality assumptions. Despite the fact that the *F*-test is currently used by default, better alternatives exist, such as the Welch's *W* ANOVA (*W*-test), the Alexander-Govern test, James' second order test, and the Brown-Forsythe ANOVA (*F**-test). Although not the focus of the current article, additional tests exist that allow researchers to compare groups either based on other estimators of central tendency than the mean (see for example Erceg-Hurn & Mirosevich, 2008; Wilcox, 1998), or based on other relevant parameters of distribution than the central tendency, such as standard deviations and the shape of the distribution (Grissom, 2000; Tomarken & Serlin, 1986). However, since most

researchers currently generate hypotheses about differences between means (Erceg-Hurn & Mirosevich, 2008; Keselman et al., 1998), we think that a first realistic first step towards progress would be to get researchers to correctly test the hypothesis they are used to.

Although the debate surrounding the assumptions of the *F*-test has been widely explored (see for example the meta-analysis of Harwell et al., 1992), applied researchers still largely ignore the consequences of assumption violations. Non-mathematical pedagogical papers summarizing the arguments seem to be lacking from the literature, and the current paper aims to fill this gap. We will discuss the pertinence of the assumptions of the *F*-test, and focus on the question of heteroscedasticity (that, as we will see, can have major consequences on error rates). We will provide a non-mathematical explanation of how alternatives to the classical *F*-test cope with heteroscedasticity violations. We conducted simulations in which we compare the *F*-test with the most promising alternatives. We argue that when variances are equal between groups, the *W*-test has nearly the same empirical Type I error rate and power as the *F*-test, but when variances are unequal, it provides empirical Type I and Type II error rates that are closer to the expected levels compared to the *F*-test. Since the *W*-test is available in practically all statistical software packages, researchers can immediately improve their statistical inferences by replacing the *F*-test by the *W*-test.

Normality and Homogeneity of Variances under Ecological Conditions

For several reasons, assumptions of homogeneity of variances and normality are always more or less violated (Glass et al., 1972). In this section we will summarize the specificity of the methods used in our discipline that can account for this situation.

Normality Assumption

It has been argued that there are many fields in psychology where the assumption of normality does not hold (Cain, Zhang & Yuan, 2017; Micceri, 1989; Yuan, Bentler & Chan, 2004). As argued by Micceri (1989), there are several factors that could explain departures from the normality assumption, and we will focus on three of them: treatment effects, the presence of subpopulations, and the bounded measures underlying residuals.

First, although the mean can be influenced by the treatment effects, experimental treatment can also change the shape of a distribution, either by influencing the *skewness*, quantifying the asymmetry of the shape of the distribution, and *kurtosis*, a measure of the tendency to produce extreme values. A distribution with positive kurtosis will have heavier tails than the normal distribution, which means that extreme values will be more likely, while a distribution with negative kurtosis will have lighter tails than the normal distribution, meaning that extreme values will be less likely (Westfall, 2014; Wilcox, 2005). For example, a training aiming at reducing a bias perception of threat when being exposed to ambiguous words will not uniformly impact the perception of all participants, depending on their level of anxiety (Grey & Mathews, 2000). This could influence the kurtosis of the distribution of bias score.

Second, prior to any experimental treatment, the presence of several subpopulations may lead to departures from the normality assumptions. A subgroup might exist that is unequal on some characteristics relevant to the measurements, that are not controlled within the studied group, which results in mixed distributions. This unavoidable lack of control is inherent of our field given its complexity. As an illustration, Wilcox (2005) writes that pooling two normally-distributed populations that have the same mean but different variances (e.g. normally distributed scores for schizophrenic and not schizophrenic participants) could result in distributions that are very similar to the normal curve, but with thicker tails. As another example, when assessing a wellness score for the general population, data may be sampled from a left-skewed distribution, because most people are probably not depressed (see Heun et al., 1999). In this case, people who suffer from depression and people who do not suffer from depression are part of the same population, which can leads to asymmetry in the distribution.

Third, bounded measures can also explain non-normal distributions. For example, response time can be very large, but never below zero, which results in right-skewed distributions. In sum, there are many common situations in which normally distributed data is an unlikely assumption.

Homogeneity of Variances Assumption

Homogeneity of variances (or homoscedasticity) is a mathematical requirement that is also ecologically unlikely (Erceg-Hurn & Mirosevich, 2008; Grissom, 2000). In a previous paper (Delacre, Lakens & Leys, 2017), we identified three different causes of heteroscedasticity: the variability inherent to the use of measured variables, the variability induced by quasi-experimental treatments on measured variables, and the variability induced by different experimental treatments on randomly assigned subjects. One additional source of variability is the presence of unidentified moderators (Cohen et al., 2013).

First, psychologists, as many scholars from various fields in human sciences, often use measured variables (e.g. age, gender, educational level, ethnic origin, depression level, etc.) instead of random assignment to conditions. Prior to any treatment, parameters of pre-existing groups can vary largely from one population to another, as suggested by Henrich, Heine, and Norenzayan (2010). For example, Green, Deschamps, and Páez (2005) have shown that the scores of competitiveness, self-reliance and interdependence are more variable in some ethnic groups than in others. This stands true for many pre-existing groups such as gender, cultures, or religions and for various outcomes (see for example Adams et al., 2014; Beilmann et al., 2014; Church et al., 2012; Cohen & Hill, 2007; Haar et al., 2014; Montoya & Briggs, 2013). Moreover, groups are sometimes defined with the intention to have different variabilities. For example, as soon as a selective school admits its students based on the results of aptitude tests,

the variability will be smaller compared to a school that accepts all students.

Second, a quasi-experimental treatment can have different impacts on variances between pre-existing groups, that can even be of theoretical interest. For example, in the field of linguistics and social psychology, Wasserman and Weseley (2009) investigated the impact of language gender structure on sexist attitudes of women and men. They tested differences between sexist attitude scores of subjects who read a text in English (i.e. a language without grammatical gender) or in Spanish (i.e. a language with grammatical gender). The results showed that (for a reason not explained by the authors), the women's score on the sexism dimension was more variable when the text was read in Spanish than in English ($SD_{spanish} = .80 > SD_{english} = .50$). For men, the reverse was true ($SD_{spanish} = .97 < SD_{english} = 1.33$).[1]

Third, even when the variances of groups are the same before treatment (due to a complete succesful randomization in group assignment), unequal variances can emerge later, as a consequence of an experimental treatment (Box, 1954; Bryk & Raudenbush, 1988; Cumming, 2005; Erceg-Hurn & Mirosevich, 2008; Keppel & Wickens, 2004). For example, Koeser and Sczesny (2014) have compared arguments advocating either masculine generic or gender-fair language with control messages in order to test the impact of these conditions on the use of gender-fair wording (measured as a frequency). They report that the standard deviations increase after treatment in all experimental conditions.

Consequences of Assumption Violations

Assumptions violations would not be a matter per se, if the *F*-test was perfectly robust against departures from them (Glass et al., 1972). When performing a test, two types of errors can be made: Type I errors and Type II errors. A Type I error consists of falsely rejecting the null hypothesis in favour of an alternative hypothesis, and the Type I error rate (α) is the proportion of tests that, when sampling many times from the same population, reject the null hypothesis when there is no true effect in the population. A Type II error consists of failing to reject the null hypothesis, and the Type II error rate (β) is the proportion of tests, when sampling many times from the same population, that fail to reject the null hypothesis when there is a true effect. Finally, the statistical power ($1 - \beta$) is the proportion of tests, when sampling many times from the same population, that correctly reject the null hypothesis when there is a true effect in the population.

Violation of the Normality Assumption

Regarding the Type I error rate, the shape of the distribution has very little impact on the *F*-test (Harwell et al., 1992). When departures are very small (i.e. a kurtosis between 1.2 and 3 or a skewness between –0.4 and 0.4), the Type I error rate of the *F*-test is very close to expectations, even with sample sizes as small as 11 subjects per group (Hsu & Feldt, 1969).

Regarding the Type II error rate, many authors underlined that departures from normality do not seriously affect the power (Boneau, 1960; David & Johnson, 1951; Glass et al., 1972; Harwell et al., 1992; Srivastava, 1959; Tiku, 1971). However, we can conclude from Srivastava (1959) and Boneau (1960) that kurtosis has a slightly larger impact on the power than skewness. The effect of non-normality on power increases when sample sizes are unequal between groups (Glass et al., 1972). Lastly the effect of non-normality decreases when sample sizes increase (Srivastava, 1959).

Violation of Homogeneity of Variances Assumption

Regarding the Type I error rate, the *F*-test is sensitive to unequal variances (Harwell et al., 1992). More specifically, the more unequal the *SD* of the population's samples are extracted from, the higher the impact. When there are only two groups, the impact is smaller than when there are more than two groups (Harwell et al., 1992). When there are more than two groups, the *F*-test becomes more liberal, meaning that the Type I error rate is larger than the nominal alpha level, even when sample sizes are equal across groups (Tomarken & Serlin, 1986). Moreover, when sample sizes are unequal, there is a strong effect of the sample size and variance pairing. In case of a positive pairing (i.e. the group with the larger sample size also has the larger variance), the test is too conservative, meaning that the Type I error rate of the test is lower than the nominal alpha level, whereas in case of a negative pairing (i.e. the group with the larger sample size has the smaller variance), the test is too liberal (Glass et al., 1972; Nimon, 2012; Overall et al., 1995; Tomarken & Serlin, 1986).

Regarding the Type II error rate, there is a small impact of unequal variances when sample sizes are equal (Harwell et al., 1992), but there is a strong effect of the sample size and variance pairing (Nimon, 2012; Overall et al., 1995). In case of a positive pairing, the Type II error rate increases (i.e. the power decreases), and in case of a negative pairing, the Type II error decreases (i.e. the power increases).

Cumulative Violation of Normality and Homogeneity of Variance

Regarding both Type I and Type II error rates, following Harwell et al. (1992), there is no interaction between normality violations and unequal variances. Indeed, the effect of heteroscedasticity is relatively constant regardless of the shape of the distribution.

Based on mathematical explanations and Monteo Carlo simulations we chose to compare the *F*-test with the *W*-test and *F**-test and to exclude the James' second-order and Alexander-Govern's test because the latter two yield very similar results to the *W*-test, but are less readily available in statistical software packages. Tomarken and Serlin (1986) have shown that from the available alternatives, the *F**-test and the *W*-test perform best, and both tests are available in SPSS, which is widely used software in the psychological sciences (Hoekstra et al., 2012). For a more extended description of the James' second-order and Alexander-Govern's test, see Schneider and Penfield (1997).

The Mathematical Differences Between the *F*-test, *W*-test, and *F**-test

The mathematical differences between the *F*-test, *W*-test and *F**-test can be explained by focusing on how standard deviations are pooled across groups. As shown in (1) the *F* statistic is calculated by dividing the inter-group variance by a pooled error term, where s_j^2 and n_j are respectively the variance estimates and the sample sizes from each independent group, and where k is the number of independent groups:

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

The degrees of freedom in the numerator (2) and in the denominator (3) of the *F*-test are computed as follows:

$$df_n = k-1 \quad (2)$$

$$df_d = N-k, \quad (3)$$

With $N = \sum_{j=1}^{k} n_j$. As a generalization of the Student's *t*-test, the *F*-test is calculated based on a pooled error term. This implies that all samples are considered as issued from a common population variance (hence the assumption of homoscedasticity). When there is heteroscedasticity, and if the larger variance is associated with the larger sample size, the error term, which is the denominator in (1), is overestimated. The *F*-value is therefore smaller, leading to fewer significant findings than expected, and the *F*-test is too conservative. When the larger variance is associated with the smaller sample size the denominator in (1) is underestimated. The *F*-value is then inflated, which yields more significant results than expected.

The *F** statistic proposed by Brown and Forsythe (1974) is computed as follows:

$$F^* = \frac{\sum_{j=1}^{k}\left[n_j\left(\bar{X}_j - \bar{X}_{..}\right)^2\right]}{\sum_{j=1}^{k}\left[\left(1-\frac{n_j}{N}\right)s_j^2\right]} \quad (4)$$

Where x_j and s_j^2 are respectively the group mean and the group variance, and $\bar{X}_{..}$ is the overall mean. As it can be seen in (4) the numerator of the *F** statistic is equal to the sum of squares between groups (which is equal to the numerator of the *F* statistic when one compares two groups). In the denominator, the variance of each group is weighted by 1 minus the relative frequency of each group. This adjustment implies that the variance associated with the group with the smallest sample size is given more weight compared to the *F*-test. As a result, when the larger variance is associated with the larger sample size, *F** is larger than *F*, because the denominator decreases, leading to more significant findings compared to the *F*-test. On the other hand, when the larger variance is associated with the smaller sample size, *F** is smaller than *F*, because the denominator increases, leading to fewer significant findings compared to the *F*-test. The degrees of freedom in the numerator and in the denominator of *F**-test are computed as follows (with the same principle as the denominator computation of the *F** statistic):

$$df_n = k-1 \quad (5)$$

$$df_d = \frac{1}{\sum_{j=1}^{k}\left[\frac{\left(\frac{\left(1-\frac{n_j}{N}\right)s_j^2}{\sum_{j=1}^{k}\left[\left(1-\frac{n_j}{N}\right)s_j^2\right]}\right)^2}{n_j-1}\right]} \quad (6)$$

Formula (7) provides the computation of the *W*-test, or Welch's *F*-test. In the numerator of the *W*-test the squared deviation between group means and the general mean are weighted by $\frac{n_j}{s_j^2}$ instead of n_j (Brown & Forsythe, 1974). As a consequence, for equal sample sizes, the group with the highest variance will have smaller weight (Liu, 2015).

$$W = \frac{\frac{1}{k-1}\sum_{j=1}^{k}\left[w_j\left(\bar{X}_j - \bar{X}'\right)^2\right]}{1+\frac{2(k-2)}{k^2-1}\sum_{j=1}^{k}\left[\left(\frac{1}{n_j-1}\right)\left(1-\frac{w_j}{w}\right)^2\right]} \quad (7)$$

where:

$$w_j = \frac{n_j}{s_j^2}$$

$$w = \sum_{j=1}^{k}\left(\frac{n_j}{s_j^2}\right)$$

$$\bar{X}' = \frac{\sum_{j=1}^{k}\left(w_j\bar{X}_j\right)}{w}$$

The degrees of freedom of the *W*-test are approximated as follows:

$$df_n = k-1 \quad (8)$$

$$df_d = \frac{k^2-1}{3\sum_{j=1}^{k}\left[\frac{(1-\frac{w_j}{w})^2}{n_j-1}\right]} \quad (9)$$

When there are only two groups to compare, the *F**-test and *W*-test are identical (i.e., they have exactly the same statistical value, degrees of freedom and significance). However, when there are more than two groups to compare, the tests differ. In the appendix we illustrate the calculation of all three statistics in detail for a fictional three-group design for educational purposes.

Monte Carlo simulations: *F*-test versus *W*-test versus *F**-test

We performed Monte Carlo simulations using R (version 3.5.0) to assess the Type I and Type II error rates for the three tests. One million datasets were generated for 3840 scenarios that address the arguments present in the literature. In 2560 scenarios, means were equal across all groups (i.e. the null hypothesis is true), in order to assess the Type I error rate of the tests. In 1280 scenarios, there were differences between means (i.e. the alternative hypothesis is true) in order to assess the power of the tests. In all scenarios, when using more than 2 samples, all samples but one was generated from the same population, and only one group had a different population mean.

Population parameter values were chosen in order to illustrate the consequences of factors known to play a key role on both the Type I error rate and the statistical power when performing an ANOVA. Based on the literature review presented above, we manipulated the number of groups, the sample sizes, the sample size ratio (n-ratio = $\frac{n_k}{n_j}$), the *SD*-ratio (*SD*-ratio = $\frac{\sigma_k}{\sigma_j}$), and the sample size and variance pairing. In our scenarios, the number of compared groups (k) varied from 2 to 5. Sample sizes of k-1 groups (n_j) were 20, 30, 40, 50, or 100. The sample size of the last group was a function of the n-ratio, ranging from 0.5 to 2, in steps of 0.5. The simulations for which the n-ratio equals 1 are known as a balanced design (i.e. sample sizes are equal across all groups). The *SD* of the population from which was extracted last group was a function of the *SD*-ratio, with values of 0.5, 1, 2 or 4. The simulations for which the *SD*-ratio equals 1 are the particular case of homoscedasticity (i.e. equal variances across groups).

All possible combinations of n-ratio and *SD*-ratio were performed in order to distinguish positive pairings (the group with the largest sample size is extracted from the population with the largest *SD*), negative pairings (the group with the smallest sample size is extracted from the population with the smallest *SD*), and no pairing (sample sizes and/or population *SD* are equal across all groups). All of those conditions were tested with normal and non-normal distributions. When two groups are compared, conclusions for the three ANOVA tests (*F*, *F**, *W*) should yield identical error rates when compared to their equivalent *t*-tests (the *F*-test is equivalent to Student's *t*-test, and the *F**-test and *W*-test are equivalent to Welch's *t*-test; Delacre et al., 2017). When there are more than three groups, the *F*-test becomes increasingly liberal as soon as the variances of the distributions in each group are not similar, even when sample sizes are equal between groups (Harwell et al., 1992; Quensel, 1947).

For didactic reasons, we will report only the results where we compared three groups (k = 3). Increasing the number of groups increases how liberal all tests are. For interested readers, all figures for cases where we compare more than three groups are available here: https://osf.io/h4ks8/. Overall, the larger the sample sizes, the less the distributions of the population underlying the samples impact the robustness of the tests (Srivastava, 1959). However, increasing the sample sizes does not improve the robustness of the test when there is heteroscedasticity. Interested reader can see all details in the following Excel spreadsheet, available on github: « Type I error rate.xlsx ».

In sum, the simulations grouped over different sample sizes yield 9 conditions based on the n-ratio, *SD*-ratio, and sample size and variance pairing, as summarized in **Table 1**.

In all Figures presented below, averaged results for each sub-condition are presented under seven different configurations of distributions, using the following legend.

Type I Error Rate of the *F*-test, *W*-test, and *F**-test

As previously mentioned, the Type I error rate (α) is the long-run frequency of observing significant results when the null-hypothesis is true. When means are equal across all groups the Type I error rate of all test should be equal to the nominal alpha level. We assessed the Type I error rate of the *F*-test, *W*-test and *F**-test under 2560 scenarios using a nominal alpha level of 5%.

When there is no difference between means, the nine cells of **Table 1** simplify into five sub-conditions:

- Equal n and *SD* across groups (a)
- Unequal n but equal *SD* across groups (b and c)
- Unequal *SD* but equal n across groups (d and g)
- Unequal n and *SD* across groups, with positive correlation between n and *SD* (e and i)
- Unequal n and *SD* across groups, with negative correlation between n and *SD* (f and h)

Table 1: 9 conditions based on the n-ratio, *SD*-ratio, and sample size and variance pairing.

		***n*-ratio**		
		1	**>1**	**<1**
SD-ratio	1	a	b	c
	>1	d	e	f
	<1	g	h	i

Note: The n-ratio is the sample size of the last group divided by the sample size of the first group. When all sample sizes are equal across groups, the n-ratio equals 1. When the sample size of the last group is higher than the sample size of the first group, n-ratio >1, and when the sample size of the last group is smaller than the sample size of the first group, n-ratio <1. *SD*-ratio is the population *SD* of the last group divided by the population *SD* of the first group. When all samples are extracted from populations with the same *SD*, the *SD*-ratio equals 1. When the last group is extracted from a population with a larger *SD* than all other groups, the *SD*-ratio >1. When the last group is extracted from a population with a smaller *SD* than all other groups, the *SD*-ratio <1.

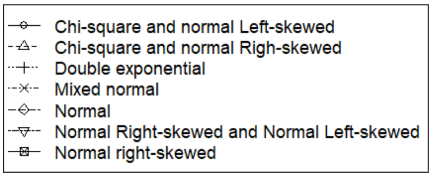


Figure 1: Legend.

In **Figures 2** to **6** (see **Figure 1** for the legend), we computed the average Type I error rate of the three tests under these five subcategories. The light grey area corresponds to the liberal criterion from Bradley (1978), who regards a departure from the nominal alpha level as acceptable whenever the Type I error rate falls within the interval [0.5 × α; 1.5 × α]. The dark grey area corresponds to the more conservative criterion from which departures from the nominal alpha is considered negligible as long as the Type I error rate falls within the interval [0.9 × α; 1.1 × α].

In **Figures 2** and **3** (cells a, b, and c in **Table 1**), the population variance is equal between all groups, so the homoscedasticity assumption is met. The *F*-test and *F**-test only marginally deviate from the nominal 5%, regardless of the underlying distribution and the *SD*-ratio. The *W*-test also only marginally deviates from the nominal 5%, except under asymmetry (the tests becomes a little more liberal) or extremely heavy tails (the test becomes a bit more conservative), consistently with observations in Harwell et al. (1992). However, deviations don't exceed the liberal criterion of Bradley (1978).

In **Figures 4, 5** and **6** (cells d to i, **Table 1**) the population variance is unequal between groups, so that the homoscedasticity assumption is not met. When sample sizes are equal across groups (**Figure 4**) and when there is a positive correlation between sample sizes and *SDs* (**Figure 5**), the Type I error rate of the *W*-test is closer to the nominal 5% than the Type I error rate of the *F**-test and the *F*-test, the latter which is consistently at the lower limit of the liberal interval suggested by Bradley, in line with Harwell et al. (1992), Glass et al. (1972), Nimon (2012) and Overall et al. (1995). Heteroscedasticity does not impact the Type I error rate of the *W*-test, regardless of the distribution (the order of the distribution shape remains the same in all conditions).

When there is a negative correlation between sample sizes and *SDs* (**Figure 6**), the Type I error rate of the *F**-test is slightly closer of the nominal 5% than the Type I error rate of the *W*-test, for which the distributions (more specifically, the skewness) has a larger impact on the Type I error rate than when there is homoscedasticity. This is consistent with conclusions of Lix et al. (1996) about

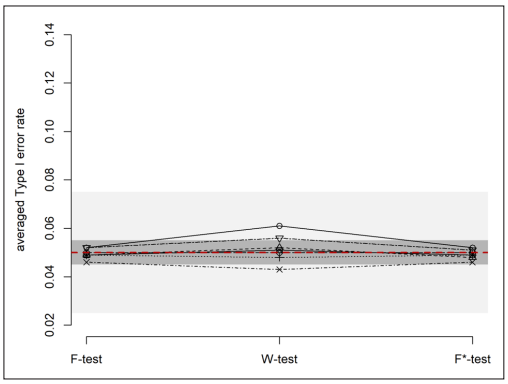


Figure 2: Type I error rate of the F-test, W-test and F*-test when there are equal SDs across groups and equal sample sizes (cell a in Table 1).

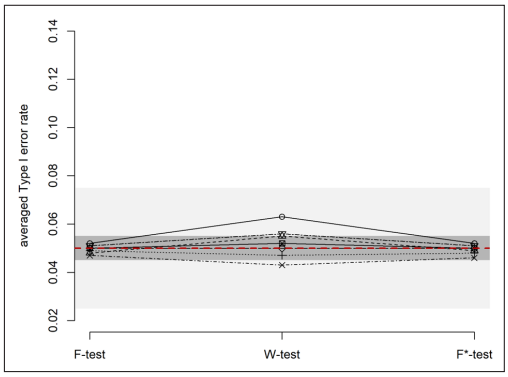


Figure 3: Type I error rate of the F-test, W-test and F*-test when there are equal SDs across groups and unequal sample sizes (cells b and c in Table 1).

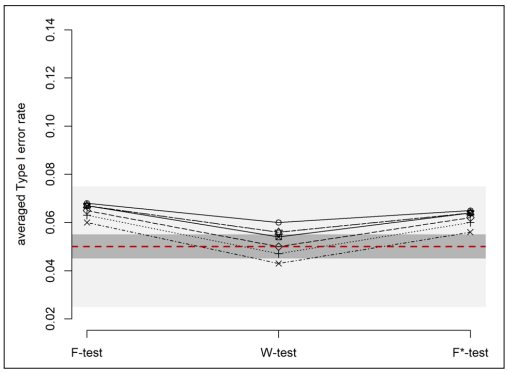


Figure 4: Type I error rate of the F-test, W-test and F*-test when there are unequal SDs across groups and equal sample sizes (cells d and g in Table 1).

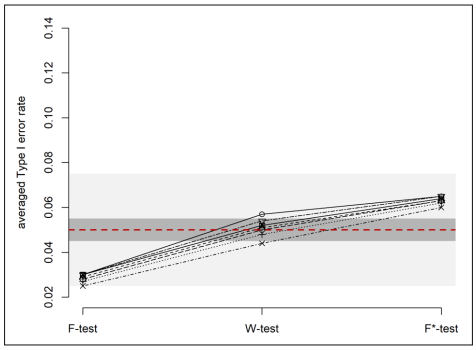


Figure 5: Type I error rate of the F-test, W-test and F*-test when there are unequal SDs across groups, and positive correlation between sample sizes and SDs (cells e and i in Table 1).

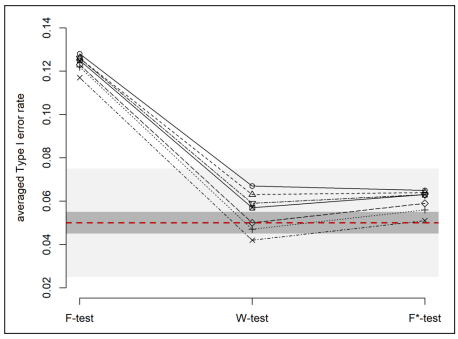


Figure 6: Type I error rate of the F-test, W-test and F*-test when there are unequal SDs across groups, and negative correlation between sample sizes and SDs (cells f and g in Table 1).

the Alexander-Govern and the James' second order tests (which return very similar results as the *W*-test, as we already mentioned). However, both tests still perform relatively well, contrary to the *F*-test that is much too liberal, in line with observations by Harwell et al. (1992), Glass et al. (1972), Nimon (2012) and Overall et al. (1995).

Conclusions

We can draw the following conclusions for the Type I error rate:

1) When all assumptions are met, all tests perform adequately.
2) When variances are equal between groups and distributions are not normal, the *W*-test is a little less efficient than both the *F*-test and the *F**-test, but departures from the nominal 5% Type I error rate never exceed the liberal criterion of Bradley (1978).
3) When the assumption of equal variances is violated, the *W*-test clearly outperforms both the *F**-test (which is more liberal) and the *F*-test (which is either more liberal or more conservative, depending on the *SDs* and *SD* pairing).
4) The last conclusion generally remains true when both the assumptions of equal variances and normality are not met.

Statistical power for the *F*-test, *W*-test, and *F**-test

As previously mentioned, the statistical power ($1 - \beta$) of a test is the long-run probability of observing a statistically significant result when there is a true effect in the population. We assessed the power of the *F*-test, *W*-test and *F**-test under 1280 scenarios, while using the nominal alpha level of 5%. In all scenarios, the last group was extracted from a population that had a higher mean than the population from where were extracted all other groups ($\mu_k = \mu_j + 1$). Because of that, in some scenarios there is a positive correlation between the *SD* and the mean (i.e. the last group has the largest *SD* and the largest mean) and in other scenarios, there is a negative correlation between *SD* and the mean (i.e. the last group has the smallest *SD* and the largest mean). As we know that the correlation between the *SD* and the mean matters for the *W*-test (see Liu, 2015), the 9 sub-conditions in **Table 1** were analyzed separately.

We computed two main outcomes: the consistency and the power. The consistency refers to the relative difference between the observed power and the nominal power, divided by the expected power:

$$Consistency = \frac{0 - E}{E} \tag{10}$$

When consistency equals zero, the observed power is consistent with the nominal power (under the parametric assumptions of normality and homoscedasticity); a negative consistency shows that the observed power is lower than the expected power; and a positive consistency shows that the observed power is higher than the expected power.

In **Figures 7, 8** and **9** (cells a, b, and c in **Table 1** see **Figure 1** for the legend), the population variance is equal between all groups, meaning that the homoscedasticity assumption is met. When distributions are normal,

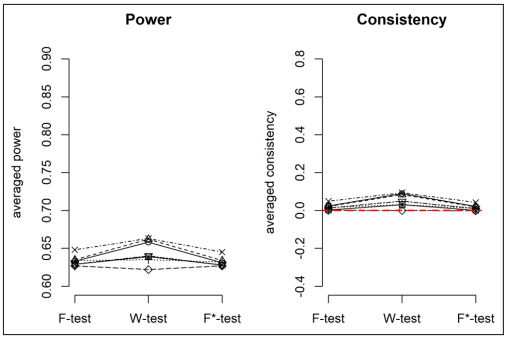


Figure 7: Power and consistency of the F-test, W-test and F*-test when there are equal SDs across groups and equal sample sizes (cell a in Table 1).

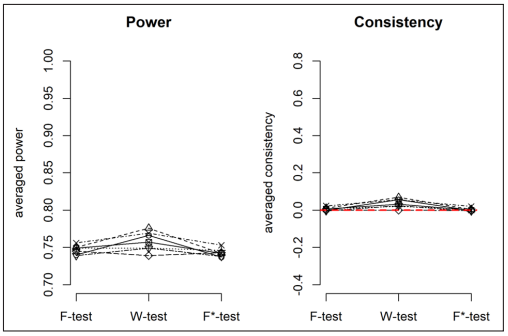


Figure 8: Power and consistency of the F-test, W-test and F*-test when there are equal SDs across groups, and positive correlation between sample sizes and means (cell b in Table 1).

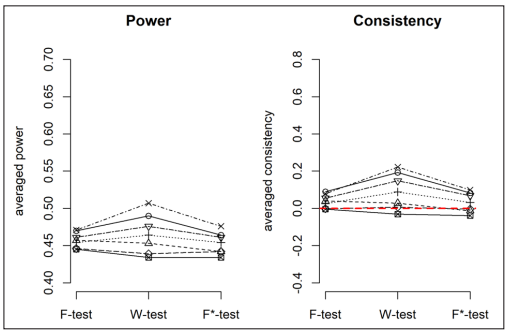


Figure 9: Power and consistency of the F-test, W-test and F*-test when there are equal SDs across groups, and negative correlation between sample sizes and means (cell c in Table 1).

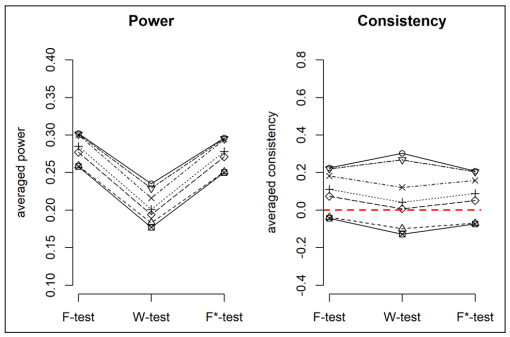


Figure 10: Power and consistency of the F-test, W-test and F*-test when there are unequal SDs across groups, positive correlation between SDs and means, and equal sample sizes across groups (cell d in Table 1).

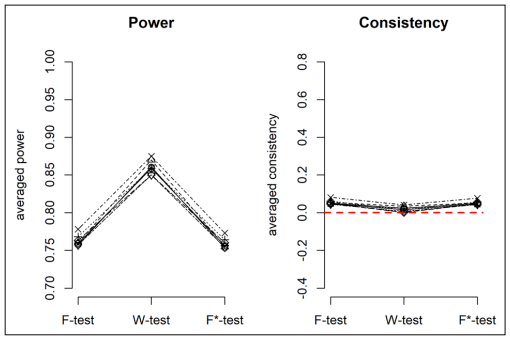


Figure 11: Power and consistency of the F-test, W-test and F*-test when there are unequal SDs across groups, negative correlation between SDs and means, and equal sample sizes across groups (cell g in Table 1).

the *W*-test is slightly less powerful than the *F*-test and *F**-test, even though differences are very small. With all other distributions, the *W*-test is generally more powerful than the *F**-test and *F*-test, even with heavy-tailed distributions, which is in contrast with previous findings (Wilcox, 1998). Wilcox (1998) concluded that there is a loss of power when means from heavy-tailed distributions (e.g. double exponential or a mixed normal distribution) are compared to means from normal distributions. This finding is based on the argument that heavy-tailed distributions are associated with bigger standard deviations than normal distributions, and that the effect size for such distributions is therefore smaller (Wilcox, 2011). However, this conclusion is based on a common conflation of kurtosis and the standard deviation, which are completely independent (DeCarlo, 1997). One can find distributions that have similar *SD* but different kurtosis (see Appendix 2). However, while the *W*-test is more powerful than the *F*-test and the *F**-test in many situations, it is a bit less consistent with theoretical expectations than both other tests in the sense that the *W*-test is generally more powerful than expected (especially with high kurtosis, or when asymmetries go in opposite directions). This is due to the fact that the *W*-test is more impacted by the distribution shape, in line with observations by Harwell et al. (1992). Note that differences between *W*-test and other tests, in terms of consistency, are very small.

In **Figures 10** to **15** (cells d to i in **Table 1** see **Figure 1** for the legend), the population variance is unequal between groups, meaning that the homoscedasticity assumption is not met. When sample sizes are equal across groups (**Figures 10** and **11**), the *F*-test and the *F**-tests are equally powerful, and have the same consistency, whatever the correlation between the *SD* and the mean. On the other hand, the power of the *W*-test depends on the correlation between the *SD* and the mean (in line with Liu, 2015). When the group with the largest mean has the largest variance (**Figure 10**), the largest deviation between group means and the general mean is given less weight, and as a consequence the *W*-test is less powerful than both other tests. At the same time, the test is slightly less consistent than both other tests. When the group with the largest mean has the smallest variance (**Figure 11**), the largest deviation between group means and the general mean is given more weight, and therefore the *W*-test is more powerful than both other tests. The test is also slightly more consistent than both other tests.

When sample sizes are unequal across groups, the power of the *F**-test and the *F*-test are a function of the correlation between sample sizes and *SDs*. When there is a negative correlation between sample sizes and *SDs* (**Figures 12** and **13**), the *F*-test is always more powerful than the *F**-test. Indeed, as was explained in the previous mathematical section, the *F*-test gives more weight to the smallest variance (the statistic is therefore increased) while the *F**-test gives more weight to the largest variance (the statistic is therefore decreased). Conversely, when there is a positive correlation between sample sizes and *SDs* (**Figures 14** and **15**), the *F*-test is always more conservative than the *F**-test, because the *F*-test gives more weight to the largest variance while the *F**-test gives more weight to the smallest variance.

The power of the *W*-test is not a function of the correlation between sample sizes and *SDs*, but rather a function of the correlation between *SDs* and means. The test is more powerful when there is a negative correlation between *SDs* and means, and less powerful when there is a positive

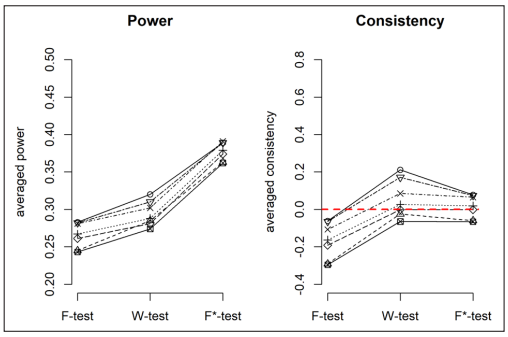


Figure 12: Power and consistency of the F-test, W-test and F*-test when there are unequal SDs across groups, negative correlation betwen sample sizes and SDs, and positive correlation between SDs and means (cell f in Table 1).

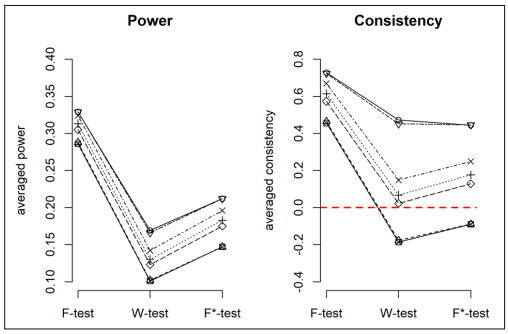


Figure 13: Power and consistency of the F-test, W-test and F*-test when there are unequal SDs across groups, negative correlation betwen sample sizes and SDs, and negative correlation between SDs and means (cell h in Table 1).

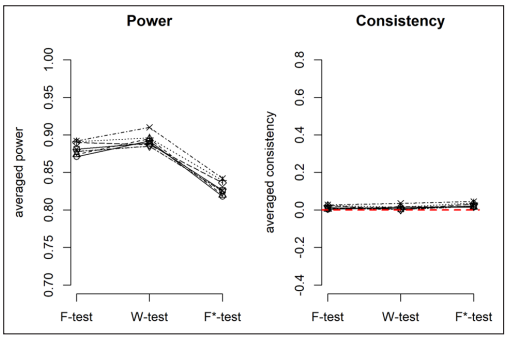


Figure 14: Power and consistency of the F-test, W-test and F*-test when there are unequal SDs across groups, positive correlation betwen sample sizes and SDs, and positive correlation between SDs and means (cell e in Table 1).

correlation between *SDs* and means. Note that for all tests, the effect of heteroscedasticity is approximately the same regardless of the shape of the distribution. Moreover,

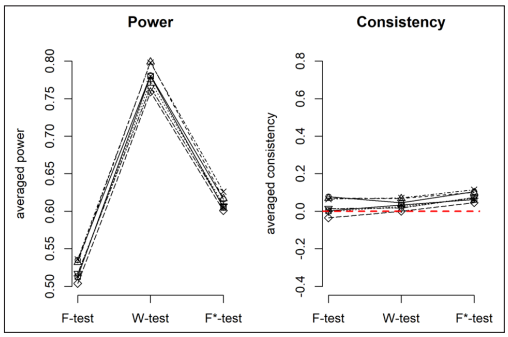


Figure 15: Power and consistency of the F-test, W-test and F*-test when there are unequal SDs across groups, positive correlation betwen sample sizes and SDs, and negative correlation between SDs and means (cell i in Table 1).

there is one constant observation in our simulations: whatever the configuration of the *n*-ratio, the consistency of the three tests is closer to zero when there is a negative correlation between the *SD* and the mean (meaning that the group with the highest mean has the lower variance).

We can draw the following conclusions about the statistical power of the three tests:

1) When all assumptions are met, the *W*-test falls slightly behind the *F*-test and the *F**-test, both in terms of power and consistency.
2) When variances are equal between groups and distributions are not normal, the *W*-test is slightly more powerful than both the *F*-test and the *F**-test, even with heavy-tailed distributions.
3) When the assumption of equal variances is violated, the *F*-test is either too liberal or too conservative, depending on the correlation between sample sizes and *SDs*. On the other side, the *W*-test is not influenced by the sample sizes and *SDs* pairing. However, it is influenced by the *SD* and means pairing.
4) The last conclusion generally remains true when both assumptions of equal variances and normality are not met.

Recommendations

Taking both the effects of the assumption violations on the alpha risk and on the power, we recommend using the *W*-test instead of the *F*-test to compare groups means. The *F*-test and *F**-test should be avoided, because a) the equal variances assumption is often unrealistic, b) tests of the equal variances assumption will often fail to detect differences when these are present, c) the loss of power when using the *W*-test is very small (and often even negligible), and d) the gain in Type I error control is considerable under a wide range of realistic conditions. Also, we recommend the use of balanced designs (i.e. same sample sizes in each group) whenever possible. When using the *W*-test, the Type I error rate is a function of criteria such as the skewness of the distributions, and whether skewness is combined with unequal variances and unequal samples sizes between groups. Our simulations show that the Type

I error rate control is in general slightly better with balanced designs.

Note that the *W*-test suffers from limitations and cannot be used in all situations. First, as previously mentioned, *W*-test, as all tests based on means, does not allow researchers to compare other relevant parameters of a distribution than the mean. For these reason, we recommend to never neglect the descriptive analysis of the data. A complete description of the shape and characteristics of the data (e.g. histograms and boxplots) is important. When at least one statistical parameter relating to the shape of the distribution (e.g. variance, skewness, kurtosis) seems to vary between groups, comparing results of the *W*-test with results of a nonparametric procedure is useful in order to better understand the data. Second, with small sample sizes (i.e. less than 50 observations per group when comparing at most four groups, 100 observations when comparing more than four groups), the *W*-test will not control Type I error rate when skewness is present and detecting departures for normality is therefore especially important in small samples. Unless you have good reasons to believe that distributions underlying the data have small kurtosis and skewness, we recommend to avoid alternative tests that are based on means comparison, in favour of alternatives such as the trimmed means test (Erceg-Hurn & Mirosevich, 2008)[2] or nonparametric tests. For more information about robust alternatives that are based on other parameters than the mean, see Erceg-Hurn and Mirosevich (2008).

Notes

[1] Note that this is a didactic example, the differences have not been tested and might not differ statistically.

[2] The null hypothesis of the trimmed means test assumes that trimmed means are the same between groups. A trimmed mean is a mean computed on data after removing the lowest and highest values of the distribution. Trimmed means and means are equal when data are symmetric. On the other hand, when data are asymmetric, trimmed means and means differ.

Additional File

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

- **Supplemental Materials.** A numerical example of the mathematical development of the F-test, W-test, and F*-test (Appendix 1) and justification for the choice of distributions in simulation (Appendix 2). DOI: https://doi.org/10.5334/irsp.198.s1

Competing Interests

The authors have no competing interests to declare.

Author Contribution

The first author performed simulations. The first, second and fourth authors contributed to the design. All authors contributed to the writing and the review of the literature. The Supplemental Material, including the full R code for the simulations and plots can be obtained from https://github.com/mdelacre/W-ANOVA. This work was supported by the Netherlands Organization for Scientific Research (NWO) VIDI grant 452-17-013. The authors declare that they have no conflicts of interest with respect to the authorship or the publication of this article.

References

Adams, B. G., Van de Vijver, F. J., de Bruin, G. P., & **Bueno Torres, C.** (2014). Identity in descriptions of others across ethnic groups in South Africa. *Journal of Cross-Cultural Psychology*, *45*(9), 1411–1433. DOI: https://doi.org/10.1177/0022022114542466

Beilmann, M., Mayer, B., Kasearu, K., & **Realo, A.** (2014). The relationship between adolescents' social capital and individualism-collectivism in Estonia, Germany, and Russia. *Child Indicators Research*, *7*(3), 589–611. DOI: https://doi.org/10.1007/s12187-014-9232-z

Boneau, C. (1960). The effects of violations of assumptions underlying the t test. *Psychological Bulletin*, *57*(1), 49–64. DOI: https://doi.org/10.1037/h0041412

Box, G. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, i. Effect of inequality of variance in the one-way classification. *The Annals of Mathematical Statistics*, *25*(2), 290–302. DOI: https://doi.org/10.1214/aoms/1177728786

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*(2), 144–152. DOI: https://doi.org/10.1111/j.2044-8317.1978.tb00581.x

Brown, M. B., & **Forsythe, A. B.** (1974). Robust tests for the equality of variances. *Journal of the American Statistical Association*, *69*(346), 364–367. DOI: https://doi.org/10.2307/2285659

Bryk, A. S., & **Raudenbush, S. W.** (1988). Heterogeneity of variance in experimental studies: A challenge to conventional interpretations. *Psychological Bulletin*, *104*(3), 396–404. DOI: https://doi.org/10.1037/0033-2909.104.3.396

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. DOI: https://doi.org/10.3758/s13428-016-0814-1

Church, A. T., Willmore, S. L., Anderson, A. T., Ochiai, M., Porter, N., Mateo, N. J., Ortiz, F. A., et al. (2012). Cultural differences in implicit theories and self-perceptions of traitedness: Replication and extension with alternative measurement formats and cultural dimensions. *Journal of Cross-Cultural Psychology*, *43*(8), 1268–1296. DOI: https://doi.org/10.1177/0022022111428514

Cohen, A. B., & **Hill, P. C.** (2007). Religion as culture: Religious individualism and collectivism among American Catholics, Jews, and Protestants. *Journal of Personality*, *75*(4), 709–742. DOI: https://doi.org/10.1111/j.1467-6494.2007.00454.x

Cohen, J., Cohen, P., West, S. G., & **Aiken, L. S.** (2013). *Applied multiple regression/correlation analysis for the behavioural sciences*. Mahwah, NJ: Erlbaum. DOI: https://doi.org/10.4324/9780203774441

Cumming, G. (2005). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge.

David, F. N., & **Johnson, N. L.** (1951). The effect of non-normality on the power function of the f-test in the analysis of variance. *Biometrika*, *38*(1–2), 43–57. DOI: https://doi.org/10.2307/2332316

DeCarlo, L. T. (1997). On the meaning and use of kurtosis. *Psychological Methods*, *2*(3), 292–307. DOI: https://doi.org/10.1037//1082-989X.2.3.292

Delacre, M., Lakens, D., & **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

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), 591–601. DOI: https://doi.org/10.1037/0003-066X.63.7.591

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

Green, E. G., Deschamps, J.-C., & **Páez, D.** (2005). Variation of individualism and collectivism within and between 20 countries: A typological analysis. *Journal of Cross-Cultural Psychology*, *36*(3), 321–339. DOI: https://doi.org/10.1177/0022022104273654

Grey, S., & **Mathews, A.** (2000). Effects of training on interpretation of emotional ambiguity. *The Quarterly Journal of Experimental Psychology*, *53*(4), 1143–1162. DOI: https://doi.org/10.1080/713755937

Grissom, R. (2000). Heterogeneity of variance in clinical data. *Journal of Consulting and Clinical Psychology*, *68*(1), 155–165. DOI: https://doi.org/10.1037//0022-006X.68.1.155

Haar, J. M., Russo, M., Suñe, A., & **Ollier-Malaterre, A.** (2014). Outcomes of work-life balance on job satisfaction, life satisfaction and mental health: A study across seven cultures. *Journal of Vocational Behavior*, *85*(3), 361–373. DOI: https://doi.org/10.1016/j.jvb.2014.08.010

Harwell, M. R., Rubinstein, E. N., Hayes, W. S., & **Olds, C. C.** (1992). Summarizing Monte Carlo results in methodological research: The one- and two-factor fixed effects anova cases. *Journal of Educational Statistics*, *17*(4), 315–339. DOI: https://doi.org/10.3102/10769986017004315

Henrich, J., Heine, S. J., & **Norenzayan, A.** (2010). Most people are not weird. *Nature*, *466*, 29–29. DOI: https://doi.org/10.1038/466029a

Heun, R., Burkart, M., Maier, W., & **Bech, P.** (1999). Internal and external validity of the who well-being scale in the elderly general population. *Acta Psychiatrica Scandinavica*, *99*(3), 171–178. DOI: https://doi.org/10.1111/j.1600-0447.1999.tb00973.x

Hoekstra, R., Kiers, H. A., & **Johnson, A.** (2012). Are assumptions of well-known statistical techniques checked, and why (not)? *Frontiers in Psychology*, *3*(137), 1–9. DOI: https://doi.org/10.3389/fpsyg.2012.00137

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*(4), 515–527. DOI: https://doi.org/10.3102/00028312006004515

Keppel, G., & **Wickens, T. D.** (2004). *Design and analysis: A researcher's handbook*. Upper Saddle River, New Jersey: Prentice Hall.

Keselman, H., 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 analy. *Review of Educational Research*, *68*(3), 350–386. DOI: https://doi.org/10.3102/00346543068003350

Koeser, S., & **Sczesny, S**. (2014). Promoting gender-fair language: The impact of arguments on language use, attitudes, and cognitions. *Journal of Language and Social Psychology*, *33*(5), 548–560. DOI: https://doi.org/10.1177/0261927X14541280

Liu, H. (2015). *Comparing welch anova, a kruskal-wallis test, and traditional anova in case of heterogeneity of variance* (PhD thesis). Virginia Commonwealth University.

Lix, L. M., Keselman, J. C., & **Keselman, H.** (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance *f* test. *Review of Educational Research*, *66*(4), 579–619. DOI: https://doi.org/10.3102/00346543066004579

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*(1), 156–166. DOI: https://doi.org/10.1037/0033-2909.105.1.156

Montoya, D. Y., & **Briggs, E.** (2013). Shared ethnicity effects on service encounters: A study across three us subcultures. *Journal of Business Research*, *66*(3), 314–320. DOI: https://doi.org/10.1016/j.jbusres.2011.08.011

Nimon, K. F. (2012). Statistical assumptions of substantive analyses across the general linear model: A mini-review. *Frontiers in Psychology*, *3*(322), 1–5. DOI: https://doi.org/10.3389/fpsyg.2012.00322

Overall, J. E., Atlas, R. S., & **Gibson, J. M.** (1995). Tests that are robust against variance heterogeneity in k × 2 designs with unequal cell frequencies. *Psychological Reports*, *76*(3), 1011–1017. DOI: https://doi.org/10.2466/pr0.1995.76.3.1011

Quensel, C.-E. (1947). The validity of the *z*-criterion when the variates are taken from different normal populations. *Scandinavian Actuarial Journal*, *30*(1), 44–55. DOI: https://doi.org/10.1080/03461238.1947.10419648

Schneider, P. J., & **Penfield, D. A.** (1997). Alexander and Govern's approximations: Providing an alternative to anova under variance heterogeneity. *The Journal of Experimental Education*, *65*(3), 271–286. DOI: https://doi.org/10.1080/00220973.1997.9943459

Srivastava, A. B. L. (1959). Effects of non-normality on the power of the analysis of variance test. *Biometrika*, *46*(1–2), 114–122. DOI: https://doi.org/10.2307/2332813

Tiku, M. (1971). Power function of the f-test under non-normal situations. *Journal of the American Statistical Association*, *66*, 913–916. DOI: https://doi.org/10.1080/01621459.1971.10482371

Tomarken, A. J., & **Serlin, R. C.** (1986). Comparison of anova alternatives under variance heterogeneity and specific noncentrality structures. *Psychological Bulletin*, *99*(1), 90–99. DOI: https://doi.org/10.1037//0033-2909.99.1.90

Wasserman, B. D., & **Weseley, A. J.** (2009). ?'Qué? Quoi? Do languages with grammatical gender promote sexist attitudes? *Sex Roles*, *61*, 634–643. DOI: https://doi.org/10.1007/s11199-009-9696-3

Westfall, P. H. (2014). Kurtosis as peakedness, 1905–2014. R.I.P. *The American Statistician*, *68*(3), 191–195. DOI: https://doi.org/10.1080/00031305.2014.917055

Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist*, *53*(3), 300–314. DOI: https://doi.org/10.1037/0003-066X.53.3.300

Wilcox, R. R. (2005). Comparing medians: An overview plus new results on dealing with heavy-tailed distributions. *The Journal of Experimental Education*, *73*(3), 249–263. DOI: https://doi.org/10.3200/JEXE.73.3.249-263

Wilcox, R. R. (2011). *Introduction to robust estimation and hypothesis testing*. Cambridge, Massachusetts, US: Academic Press. DOI: https://doi.org/10.1016/B978-0-12-386983-8.00010-X

Yuan, K.-H., Bentler, P. M., & **Chan, W.** (2004). Structural equation modeling with heavy tailed distributions. *Psychometrika*, *69*(3), 421–436. DOI: https://doi.org/10.1007/BF02295644

How to cite this article: Delacre, M., Leys, C., Mora, Y. L., & Lakens, D. (2019). Taking Parametric Assumptions Seriously: Arguments for the Use of Welch's *F*-test instead of the Classical *F*-test in One-Way ANOVA. *International Review of Social Psychology*, 32(1): 13, 1–12. DOI: https://doi.org/10.5334/irsp.198

Submitted: 05 June 2018 **Accepted:** 20 May 2019 **Published:** 01 August 2019

]u[*International Review of Social Psychology* is a peer-reviewed open access journal published by Ubiquity Press. OPEN ACCESS

Note de fin de chapitre

[7]Un erratum de cet article figure dans l'annexe B.

Chapitre 4: utiliser le g_s^* de Hedges basé sur l'écart-type non poolé

Why Hedges' g_s^* based on the non-pooled standard deviation should be reported with Welch's t-test

Marie Delacre[1], Daniel Lakens[2], Christophe Ley[3], Limin Liu[3], & Christophe Leys[1]

[1] Université Libre de Bruxelles, Service of Analysis of the Data (SAD), Bruxelles, Belgium

[2] Eindhoven University of Technology, Human Technology Interaction Group, Eindhoven, the Netherlands

[3] Universiteit Gent, Department of Applied Mathematics, Computer Science and Statistics, Gent, Belgium

Author Note

A Shiny app to compute point estimators and confidence intervals based on descriptive statistics is available from https://effectsize.shinyapps.io/deffsize/. We would like to thank Aaron Caldwell for his help creating the figures in the Shiny App. Daniël Lakens was funded by VIDI Grant 452-17-013 from the Netherlands Organisation for Scientific Research.

Correspondence concerning this article should be addressed to Marie Delacre, CP191, avenue F.D. Roosevelt 50, 1050 Bruxelles. E-mail: marie.delacre@ulb.ac.be

Commentaire: l'article présenté ci-dessous n'a pas encore été accepté pour publication. Nous tenterons prochainement une nouvelle soumission, après avoir inclus une série de modifications suggérées par Geoff Cumming, conformément aux échanges reportés dans l'annexe C.

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 (Association, 2010 ; Duran et al., 2006).

In "between-subject" designs where individuals are randomly assigned into one of two independent groups and group scores are compared based on their means, the dominant estimator of effect size is Cohen's d_s, where the sample mean difference is divided by the pooled sample standard deviation (Peng, Chen, Chiang, & Chiang, 2013 ; Shieh, 2013). This estimator is available in many statistical software packages, such as SPSS and Stata. However, computing the pooled sample standard deviation assumes that both sample variances are estimates of a common population variance, which is known as the homogeneity of variance assumption. It has been widely argued that there are many fields in psychology where this assumption is ecologically unlikely (Delacre, Lakens, & Leys, 2017 ; Erceg-Hurn & Mirosevich, 2008 ; Grissom, 2000). The question how to deal with the assumption of equal variances has been widely explored in the context of hypothesis testing, and it is becoming increasingly common to by default report a t-test that does not assume equal variances, such as Welch's t-test.

However, the question which effect size to report when equal variances are not assumed has received less attention. One possible reason is that researchers have not found consensus on which of the available options should be used (Shieh, 2013). Even within the very specific context of an estimate for the standardized sample mean difference there is little agreement about which estimator is the best choice. In this article, we will review the main candidates that have been proposed in the literature in the d family of effect sizes, without (Cohen's d_s, Glass's d_s, Shieh's d_s and Cohen's d_s^*) and with correction for bias (Hedges' g_s, Glass's g_s, Shieh's g_s and Hedges' g_s^*). We provide an R package and Shiny app to compute relevant effect size measures and their confidence intervals.

Before reviewing the most important effect size measures in the d-family, we will first list the different purposes effect size measures serve, and discuss the relationship between effect sizes, statistical, and practical significance. Based on a detailed description of the good properties an effect size measure should possess, we will evaluate these properties in the Monte Carlo simulations we performed to compare the different effect size estimators with correction for

bias.

Three purposes of effect size estimators

The effect size is a measure of the magnitude of an effect. In the context of the comparison of two groups based on their means, when the null hypothesis is the absence of effect, d-family effect size estimators estimate the magnitude of the differences between parameters of two populations groups are extracted from (e.g. the mean, Peng & Chen, 2014). Such a measure can be used for three different purposes.

First, effect size measures can be used for *interpretative* purposes. They allow researchers to assess the practical significance of a result (i.e. statements about the relevance of an effect in real life). In order to assess the meaningfulness of an effect, we should be able to relate this effect size estimate with behaviors/meaningful consequences in the real world (Andersen, McCullagh, & Wilson, 2007). This typically involves an analysis of the costs (determined by a specific context) and the benefits (in part determined by the size of the effect). It is important to remember an effect size is just a mathematical indicator of the magnitude of a difference, which depends on the way a variable is converted into numerical indicator. An effect size in itself is not a measure of the importance or the relevance of an effect for real life (even if benchmarks for small, medium, or large effect sizes might have contributed to such a misinterpretation, Stout & Ruble, 1995).

Second, effect size measures can be used for *comparative* purposes. They allow researchers to assess the stability of results across designs, analyses, and sample sizes. This includes statistically comparing and combining the results from two or more studies in a meta-analysis.

Third, effect size measures can be used for *inferential* purposes. Hypothesis tests and confidence intervals based on the same statistical quantity are directly related: if the area of the null hypothesis is out of the $(1-\alpha)$-confidence interval, then the hypothesis test would also result in a p-value below the nominal alpha level. At the same time, the interval provides extra information about the precision of the sample estimate for inferential purposes (Altman, 2005 ; Ellis, 2010), and which effect sizes are excluded. The narrower the interval, the higher the precision, and the wider the confidence interval, the more the data lack precision. Effect size measures are also indirectly related to the hypothesis tests as effect sizes from previous studies can be used in an

a-priori power analysis when planning a new study (Lakens, 2013 ; Prentice & Miller, 1992 ; Stout & Ruble, 1995 ; Sullivan & Feinn, 2012 ; Wilkinson, 1999).

Properties of a good effect size estimator

The empirical value of an estimator (called the *estimate*) depends on the sample value. Different samples extracted from the same population will lead to different sample estimates of the population value. The *sampling distribution* of the estimator is the distribution of all estimates, based on all possible samples of size n extracted from one population. Studying the sampling distribution is useful, as it allows us to assess the qualities of an estimator. More specifically, three desirable properties a good estimator should possess for inferential purposes are: *unbiasedness*, *consistency* and *efficiency* (Wackerly, Mendenhall, & Scheaffer, 2008).

An estimator is unbiased if the distribution of estimates is centered around the true population parameter. On the other hand, an estimator is positively (or negatively) biased if the distribution is centered around a value that is higher (or lower) than the true population parameter (see Figure 1). In other words, examining the bias of an estimator tells us if estimates are on average accurate. The *bias* of a point estimator $\hat{\delta}$ can be computed as

$$\delta_{bias} = E(\hat{\delta}) - \delta \tag{1}$$

where $E(\hat{\delta})$ is the expectation of the sampling distribution of the estimator and δ is the true (population) parameter.

As we can see in Tables 1 and 2 the bias is directly related to the population effect size. The larger the population effect size, the larger the bias. It is therefore also interesting to examine the *relative bias*, defined as the ratio between the bias and the population effect size:

$$\delta_{relative\ bias} = \frac{E(\hat{\delta}) - \delta}{\delta} \tag{2}$$

While the bias informs us about the quality of estimates on average, in particular their capacity of lying close to the true value, it says nothing about individual estimates. Imagine a situation where the distribution of estimates is centered around the real parameter but with such a large variance that some point estimates are very far from the center. This would be problematic,

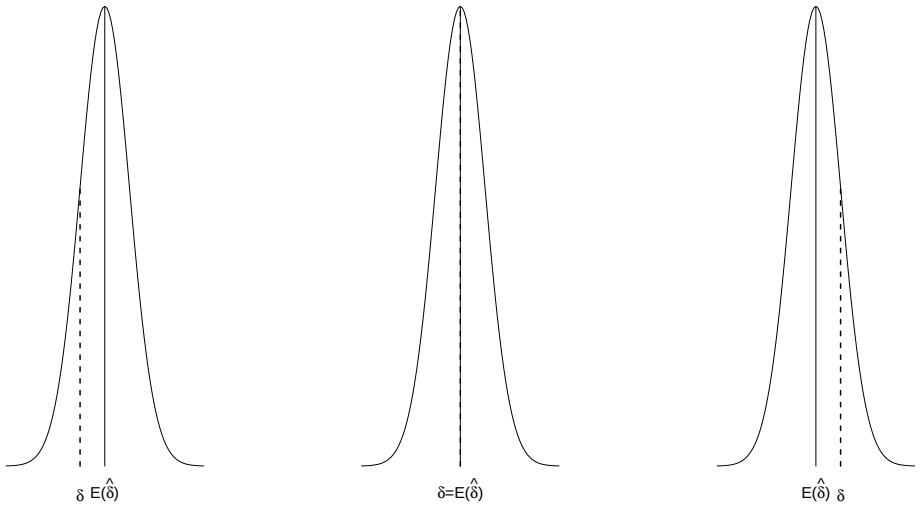


FIGURE 1 – Sampling distribution for a positively biased (left), an unbiased (center) and a negatively biased estimator (right)

since any single estimate might be very far from the true population value. Therefore it is not only essential for an estimator to be unbiased, but it is also desirable that the variability of its sampling distribution is small. Ideally, sample estimates are close to the true population parameter. Among two unbiased estimators $\hat{\delta_1}$ and $\hat{\delta_2}$, we therefore say that $\hat{\delta_1}$ is *more efficient* than $\hat{\delta_2}$ if

$$Var(\hat{\delta}_1) \leq Var(\hat{\delta}_2) \tag{3}$$

where $Var(\hat{\delta})$ is the variance of the sampling distribution of the estimator $\hat{\delta}$. Among all unbiased estimators, the more efficient estimator will be the one with the smallest variance [8]. The variance of an estimator $\hat{\delta}$ is a function of its size (the larger the estimator, the larger the variance) and, therefore, we might be interested in evaluating the *relative variance* as the ratio between the variance and the square of the population estimator:

$$relative\ var(\hat{\delta}_1) = \frac{Var(\hat{\delta})}{\delta^2} \tag{4}$$

Note that both unbiasedness and efficiency are very important when choosing an estimator. In some situations, it might be better to have a slightly biased estimator with low variance, (so that each estimate remains relatively close to the true parameter and one might be able to apply bias correction techniques) rather than an unbiased estimator with a large variance

(Raviv, 2014).

Finally, the last property of a good point estimator is *consistency*. Consistency means that the bigger the sample size, the closer the estimate is to the population parameter. In other words, the estimates *converge* to the true population parameter.

Different measures of effect sizes

The d-family effect sizes are commonly used for mean differences between groups or conditions. The population effect size is defined as

$$\delta = \frac{\mu_1 - \mu_2}{\sigma} \tag{5}$$

where both populations follow a normal distribution with mean μ_j in the j^{th} population ($j = 1, 2$) and standard deviation σ. There exist different estimators of this effect size measure. For all, the mean difference is estimated by the difference $\bar{X}_1 - \bar{X}_2$ of both sample means. When the equality of variances assumption is assumed, σ is estimated by pooling both sample standard deviations (S_1 and S_2). When the equality of variances assumption cannot be assumed, alternatives to the pooled standard deviation are available. In the next section we will present effect sizes that assume equal variances between groups (Cohen's d_s and Hedges' g_s), and effect sizes that do not assume equal variances (Glass's d_s, Shieh's d_s, Hedges' d_s^*, Glass's g_s, Shieh's g_s, and Hedges' g_s^*). For each effect size we will provide information about their theoretical bias, variance, and consistency.

When variances are equal between groups

When we have good reasons to assume equality of variances between groups then the most common estimator of δ is Cohen's d_s, where the sample mean difference is divided by a pooled error term (Cohen, 1965):

$$Cohen's\ d_s = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1-1)\times S_1^2 + (n_2-1)\times S_2^2}{n_1+n_2-2}}}$$

where S_j is the standard deviation, and n_j the sample size of the j^{th} sample ($j = 1, 2$). The reasoning behind this measure is to make use of the fact that both samples share the same population variance (H. J. Keselman, Algina, Lix, Wilcox, & Deering, 2008), which means a more accurate estimation of the population variance can be achieved by pooling both estimates of this parameter (i.e. S_1 and S_2). Since the larger the sample size, the more accurate the estimate, we give more weight to the estimate based on the larger sample size. Cohen's d_s is directly related to Student's t-statistic:

$$t_{Student} = \frac{Cohen's\ d_s}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \leftrightarrow Cohen's\ d_s = t_{Student} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{6}$$

Under the assumption of normality and equal variances between groups, Student's t-statistic follows a t-distribution with known degrees of freedom

$$df_{Student} = n_1 + n_2 - 2 \tag{7}$$

and noncentrality parameter [9]

$$ncp_{Student} = \frac{\delta_{Cohen}}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where $\delta_{Cohen} = \frac{\mu_1 - \mu_2}{\sigma_{pooled}}$ and $\sigma_{pooled} = \sqrt{\frac{(n_1-1)\times\sigma_1^2 + (n_2-1)\times\sigma_2^2}{n_1+n_2-2}}$. The relationship described in equation 6 and the theoretical distribution of Student's t-statistic allow us to determine the sampling distribution of Cohen's d_s, and therefore, its expectation and variance when the assumptions of normality and equal variances are met. All these equations are provided in Table 1. For interested readers, Supplemental Material 1 provides a detailed examination of the theoretical bias and variance of Cohen's d_s based on Table 1, as well as the bias and variance of all other estimators described later, based on Tables 2 and 3), with the goal to determine which parameters influence the bias and variance of different estimators. The main results will be discussed in the section "Monte Carlo Simulations: assessing the bias, efficiency and consistency of 5 estimators" below.

While Cohen's d_s is a consistent estimator, its bias and variance are substantial with small sample sizes, even under the assumptions of normality and equal variances (Lakens, 2013). In order to compensate for Cohen's d_s bias with small sample sizes, Hedges & Olkin (1985) defined

a bias-corrected version:

$$Hedges'\ g_s = Cohen's\ d_s \times \frac{\Gamma(\frac{df_{Student}}{2})}{\sqrt{\frac{df_{Student}}{2}} \times \Gamma(\frac{df_{Student}-1}{2})}$$

where $df_{Student}$ has been defined in equation 7 and $\Gamma()$ is the gamma function [for integers, $\Gamma(x)$ is the factorial of x minus 1: $\Gamma(x) = (x-1)!$; Goulet-Pelletier & Cousineau (2018)]. This equation can be approximated as follows:

$$Hedges'\ g_s = Cohen's\ d_s \times \left(1 - \frac{3}{4N-9}\right)$$

where N is the total sample size. Hedges' g_s is theoretically unbiased when the assumptions of normality and equal variances are met (see Table 1). Moreover, while the variance of both Cohen's d_s and Hedges' g_s depend on the same parameters (i.e. the total sample size (N) and the sample sizes ratio $\left(\frac{n_1}{n_2}\right)$), Hedges' g_s is less variable, especially with small sample sizes [10].

While the pooled error term is the best choice when variances are equal between groups (Grissom & Kim, 2001), it may not be well advised for use with data that violate this assumption (Cumming, 2013 ; Grissom & Kim, 2001, 2005 ; Kelley, 2005 ; Shieh, 2013). When variances are unequal between groups, the expression in equation 5 is no longer valid because both groups do not share a common population variance. If we pool the estimates of two unequal population variances, the estimator of effect size will be smaller as it should be in case of positive pairing (i.e. the group with the larger sample size is extracted from the population with the larger variance) and larger as it should be in case of negative pairing (i.e. the group with the larger sample size is extracted from the population with the smaller variance). Because the assumption of equal variances across populations is rarely realistic in practice (Cain, Zhang, & Yuan, 2017 ; Delacre, Lakens, & Leys, 2017 ; Delacre, Leys, Mora, & Lakens, 2019 ; Erceg-Hurn & Mirosevich, 2008 ; Gene V. Glass, Peckham, & Sanders, 1972 ; Grissom, 2000 ; Micceri, 1989 ; Yuan, Bentler, & Chan, 2004), both Cohen's d_s and Hedges' g_s should be abandoned in favor of an alternative robust to unequal population variances.

TABLE 1: Expectation, bias and variance of Cohen's d_s and Hedges' g_s under the assumptions that independent residuals are normally distributed with equal variances across groups.

	df	Expectation	Variance
Cohen's d_s	$N-2$	$\delta_{Cohen} \times c_f$	$\frac{N \times df}{n_1 n_2 \times (df-2)} + \delta^2_{Cohen}\left[\frac{df}{df-2} - c_f^2\right]$
		$\approx \frac{\delta_{Cohen}}{\left(1-\frac{3}{4N-9}\right)}$	$\approx \frac{N \times df}{n_1 n_2 \times (df-2)} + \delta^2_{Cohen}\left[\frac{df}{df-2} - \left(\frac{1}{1-\frac{3}{4N-9}}\right)^2\right]$
Hedges' g_s	$N-2$	δ_{Cohen}	$Var(Cohen's\ d_s) \times \left[\frac{\Gamma(\frac{df}{2})}{\sqrt{\frac{df}{2}} \times \Gamma(\frac{df-1}{2})}\right]^2$
			$\approx Var(Cohen's\ d_s) \times \left[1 - \frac{3}{4N-9}\right]^2$

Note. $\delta_{Cohen} = \frac{\mu_1 - \mu_2}{\sigma_{pooled}}$ and $c_f = \frac{\sqrt{\frac{df}{2}} \times \Gamma\left(\frac{df-1}{2}\right)}{\Gamma\left(\frac{df}{2}\right)}$; Cohen's d_s is a biased estimator, because its expectation differs from the population effect size. Moreover, the larger the population estimator (δ_{Cohen}), the larger the bias. Indeed, the bias is the difference between the expectation and δ_{Cohen}: $\delta_{bias} = \delta_{Cohen} \times (c_f - 1)$. On the other hand, Hedges' g_s is an unbiased estimator, because its expectation equals δ_{Cohen}; equations in Table 1 require $df \geq 3$ (i.e. $N \geq 5$).

When variances are unequal between populations

In his review, Shieh (2013) mentions three options available in the literature to deal with the case of unequal variances: (A) the Glass's d_s, (B) the Shieh's d_s and (C) the Cohen's d_s^*.

Glass's $\mathbf{d_s}$ When comparing one control group with one experimental group, G. V. Glass, McGav, & Smith (1981) recommend using the standard deviation of the control group as standardizer. This yields

$$Glass's\ d_s = \frac{\bar{X}_e - \bar{X}_c}{S_c}$$

where $\bar{X}_e$ *and* $\bar{X}_c$ are the sample means of the experimental and control groups, and S_c is the sample SD of the control group. One argument in favour of using S_c as standardizer is the fact that it is not affected by the experimental treatment. When it is easy to identify which group is the "control" one, it is therefore convenient to compare the effect size estimation of different designs studying the same effect (Cumming, 2013). However, defining this group is not always obvious (Coe, 2002). This could induce large ambiguity because depending on the chosen SD as standardizer, measures could be substantially different (Shieh, 2013). The distribution of Glass's d_s is defined as in Algina, Keselman, & Penfield (2006):

$$Glass's\ d_s \sim \sqrt{\frac{1}{n_c} + \frac{\sigma_e^2}{n_e \times \sigma_c^2}} \times t_{df,ncp} \tag{8}$$

where n_c and n_e are the sample sizes of the control and experimental groups, and df and ncp are defined as follows:

$$df = n_c - 1 \tag{9}$$

$$ncp = \frac{\delta_{Glass}}{\sqrt{\frac{1}{n_c} + \frac{\sigma_e^2}{n_e \times \sigma_c^2}}}$$

where $\delta_{Glass} = \frac{\mu_c - \mu_e}{\sigma_c}$ and μ_c and μ_e are respectively the mean of the populations control and experimental groups are extracted from. Thanks to equation 8, we can compute its theoretical expectation and variance when the assumption of normality is met (see Table 2), and therefore determine which factors influence bias and variance, and how they do so (see Supplemental Material 1).

TABLE 2: Expectation, bias and variance of Glass's d_s and Cohen's d_s^* and Shieh's d_s under the assumption that independent residuals are normally distributed.

	df	Expectation	Variance
Glass's d_s	$n_c - 1$	$\delta_{Glass} \times c_f$	$\frac{df}{df-2} \times \left(\frac{1}{n_c} + \frac{\sigma_e^2}{n_e \sigma_c^2}\right) + \delta_{Glass}^2 \left(\frac{df}{df-2} - c_f^2\right)$
Cohen's d_s^*	$\frac{(n_1-1)(n_2-1)(\sigma_1^2+\sigma_2^2)^2}{(n_2-1)\sigma_1^4+(n_1-1)\sigma_2^4}$	$\delta_{Cohen}^* \times c_f$	$\frac{df}{df-2} \times \frac{2\left(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)}{\sigma_1^2+\sigma_2^2} + (\delta_{Cohen}^*)^2 \left(\frac{df}{df-2} - c_f^2\right)$
		$\approx \delta_{Cohen}^* \times \frac{4df-1}{4(df-1)}$	$\approx \frac{df}{df-2} \times \frac{2\left(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)}{\sigma_1^2+\sigma_2^2} + (\delta_{Cohen}^*)^2 \left[\frac{df}{df-2} - \left(\frac{4\,df-1}{4(df-1)}\right)^2\right]$
Shieh's d_s	$\frac{\left(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)^2}{\frac{(\sigma_1^2/n_1)^2}{n_1-1}+\frac{(\sigma_2^2/n_2)^2}{n_2-1}}$	$\delta_{Shieh} \times c_f$	$\frac{df}{(df-2)N} + \delta_{Shieh}^2 \left(\frac{df}{df-2} - c_f^2\right)$

Note. $c_f = \frac{\sqrt{\frac{df}{2}} \times \Gamma\left(\frac{df-1}{2}\right)}{\Gamma\left(\frac{df}{2}\right)}$; $\delta_{Glass} = \frac{\mu_c - \mu_e}{\sigma_c}$, $\delta_{Shieh} = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n_1/N} + \frac{\sigma_2^2}{n_2/N}}}$ and $\delta_{Cohen}^* = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2+\sigma_2^2}{2}}}$; all estimators are biased estimators, because their expectations differ from the population effect size δ. Moreover, the larger the population estimator (δ), the larger the bias. Indeed, the bias is the difference between the expectation and δ: $\delta_{bias} = \delta \times (c_f - 1)$. Equations require $df \geq 3$ and at least 2 subjects per group.

Shieh's $\mathbf{d_s}$ Kulinskaya & Staudte (2007) were the first to recommend the use of a standardizer that takes the sample sizes allocation ratios into account, in addition to the variance of both samples. Shieh (2013), following Kulinskaya & Staudte (2007), proposed a modification of the exact *SD* of the sample mean difference:

$$Shieh's\ d_s = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_1^2/q_1 + S_2^2/q_2}}; \quad q_j = \frac{n_j}{N} (j = 1, 2)$$

where $N = n_1 + n_2$. Shieh's d_s is directly related with Welch's t-statistic:

$$Shieh's\ d_s = \frac{t_{Welch}}{\sqrt{N}} \leftrightarrow t_{welch} = Shieh's\ d_s \times \sqrt{N} \tag{10}$$

The exact distribution of Welch's t-statistic is more complicated than the exact distribution of Student's t-statistic, but it can be approximated, under the assumption of normality, by a t-distribution with degrees of freedom and noncentrality parameters (Shieh, 2013 ; Welch, 1938):

$$df_{Welch} = \frac{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^2}{\frac{(\sigma_1^2/n_1)^2}{n_1 - 1} + \frac{(\sigma_2^2/n_2)^2}{n_2 - 1}} \tag{11}$$

$$ncp_{Welch} = \delta_{Shieh} \times \sqrt{N} = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n1} + \frac{\sigma_2^2}{n_2}}}$$

where $\delta_{Shieh} = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n_1/N} + \frac{\sigma_2^2}{n_2/N}}}$. The relationship described in equation 10 and the theoretical distribution of Welch's t-statistic allow us to approximate the sampling distribution of Shieh's d_s. Based on the sampling distribution of Shieh's d_s, we can estimate its theoretical expectation and variance under the assumption of normality (see Table 2), and thereby determine which factors influence bias and variance, and how they do so (see Supplemental Material 1).

As demonstrated in the Appendix, when variances and sample sizes are equal across groups, the biases and variances of Cohen's d_s and Shieh's d_s are identical except for multiplication by a constant. The same is true for the estimators δ_{Cohen} and δ_{Shieh}:

$$\delta_{Cohen} = 2 \times \delta_{Shieh} \quad \text{(considering } \sigma_1 = \sigma_2 \text{ and } n_1 = n_2) \tag{12}$$

$$Bias_{Cohen's\ d_s} = 2 \times Bias_{Shieh's\ d_s} \quad \text{(considering } \sigma_1 = \sigma_2 \text{ and } n_1 = n_2) \tag{13}$$

$$Var_{Cohen's\ d_s} = 4 \times Var_{Shieh's\ d_s} \quad \text{(considering } \sigma_1 = \sigma_2 \text{ and } n_1 = n_2) \tag{14}$$

We can deduce from equations 12, 13 and 14 that relative to their respective population effect size, Cohen's d_s and Shieh's d_s are equally accurate. In other words, their relative bias and variance are identical. This is a good illustration of our motivation to favor relative bias and variance (previously defined in equations 2 and 4) over the most commonly used raw bias and variance (previously defined in equations 1 and 3).

When sample sizes are not equal, according to the statistical properties of Welch's statistic under heteroscedasticity, Shieh's d_s accounts for the allocation ratio of sample sizes to each condition. The lack of generality caused by taking this specificity of the design into account has led Cumming (2013) to question its usefulness in terms of interpretability: when the mean difference $(\bar{X}_1 - \bar{X}_2)$, S_1, and S_2 remain constant, Shieh's d_s will vary as a function of the sample sizes allocation ratio (unlike Cohen's d_s^* that we will define below). At the population level, δ_{Shieh} also depends on the sample sizes allocation ratio, as illustrated in the following shiny application: https://effectsize.shinyapps.io/ShiehvsCohen/.

Cohen's $\mathbf{d_s^*}$ An effect size estimator based on the sample mean difference divided by the square root of the non pooled average of both variance estimates was suggested by Welch (1938). Here, we indicate the difference between Cohen's d_s (based on the pooled standard deviations) and Cohen's d_s^* with an asterisk. This yields:

$$Cohen's\ d_s^* = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\left(S_1^2 + S_2^2\right)}{2}}}$$

where $\bar{X}_j$ is the mean and S_j is the standard deviation of the j^{th} sample (j = 1,2). We know the distribution of Cohen's d_s^* (Huynh, 1989):

$$Cohen's\ d_s^* \sim \sqrt{\frac{2(n_2 \times \sigma_1^2 + n_1 \times \sigma_2^2)}{n_1 n_2 (\sigma_1^2 + \sigma_2^2)}} \times t_{df^*, ncp^*} \tag{15}$$

Where df^* and ncp^* are defined as follows:

$$df^* = \frac{(n_1 - 1)(n_2 - 1)(\sigma_1^2 + \sigma_2^2)^2}{(n_2 - 1)\sigma_1^4 + (n_1 - 1)\sigma_2^4} \tag{16}$$

$$ncp^* = \delta^*_{Cohen} \times \sqrt{\frac{n_1 n_2(\sigma_1^2 + \sigma_2^2)}{2(n_2\sigma_1^2 + n_1\sigma_2^2)}} = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

where $\delta^*_{Cohen} = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2 + \sigma_2^2}{2}}}$. Using equation 15 we can compute its theoretical expectation and variance when the assumption of normality is met (see Table 2), and therefore determine which factors influence bias and variance, and how they do so (see Supplemental Material 1). This estimator has been widely criticized, because it results in a variance term of an artificial population (i.e. since the variance term does not estimate the variance of one or the other group, the composite variance is an estimation of the variance of an artificial population) and is therefore very difficult to interpret (Grissom & Kim, 2001), and unless both sample sizes are equal, the variance term does not correspond to the variance of the mean difference (Shieh, 2013).

However, we will show throughout the simulation section that this estimator exhibits very good inferential properties. Moreover, it has a constant value across sample sizes ratios, as shown in the Shiny App at https://effectsize.shinyapps.io/ShiehvsCohen/.

Glass's g_s, Shieh's g_s and Hedges' g^*_s As for Cohen's d_s, a Hedges' correction can be applied in order to compensate for the bias of Glass's d_s, Shieh's d_s and Cohen's d^*_s with small sample sizes (see Table 2). This correction has the following general form:

$$g_s = d_s \times \frac{\Gamma(\frac{\nu}{2})}{\sqrt{\frac{\nu}{2}} \times \Gamma(\frac{\nu-1}{2})}$$

where the distinct values of ν are provided in equation 9 for Glass's g_s, in equation 16 for Hedges' g^*_s and in equation 11 for Shieh's g_s. The three corrected estimators are theoretically unbiased when the assumption of normality is met. Their variance is a function of the same parameters as their biased equivalent. However, due to the correction they have a smaller variance, especially with small sample sizes, as shown in Table 3. In summary:

— The variances of Hedges' g^*_s and Shieh's g_s depend on the total sample size (N), their respective population effect size (δ), and the interaction between the sample sizes ratio and the SD-ratio $\left(\frac{n_1}{n_2} \times \frac{\sigma_1}{\sigma_2}\right)$.

— The variance of Glass's g_s also depends on N, δ and $\frac{n_c}{n_e} \times \frac{\sigma_c}{\sigma_e}$. In addition, there is also a main effect of the SD-ratio $\left(\frac{\sigma_c}{\sigma_e}\right)$ on its variance.

How these parameters influence the variance of the estimators will be summarized and illustrated in Monte Carlo simulations below.

TABLE 3: Expectation, bias and variance of Glass's d_s and Cohen's d_s^* and Shieh's d_s under the assumption that independent residuals are normally distributed.

	df	Expectation	Variance
Glass's g_s	$n_c - 1$	δ_{glass}	$Var(Glass's\ d_s) \times \left(\frac{\Gamma\left(\frac{df}{2}\right)}{\sqrt{\frac{df}{2}} \times \Gamma\left(\frac{df-1}{2}\right)} \right)^2$
Cohen's g_s^*	$\frac{(n_1-1)(n_2-1)(\sigma_1^2+\sigma_2^2)^2}{(n_2-1)\sigma_1^4+(n_1-1)\sigma_2^4}$	δ_{Cohen}^*	$Var(Cohen's\ d_s^*) \times \left(\frac{\Gamma\left(\frac{df}{2}\right)}{\sqrt{\frac{df}{2}} \times \Gamma\left(\frac{df-1}{2}\right)} \right)^2$
Shieh's g_s	$\approx \frac{\left(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)^2}{\frac{(\sigma_1^2/n_1)^2}{n_1-1}+\frac{(\sigma_2^2/n_2)^2}{n_2-1}}$	δ_{Shieh}	$Var(Shieh's\ d_s) \times \left(\frac{\Gamma\left(\frac{df}{2}\right)}{\sqrt{\frac{df}{2}} \times \Gamma\left(\frac{df-1}{2}\right)} \right)^2$

Note. $c_f = \frac{\sqrt{\frac{df}{2}} \times \Gamma\left(\frac{df-1}{2}\right)}{\Gamma\left(\frac{df}{2}\right)}$; $\delta_{Glass} = \frac{\mu_c - \mu_e}{\sigma_c}$, $\delta_{Shieh} = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n_1/N} + \frac{\sigma_2^2}{n_2/N}}}$ and $\delta_{Cohen}^* = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2+\sigma_2^2}{2}}}$; all estimators are unbiased estimators, because their expectations equal the population effect size δ; equations require $df \geq 3$ and at least 2 subjects per group.

Monte Carlo Simulations

Assessing the bias, efficiency and consistency of 5 estimators

Method We performed Monte Carlo simulations using R (version 3.5.0) to assess the bias, efficiency and consistency of Cohen's g_s, Glass's g_s (using respectively the sample SD of the first or second group as a standardizer), Hedges' g_s^* and Shieh's g_s.

A set of 100,000 datasets was generated for 1,008 scenarios as a function of different criteria. In 252 scenarios, samples were extracted from a normally distributed population (in order to ensure the reliability of our calculation method) and in 756 scenarios, samples were extracted from non normal population distributions. In order to assess the quality of estimators under realistic deviations from the normality assumption, we referred to the review of Cain, Zhang, & Yuan (2017). Cain, Zhang, & Yuan (2017) investigated 1,567 univariate distributions from 194 studies published by authors in Psychological Science (from January 2013 to June 2014) and the American Education Research Journal (from January 2010 to June 2014). For each distribution, they computed Fisher's skewness

$$G_1 = \frac{\sqrt{n(n-1)}}{n-2} \frac{m_3}{\sqrt{(m_2)^3}}$$

and kurtosis

$$G_2 = \frac{n-1}{(n-2)(n-3)} \times \left[(n+1) \left(\frac{m_4}{(m_2)^2} - 3 \right) + 6 \right]$$

where n is the sample size and m_2, m_3 and m_4 are respectively the second, third and fourth centered moments. They found values of kurtosis from G2 = -2.20 to 1,093.48. According to their suggestions, throughout our simulations, we kept constant the population kurtosis value at the 99th percentile of their distribution of kurtosis, i.e. G2=95.75. Regarding skewness, we simulated population parameter values which correspond to the 1st and 99th percentile of their distribution of skewness, i.e. respectively G1 = -2.08 and G1 = 6.32. We also simulated samples extracted from population where G1 = 0, in order to assess the main effect of high kurtosis on the quality of estimators. All possible combinations of population skewness and kurtosis and the number of scenarios for each combination are summarized in Table 4.

TABLE 4: Number of combinations of skewness and kurtosis in our simulations.

		Kurtosis		
		0	95.75	**TOTAL**
	0	252	252	**504**
Skewness	-2.08	/	252	**252**
	6.32	/	252	**252**
	TOTAL	**252**	**756**	**1008**

Note. Fisher's skewness (G1) and kurtosis (G2) are presented in Table 4. The 252 combinations where both G1 and G2 equal 0 correspond to the normal case.

For the 4 resulting combinations of skewness and kurtosis (see Table 4), all other parameter values were chosen in order to illustrate the consequences of factors identified as playing a key role on the variance of unbiased estimators. We manipulated the population mean difference ($\mu_1 - \mu_2$), the sample sizes (n), the sample size ratio (n-ratio = $\frac{n_1}{n_2}$), the population *SD*-ratio (i.e. $\frac{\sigma_1}{\sigma_2}$), and the sample size and population variance pairing $\left(\frac{n_1}{n_2} \times \frac{\sigma_1}{\sigma_2}\right)$. In our scenarios, μ_2 was always 0 and μ_1 varied from 1 to 4, in steps of 1 (so does $\mu_1 - \mu_2$)[11]. Moreover, σ_1 always equals 1, and σ_2 equals .1, .25, .5, 1, 2, 4 or 10, and therefore, so were the ratios of $\frac{\sigma_1}{\sigma_2}$. The simulations for which both σ_1 and σ_2 equal 1 are the particular case of homoscedasticity, or equal population variances across groups. The sample sizes of both groups (n_1 and n_2) were 20, 50 or 100. When sample sizes of both groups are equal, the n-ratio equals 1 (this is known as a balanced design). All possible combinations of n-ratio and population *SD*-ratio were simulated in order to distinguish scenarios where both sample sizes and population variances are unequal across groups (with positive pairing when the group with the largest sample size is extracted from the population with the largest *SD*, and negative pairing when the group with the smallest sample size is extracted from the population with the smallest *SD*) and scenarios with no pairing

between sample sizes and variances (sample sizes and/or population *SD* are equal across all groups). In sum, the simulations grouped over different sample sizes yield 4 conditions (a, b, c and d) based on the n-ratio, population *SD*-ratio, and sample size and population variance pairing, as summarized in Table 5. We chose to divide scenarios into these 4 conditions because analyses in Supplemental Material 1 revealed main and interaction effects of sample sizes ratio and *SD*-ratio on the bias and variance of some estimators.

TABLE 5: 4 conditions based on the n-ratio and the SD-ratio.

		***n*-ratio**		
		1	**>1**	**<1**
	1	a	b	b
***SD*-ratio**	**>1**	c	d	d
	<1	c	d	d

Results Before presenting the comparison of the estimators for each condition, it is useful to make some general comments.

1) We previously discussed the fact that raw bias and variances are sometimes misleading. They can give the illusion of huge differences between two estimators, even if these differences only reflect a change of unit (i.e. different population effect sizes). To better understand this, imagine a sample of 15 people for whom we know the height (in meters) and we compute a sample variance of 0.06838. If we convert sizes to centimeters and compute the sample variance again, we find a measure of 683.8 (i.e. 10^4 larger). Both measures represent the same amount of variability, but they are expressed in different units. The same issue due to a change in scales occurs when comparing the estimates of different population measures. To avoid this possible confusion, we will only present the relative bias and relative variance in all Figures (and anytime we will mention the biases

and variances in the results section, we will be referring to relative bias and variance). For interested readers, illustrations of the raw bias and variance are available on Github: https://anonymous.4open.science/r/deffectsize-B1D2/.

2) For the sake of readability, the vertical axis differs across plots.

3) Throughout this section, we will *compare* the relative bias and variance of different estimators, but we do not present bias and variance in absolute terms. We chose very extreme (although realistic) conditions, and we know that none of the parametric measures of effect size will be robust against such extreme conditions. Our goal is therefore to study the robustness of the estimators against normality violations only in comparison with the robustness of other indicators, but not in absolute terms.

After these general remarks, we will analyze each condition separately. In all Figures presented below, for different sub-conditions, the averaged relative bias and relative variance of five estimators are presented. When describing the Glass's g_s estimators, we will systematically refer to the "control group" as the condition the standardizer is based on (i.e. the first group when using S_1 as standardizer, the second group when using S_2 as standardizer). The other condition will be referred to as the "experimental group".

When variances are equal across groups

Figures 2 and 3 represent configurations where the equality of variances assumption is met. According to our expectations, one observes that the bias of all estimators is approximately zero as long as the normality assumption is met (first column in both Figures)[12]. However, the more the data generation process deviations from the normality assumption (i.e. when moving from left to right in the Figures), the larger the bias in the estimators.

We will observe that Glass's g_s should always be avoided when the equality of variance assumption is met. Hedges' g_s, Hedges' g_s^* and Shieh's g_s perform equally well as long as the sample size ratio is close to 1 (condition a; see Figure 2). However, when designs are highly unbalanced (condition b; see Figure 3), Shieh's g_s is not consistent anymore, while Hedges' g_s^* remain consistent, Hedges's g_s is a better estimator. For interested readers, these findings are detailed in the three paragraphs below.

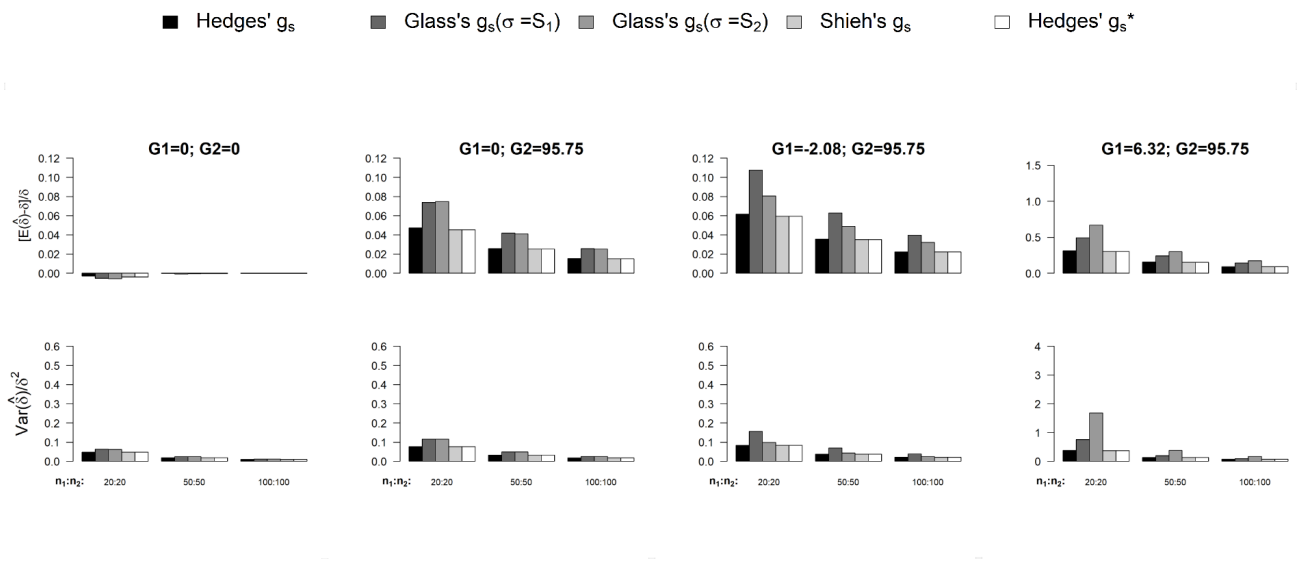


FIGURE 2 – Bias and efficiency of estimators of standardized mean difference, when variances and sample sizes are equal across groups (condition a)

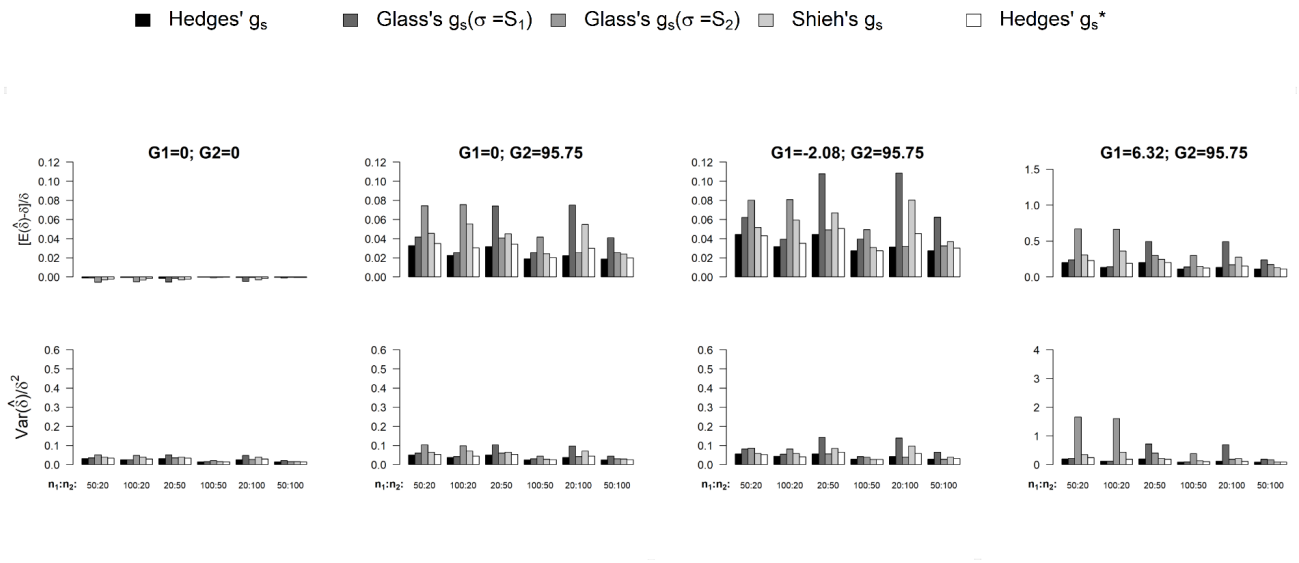


FIGURE 3 – Bias and efficiency of estimators of standardized mean difference, when variances are equal across groups and sample sizes are unequal (condition b)

Figure 2 illustrates scenarios where both population variances and sample sizes are equal across groups (condition a). One can first notice that all estimators are consistent, as their bias and variance decrease when the total sample size increases. For any departure from the normality assumption, both bias and variance of Hedges' g_s, Shieh's g_s and Hedges' g_s^* are similar[13] and smaller than the bias and variance of Glass's g_s estimates using either S_1 or S_2 as a standardizer. Moreover, when samples are extracted from skewed distributions, Glass's g_s will show different bias and variance as a function of the chosen standardizer (S_1 or S_2), even if both S_1 and S_2 are estimates of the same population variance, based on the same sample size. This is due to non-null correlations of opposite sign between the mean difference ($\bar{X}_1 - \bar{X}_2$) and respectively S_1 and S_2. In Supplemental Material 3, we detailed in which situation a non-null correlation occurs between the sample mean difference ($\bar{X}_1 - \bar{X}_2$) and the standardizer of compared estimators as well as the way this correlation impacts the bias and variance of estimators.

Figure 3 illustrates scenarios where population variances are equal across groups, but sample sizes are unequal (condition b). For any departures from the normality assumptions, Hedges' g_s shows the smallest bias and variance. Hedges' g_s and Hedges' g_s^* are consistent estimators (i.e. the larger the sample sizes, the lower the bias and the variance), unlike Shieh's g_s and Glass's g_s. The bias of Glass's g_s does not depend either on the size of the experimental group or on the total sample size. The only way to decrease the bias of Glass's g_s is therefore to add subjects in the control group. On the other hand, the variance of Glass's g_s depends on both sample sizes, but not in an equivalent way: in order to reduce the variance, it is much more efficient to add subjects in the control group and when the size of the experimental group decreases so does the variance, even when the total sample size is increased. Regarding Shieh's g_s, for a given sample size ratio, the bias and variance will decrease when sample sizes increase. However, there is a large effect of the sample sizes ratio such that when the sample sizes ratio moves away from 1 by adding subjects, bias and variance might increase.[14] On the other hand, when the sample sizes ratio moves closer to 1 by adding subjects, the bias will decrease.

When samples are extracted from skewed distributions and have unequal sizes (the two last columns in Figure 3), for a constant total sample size, Glass's g_s, Shieh's g_s and Hedges' g_s^* will show different bias and variance depending on which group is the largest one (e.g. when distributions are right-skewed, the bias and variance of all these estimators when n_1 and n_2 are respectively 50 and 20 are not the same as their bias and variance when n_1 and n_2 are

respectively 20 and 50). This is due to a non-null correlation of opposite sign between the mean difference ($\bar{X}_1 - \bar{X}_2$) and their respective standardizers depending on which group is the largest one, as detailed in Supplemental Material 3. One observes that under these configurations, the bias and variance of Glass's g_s are sometimes a bit smaller and sometimes much larger than the bias and variance of Shieh's g_s and Cohen's d_s^*. [15]

When variances are unequal across groups

Figures 4 to 9 represent configurations where the equality of variances assumption is not met. According to our expectations, one observes that the bias of all estimators is approximately zero as long as the normality assumption is met (first column in all Figures), and the further from the normality assumption (i.e. when moving from left to right in Figures), the larger the bias[16]. It might be considered surprising that the bias of Hedges' g_s remains very small throughout these conditions. As discussed in the section "Different measures of effect size", Hedges' g_s should be avoided when population variances and sample sizes are unequal across groups, because of the pooled error term. When pooling the estimates of two unequal population variances, the resulting estimator will be smaller (in case of positive pairing) or larger (in case of negative pairing) than it should be. At the same time, when pooling two unequal population variances, the population effect size will also be smaller (in case of positive pairing) or larger (in case of negative pairing) as it should be. As a consequence, the distortion cannot be seen through the difference between the expected estimator and the population effect size measure. For this reason, the bias and variance of Hedges' g_s will not be taken into account in the following comparisons.

We will observe that when variances are unequal across populations, Glass's g_s sometimes performs better, but also sometimes performs much worst than Shieh's g_s and Hedges' g_s^*, both in terms of bias and variance. The performance of Glass's g_s highly depends on parameters that we cannot control (i.e. a triple interaction between the n-ratio, the SD-ratio and the correlation between the standardizer and the mean difference) and for this reason, we do not recommend using it. When the sample sizes ratio is close to 1, Shieh's g_s and Hedges' g_s^* are both appropriate but the further the sample sizes ratio is from 1, the larger the bias of Shieh's g_s in order that in the end, the measure that we believe performs best across scenarios is Hedges' g_s^*.

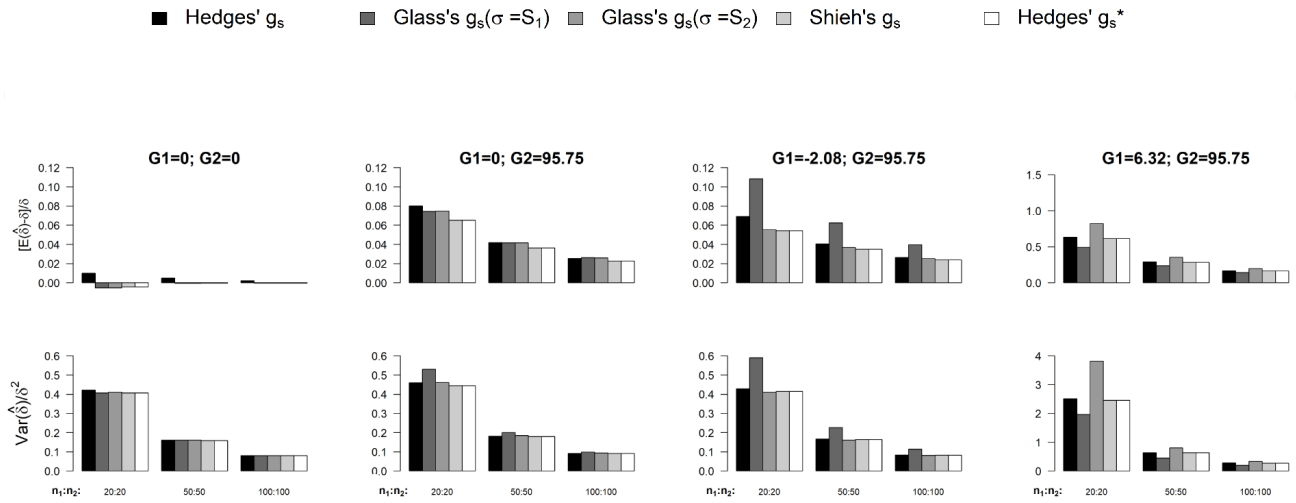


FIGURE 4 – Bias and efficiency of estimators of standardized mean difference, when variances are unequal across groups and sample sizes are equal (condition c), as a function of n-ratio

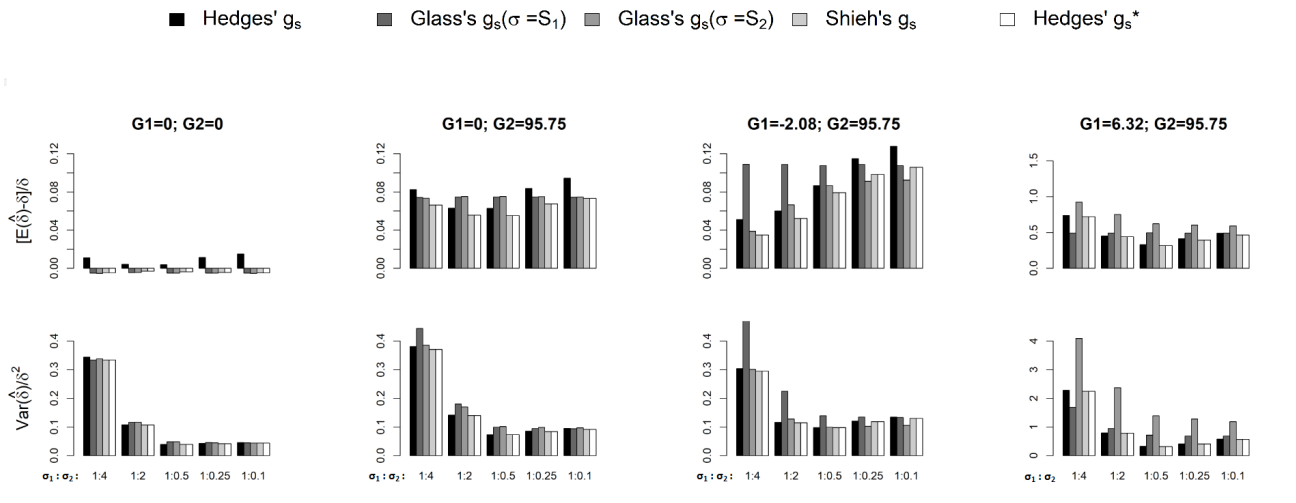


FIGURE 5 – Bias and efficiency of estimators of standardized mean difference, when variances are unequal across groups and sample sizes are equal (condition c) as a function of the SD-ratio (when $n_1 = n_2 = 20$)

Figures 4 and 5 are dedicated to scenarios where population variances are unequal between groups and sample sizes are equal (condition c). In Figure 4, scenarios are subdivided as a function of the sample sizes and one can notice that all estimators are consistent, as their bias and variance decrease when the total sample size increases. In Figure 5, scenarios are subdivided as a function of the SD-ratio. Because the comparison pattern remains very similar for all sample sizes, we present only scenarios when sample sizes equal 20. One should first notice that for all estimators in Figure 5, the relative variance seems to be much larger when $S_2 > S_1$. [17] This information should not be taken into account because it is only an artefact of our simulation conditions combined with the way we computed the relative variance. [18]

When samples are extracted from skewed distributions, the bias and variance of Glass's g_s are sometimes smaller and sometimes larger than the bias of Shieh's g_s and Hedges' g_s. This is mainly due to the fact that when two samples of same sizes are extracted from two skewed distributions with unequal variances (the two last columns in Figure 5), there will be non-null correlations of opposite sign between the mean difference $(\bar{X}_1 - \bar{X}_2)$ and the standardizer of *all* estimators, depending on which population variance is larger [19].

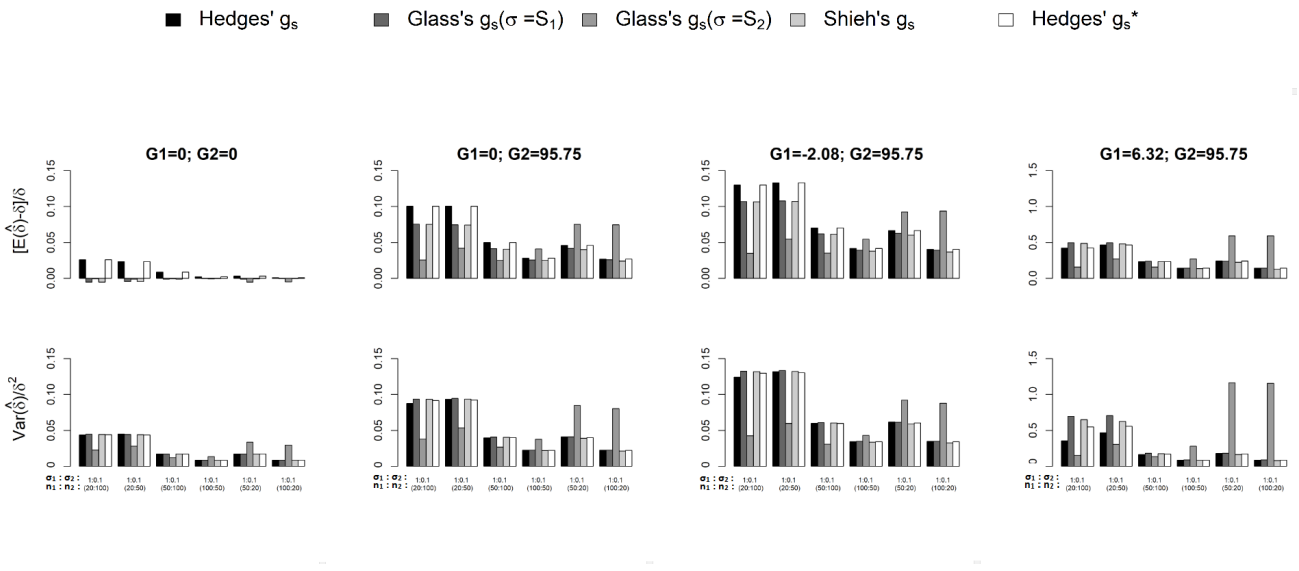


FIGURE 6 – Bias and efficiency of estimators of standardized mean difference, when variances and sample sizes are unequal across groups (condition d), $\frac{\sigma_1}{\sigma_2} = 10$ and $\sigma_1 > \sigma_2$

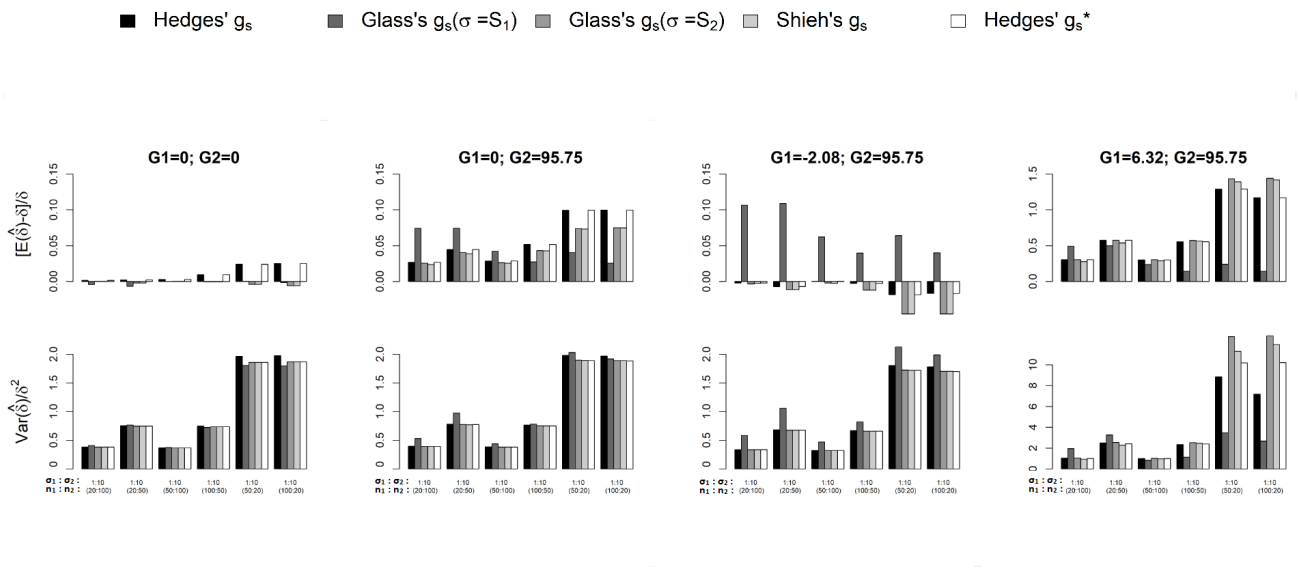


FIGURE 7 – Bias and efficiency of estimators of standardized mean difference, when variances and sample sizes are unequal across groups (condition d), $\frac{\sigma_1}{\sigma_2} = 10$ and $\sigma_1 < \sigma_2$

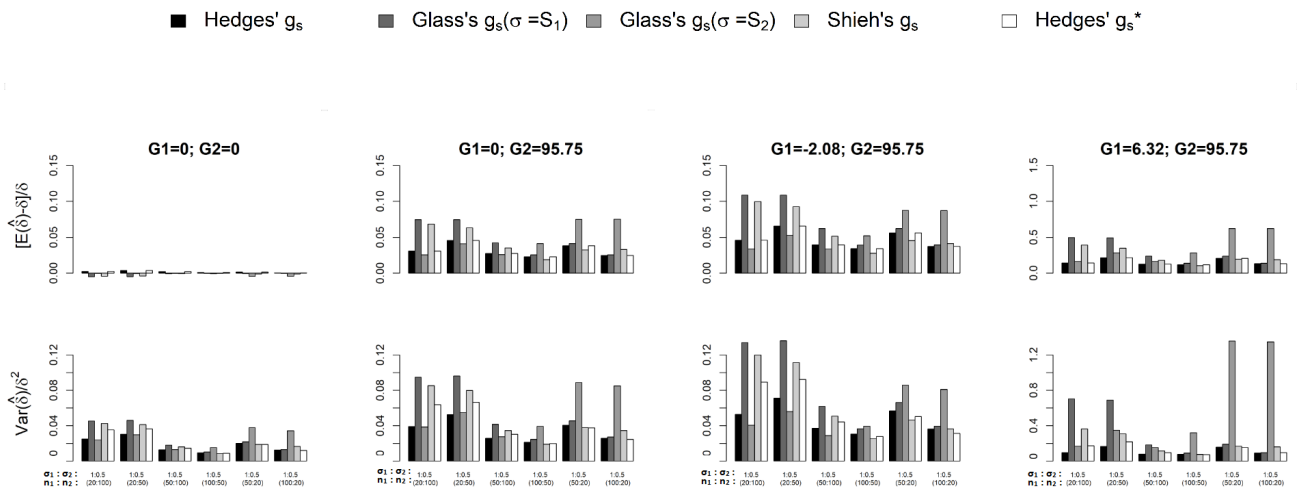


FIGURE 8 – Bias and efficiency of estimators of standardized mean difference, when variances and sample sizes are unequal across groups (condition d), $\frac{\sigma_1}{\sigma_2} = 2$ and $\sigma_1 > \sigma_2$

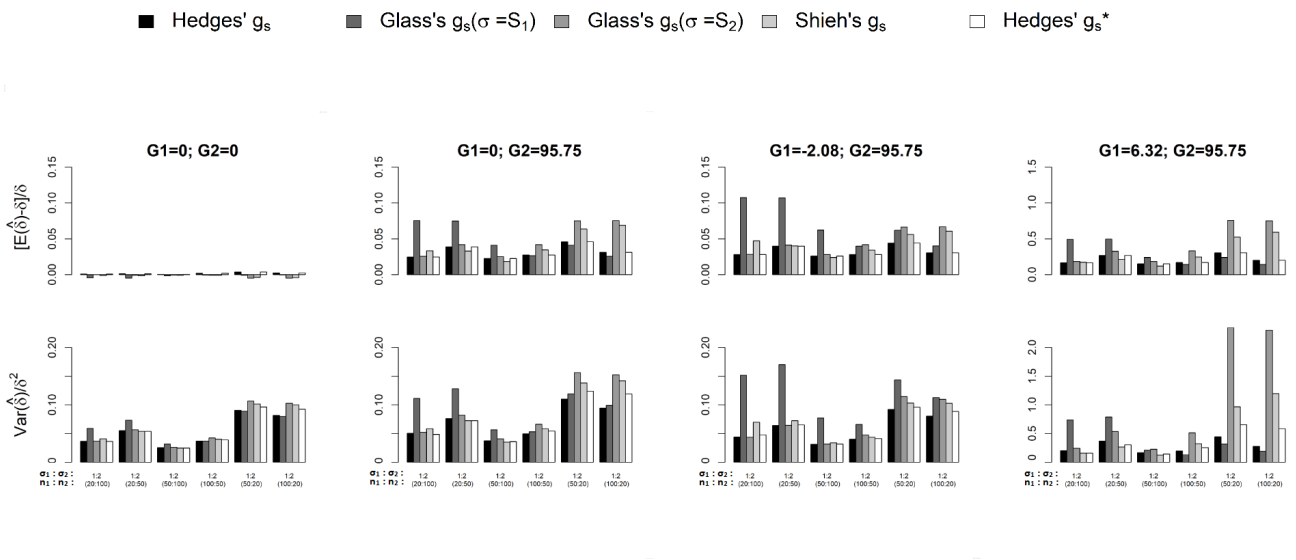


FIGURE 9 – Bias and efficiency of estimators of standardized mean difference, when variances and sample sizes are unequal across groups (condition d), $\frac{\sigma_1}{\sigma_2} = 2$ and $\sigma_1 < \sigma_2$

Figures 6 to 9 are dedicated to scenarios where both sample sizes and population variances differ across groups. Due to a high number of combinations between the sample sizes-ratio and the SD-ratio in our simulations, we decided to present only some conditions. Because equations in Table 3 revealed an interaction effect between the sample sizes ratio and the SD-ratio on the bias and variance of Hedges' g_s^* and Shieh's g_s (see Supplemental Material 1), we chose to present all configurations where the larger SD is 10 times larger than the smaller SD (Figures 6 and 7), and configurations where the larger SD is twice larger than the smaller SD (Figures 8 and 9), in order to compare the effect of the sample sizes ratio on the bias and variance of all estimators when the SD-ratio is large ($\frac{\sigma_1}{\sigma_2} = 10$ *or* .1) or medium ($\frac{\sigma_1}{\sigma_2} = 2$ *or* .5).

When distributions are symmetric, the bias of Glass's g_s only depends on the size of the control group and is therefore not impacted by either the sample sizes ratio or the total sample size. When comparing Figures 6 to 9, one can also notice that the bias of Glass's g_s does not depend on the SD-ratio either. Unlike the bias of Glass's g_s, its variance depends on both sample sizes, but not in an equivalent way. In most scenarios it is more efficient, in order to reduce the variance of Glass's g_s, to add subjects in the control group. Regarding Hedges' g_s^* and Shieh's g_s, their respective biases and variances depend on an interaction effect between the sample sizes ratio and the SD-ratio $\left(\frac{n_1}{n_2} \times \frac{\sigma_1}{\sigma_2}\right)$: the sample sizes ratio associated with the smallest bias and variance is not the same when the more variable group is 10 times more variable than the other

group (Figures 6 and 7) than when it is only twice more variable (Figures 8 and 9). However, the respective biases and variances of Hedges' g_s^* and Shieh's g_s are always smaller when there is a positive pairing between sample sizes and variances. When samples are extracted from skewed distributions, the bias and variance of Glass's g_s are sometimes smaller and sometimes larger than the bias of Shieh's g_s and Hedges' g_s^*, due to a combination of three factors: (1) which group is larger, (2) which group has the smallest standard deviation and (3) what is the correlation between the standardizer and the mean difference.

Recommendations

We recommend using Hedges' g_s^* in order to assess the magnitude of the effect when comparing two independent means, because a) it does not rely on the equality of population variances assumption (unlike Hedges' g_s), b) it is always consistent (unlike Shieh's g_s), c) it is easy to interpret (Hedges' g_s^* can be interpreted in the same way as Hedges' g_s) and d) it remains constant for any sample sizes ratio, even when population variances are unequal across groups, as shown in the Shiny App at https://effectsize.shinyapps.io/ShiehvsCohen/.

Effect sizes estimates such as Hedges' g_s^* should always be reported with a confidence interval. To help researchers compute Hedges' g_s^* and it's confidence interval we created the R package *deffectsize* (see https://anonymous.4open.science/r/deffectsize-B1D2/). The *datacohen_CI* function was built in order to compute point estimators and confidence intervals based on raw data and the *cohen_CI* function was built in order to compute point estimators and confidence intervals based on descriptive statistics (sample means, sample variances and sample sizes). By default, unbiased Cohen's g_s^* is computed but it is also possible to compute biased estimators (e.g. Cohen's d_s^*) and/or to use a pooled error term as standardizer by assuming that the equality of population variances is met (e.g. Hedges' g_s or Cohen's d_s, depending on whether we choose to compute unbiased or biased estimator). Other functions (*datashieh_CI*, *shieh_CI*, *dataglass_CI* and *glass_CI*) are available in order to compute Shieh's g_s (or Shieh's d_s) and Glass's g_s (or Glass's d_s) as well as their respective confidence intervals, even though we don't recommend to use these effect sizes by default. Researchers who do not use R can use a Shiny app to compute point estimators and confidence intervals based on descriptive statistics: https://effectsize.shinyapps.io/deffsize/.

[8]The Cramér-Rao inequality provides a theoretical lower bound for the variance of unbiased estimators. An estimator reaching this bound is therefore optimally efficient.

[9]Under the null hypothesis of no differences between sample means, Student's t-statistic will follow a central t-distribution with $n_1 + n_2 - 2$ degrees of freedom. However, when the null hypothesis is false, the distribution of this quantity will not be centered, and a noncentral t-distribution will arise.

[10]In Table 1, one can see that the variance of Hedges' g_s equals the variance of Cohen's d_s, multiplied by $\left[\frac{\Gamma(\frac{df}{2})}{\sqrt{\frac{df}{2}}\times\Gamma(\frac{df-1}{2})}\right]^2$. This term is always less than 1 and tends to 1 when the sample sizes tends to infinity ($52 \leq \left[\frac{\Gamma(\frac{df}{2})}{\sqrt{\frac{df}{2}}\times\Gamma(\frac{df-1}{2})}\right]^2 < 1$ for $3 \leq df < \infty$). As a consequence, the larger the total sample size, the smaller the difference between the variance of Cohen's d_s and Hedges' g_s.

[11]In the original plan, we had added 252 simulations in which μ_1 and μ_2 were both null. We decided not to present the results of these simulations in the main article, because the relative bias and the relative variance appeared to us to be very useful to fully understand the comparison of the estimators, and computing them is impossible when the real mean difference is zero. Indeed, for these specific configurations, both relative bias and relative variance would have infinite values due to the presence of the population effect size term in their denominator. However, these extra simulations were included in the simulation checks, in Supplemental Material 2.

[12]When looking at relative bias for all estimators, the maximum departure from zero is 0.0064 when sample sizes are equal across groups, and 0.0065 with unequal sample sizes.

[13]While the bias and variance of Cohen's d_s, Cohen's d_s^* and Shieh's d_s are identical, the bias and variance of Hedges' g_s are marginally different from the bias and variance of Hedges' g_s^* and Shieh's g_s (these last two having identical bias and variance). Indeed, because of the sampling error, differences remain between sample variances, even when population variances are equal between groups. Since the Hedges' correction applied to Cohen's d_s does not contain the sample variances (unlike the correction applied on both other estimators), the bias and variance of Hedges' g_s are slighly different from the bias and variance of Hedges' g_s^* and Shieh's g_s.

[14]Regarding variance, in Supplemental Material 1, we mentioned that when the population effect size is zero, the larger the total sample size, the lower the variance, whether the sample sizes ratio is constant or not. We also mentioned that this is no longer true when the population effect size is not zero. In our simulations the effect size is never zero. The effect size effect is partially visible in Figure 3 because we do not entirely remove the effect size effect when we divide the variance by δ^2. This is due to the fact that one term, in the equation of the variance computation, does not depend on the effect size.

[15]Supplemental Material 3 shows that when the $\mu_1 - \mu_2 > 0$ (like in our simulations), all other parameters being equal, an estimator is always less biased and variable when choosing a standardizer that is positively correlated with $\bar{X}_1 - \bar{X}_2$. Supplemental Material 3 also shows that the smaller n_c, the larger the magnitude of

correlation between S_c and $\bar{X}_1 - \bar{X}_2$. When $cor(S_c, \bar{X}_1 - \bar{X}_2)$ is positive, the positive effect of increasing the magnitude of the correlation is counterbalanced by the negative effect of reducing n_c. On the other hand, when $cor(S_c, \bar{X}_1 - \bar{X}_2)$ is negative, the negative effect of increasing the magnitude of the correlation is amplified by the negative effect of decreasing n_c. This explains why the difference between Glass's g_s and other estimators is larger when Glass's g_s is the least efficient estimator.

[16]When looking at the relative bias for all estimators, the maximum departure from zero is 0.0173 when sample sizes are equal across groups, and 0.0274 when both sample sizes and variances differ across groups.

[17]The difference between the variance of estimators when the second group is 10 times larger than the first group was so large that we decided to not present it, for the sake of readability of the Figures.

[18]We previously mentioned that when dividing the variance by δ^2, we do not entirely remove the effect size effect. Actually, we introduce δ^2 in the denominator of the first term, in the equation of the variance computation. Because we performed our simulations in order that σ_1 always equals 1, the smaller S_2, the larger the population effect size and therefore, the smaller the relative variance.

[19]When population variances are unequal, a non-null correlation occurs between standardizer estimates and $\bar{X}_1 - \bar{X}_2$. For standardizers computed based on both S_1 and S_2, the sign of the correlation between the standardizer and the mean difference will be the same as the sign of the correlation between the mean difference and the estimate of the larger population variance. For interested readers, this is detailed in Supplemental Material 3.

Chapitre 5: les tests d'équivalence

Lorsqu'on applique un test d'hypothèse, l'hypothèse nulle la plus couramment définie est celle d'absence d'effet ou de différence entre les groupes (Nickerson, 2000). Il arrive également parfois que les chercheurs définissent un intervalle de valeur comme hypothèse nulle, mais le plus souvent, cet intervalle est borné par la valeur 0 (Nickerson, 2000), on parle alors d'hypothèse unilatérale. Avec cette stratégie, le rejet de l'hypothèse nulle constitue un soutien en faveur de la présence d'un effet non nul, par contre, le non rejet de l'hypothèse nulle ne peut être interprété comme un soutien en faveur de l'absence d'effet. Pourtant, il arrive souvent que des chercheurs l'interprètent de la sorte (Anderson & Maxwell, 2016). Finch, Cumming, & Thomason (2001), par exemple, ont reporté que parmis 150 articles publiés entre 1940 et 1999 dans le *JAP* (*Journal of Applied Psychology*), 38% interprétaient un résultat non significatif comme une acceptation de l'hypothès nulle. Plus récemment, Lakens (2017) a noté que l'expression "pas d'effet" a été utilisée dans 108 articles publié dans *Social Psychological and Personality Science* avant août 2016 et que dans presque tous les cas, c'était sur base du non rejet de l'hypothèse nulle que cette conclusion était tirée. Cette erreur d'interprétation est également fréquemment commise dans le cadre des études de réplication. Anderson & Maxwell (2016), par exemple, ont analysé 50 réplications d'études publiées en 2013 dans PsycINFO. Ils ont noté que 14 études affirmaient avoir obtenu des effets "nuls" (interprété comme un échec à la réplication), et tous l'ont fait sur base de l'acceptation d'une hypothèse nulle d'absence d'effet. C'est par exemple de cette manière qu'on été réalisées la plupart des tentatives de réplications de la célèbre étude de Bem (Ritchie, Wiseman & French, 2012, cités par Anderson & Maxwell, 2016).

A travers ce chapitre, notre premier objectif sera d'expliquer pourquoi interpréter le non rejet de l'hypothèse d'absence d'effet comme un soutien en faveur d'une absence d'effet n'est pas une bonne stratégie. Nous introduirons ensuite les tests d'équivalence qui permettent d'obtenir un soutien en faveur d'un effet jugé non pertinent, et plus particulièrement le TOST (Two One-sided test). Nous verrons que l'aspect le plus compliqué de la réalisation du TOST est la définition des bornes d'équivalence. Pour cette raison, notre troisième objectif sera de fournir quelques pistes en vue de définir ces bornes. Pour finir, nous présenterons un article dans lequel nous comparons le TOST à la SGPV (Second Generation P-Value), une stratégie récemment développée par Blume, D'Agostino McGowan, Dupont, & Greevy Jr (2018).

Limites de l'approche traditionnelle

Lorsqu'on teste une hypothèse nulle, il y a deux conclusions possibles: soit ont la rejette, soit on ne la rejette pas. Si rejeter l'hypothèse nulle amène à conclure en faveur de l'hypothèse alternative, ne pas la rejeter ne permet pas de conclure en faveur de l'hypothèse nulle. Au mieux, cela nous montre que les données ne sont pas incompatibles avec l'hypothèse nulle, mais cela ne veut en aucun cas dire qu'elles ne sont compatibles avec aucune autre hypothèse. Afin de l'illustrer, la Table 1 résume les résultats de simulations Monte Carlo pour un ensemble de 42 scénarios qui varient en fonction de la taille des échantillons (n_j) et de la différence entre les moyennes des deux populations dont sont extraits les échantillons ($\mu_1 - \mu_2$). Pour chaque scénario, à 100,000 reprises, nous avons généré aléatoirement une paire d'échantillons indépendants, réalisé un test t de Student pour échantillons indépendants et extrait la p-valeur du test. Ensuite, nous avons calculé la proportion d'itérations associées à une p-valeurs supérieures à .05, nous amenant à ne pas rejeter l'hypothèse nulle lorsqu'on travaille avec un risque alpha de 5% (ce risque alpha étant communément accepté par la majorité des chercheurs, Meyners, 2012). Lorsque l'hypothèse nulle est fausse (toutes les colonnes de la Table 1, à l'exception de la première), cette proportion correspond au taux d'erreur de type II (communément appelé β).

Table 1.

Proportion de p-valeurs supérieures à .05 en fonction de la taille des échantillons (n_j) et de la différence entre les moyennes de chaque population ($\mu_1 - \mu_2$).

	Différence de moyennes dans la population ($\mu_1 - \mu_2$)					
n_j	0	.1	.2	.3	.4	.5
100	0.949	0.891	0.709	0.441	0.196	0.060
200	0.949	0.832	0.483	0.151	0.021	0.001
300	0.950	0.768	0.312	0.043	0.001	< .001
400	0.949	0.708	0.194	0.011	< .001	< .001
500	0.949	0.646	0.116	0.003	< .001	< .001
600	0.950	0.591	0.066	0.001	< .001	< .001
700	0.948	0.536	0.037	< .001	< .001	< .001

Note. Pour chaque scénario, les deux échantillons sont toujours de même taille ($n_1 = n_2 = n$) et sont extraits de populations se distribuant normalement et ayant la même variance ($\sigma_1 = \sigma_2 = \sigma$). La moyenne de la première population (μ_1) vaut systématiquement 0, et celle de la deuxième population (μ_2) varie de sorte à obtenir la différence de moyenne $\mu_1 - \mu_2$ désirée. Par ailleurs, σ vaut systématiquement 1, si bien que la différence de moyenne brute est égale au δ de Cohen.

Pour les scénarios de la première colonne, l'hypothèse nulle est vraie: il n'y a pas de différence entre les moyennes de population. Puisque les conditions d'application du test t de Student sont toutes rencontrées, on est amené à rejeter l'hypothèse nulle, c'est-à-dire à commettre une erreur de type I dans une proportion d'itération égale à $\alpha = 5\%$. Par conséquent, on est amené à *ne pas* rejeter l'hypothèse nulle 95% du temps (ce qui correspond à $(1-\alpha)\%$)[20]. Pour les scénarios envisagés dans toutes les autres colonnes de la Table 1, une vraie différence entre les moyennes de population existe, si bien que le rejet de l'hypothèse nulle est la bonne décision. Pourtant, pour plusieurs scénarios, le nombre d'itérations amenant à conclure au non rejet de l'hypothèse nulle est bien supérieur au nombre d'itérations amenant à conclure au rejet de l'hypothèse nulle, comme on peut le voir à travers les valeurs β. Par exemple, avec 100 sujets par groupes et considérant $\sigma_1 = \sigma_2 = 1$, on ne détectera pas une différence de moyenne de .1 dans près de 90% des cas. Avec 700 sujets par groupe, cette différence ne sera toujours pas détectée plus d'une fois sur deux ($\approx 54\%$ des itérations). En présence d'un effet non nul, cela se justifie par un manque de puissance des tests réalisés, ce qui démontre bien qu'un non rejet de l'hypothèse nulle peut en fait signifier deux choses: soit qu'il n'y a vraiment pas de différence entre les moyennes des populations (ou autrement dit, que les différences observées sont dûes au hasard), soit que le test n'est pas suffisamment puissant pour détecter la différence. Or, le manque de puissance des tests est récurrent dans la littérature, comme tendent à le montrer diverses méta-analyses (Bakker, Van Dijk, & Wicherts, 2012 ; Button et al., 2013 ; Funder et al., 2014).

Pour éviter d'interpréter un test peu puissant comme un soutien en faveur de l'hypothèse nulle, l'approche de la puissance est devenue l'approche par défaut dans les années 80 pour tester l'équivalence (Meyners, 2012). A travers cette approche qui est restée très populaire (Quertemont, 2011), dans un premier temps, on définit ce qu'on considère comme étant la plus petite valeur d'intérêt (en anglais, le "SESOI" pour "Smaller Effect Size of Interest"), c'est-à-dire la taille d'effet minimale requise pour considérer qu'un effet est pertinent. Ensuite, on estime la puissance de notre test à détecter un effet de cette taille[21], et si cette estimation atteind une valeur jugée satisfaisante (en général, 80%), alors on considère que l'on peut interpréter le non rejet de l'hypothèse nulle d'absence d'effet comme soutien en faveur de l'équivalence (Meyners, 2012 ; Quertemont, 2011 ; Schuirmann, 1987). L'idée sous-jacente est que si l'effet est au moins aussi grand que les bornes de la zone d'équivalence, sur le long terme, on devrait le plus souvent rejeter l'hypothèse nulle. Par conséquent, un non rejet de l'hypothèse nulle devrait généralement

signifier que l'effet n'atteint pas le SESOI et donc, que l'effet observé n'est pas pertinent. Bien que ce raisonnement puisse sembler tentant, de prime abord, il présente d'importantes limites.

Premièrement, le test n'a pas de bonnes propriétés asymptotiques. Ceci est illustré au sein de la Table 2, dans laquelle nous envisageons les mêmes scénarios que dans la Table 1 et ajoutons une contrainte de puissance: nous décidons qu'on ne peut conclure à l'équivalence que si l'on atteind une puissance de 80% pour détecter une différence de moyenne de .3. On constate qu'avec 100 sujets par groupes, aucune itération n'amènera à conclure à l'équivalence, pas même lorsque la différence entre les moyennes de population vaut 0. Cela s'explique par le fait que l'on n'atteind jamais la puissance minimale de 80%.[22] Par contre, une fois les échantillons assez grands pour s'assurer une puissance de 80% pour détecter une différence de moyennes de population de .3, lorsque la différence entre les moyennes de populations est non nulle, la proportion d'itérations qui amènent à conclure à l'équivalence diminue à mesure que la taille des échantillons augmente. Par exemple, lorsque la différence de moyennes vaut .1 au niveau des populations, on conclura à l'équivalence dans 81% des itérations avec 200 sujets par groupe, contre seulement 54% des itérations avec 700 sujets par groupe [23].

Table 2.

Proportion d'itérations qui amènent à conclure à l'équivalence en fonction de la taille des échantillons (n_j) et de la différence entre les moyennes de chaque population ($\mu_1 - \mu_2$), quand on exige une puissance minimale de 80% pour détecter une différence de moyenne de .3.

	Différence de moyennes dans la population ($\mu_1 - \mu_2$)					
n_j	0	.1	.2	.3	.4	.5
100	< .001	< .001	< .001	< .001	< .001	< .001
200	0.923	0.809	0.469	0.146	0.020	0.001
300	0.950	0.768	0.312	0.043	0.001	< .001
400	0.949	0.708	0.194	0.011	< .001	< .001
500	0.949	0.646	0.116	0.003	< .001	< .001
600	0.950	0.591	0.066	0.001	< .001	< .001
700	0.948	0.536	0.037	< .001	< .001	< .001

Note. Pour chaque scénario, les deux échantillons sont toujours de même taille ($n_1 = n_2 = n$) et sont extraits de populations se distribuant normalement et ayant la même variance ($\sigma_1 = \sigma_2 = \sigma$). La moyenne de la première population (μ_1) vaut systématiquement 0, et celle de la deuxième population (μ_2) varie de sorte à obtenir la différence de moyenne $\mu_1 - \mu_2$ désirée. Par ailleurs, σ vaut systématiquement 1, si bien que la différence de moyenne brute est égale au δ de Cohen.

Deuxièmement, pour une taille d'échantillon donnée, plus l'erreur (la variabilité des scores au sein de chaque groupe) sera plus grande (Meyners, 2012 ; Schuirmann, 1987), plus la probabilité

de conclure à l'équivalence augmentera. Ce dernier point est illustré au sein de la Figure 10, dans le contexte de la comparaison de deux moyennes. Sur l'axe des abscisses, on représente différentes estimations de la différence de moyenne ($\bar{X}_1 - \bar{X}_2$) et sur l'axe des ordonnées, la précision des estimations $\bar{X}_1 - \bar{X}_2$ ($S\sqrt{\frac{2}{n}}$ correspond à l'estimation de l'erreur standard de $\bar{X}_1 - \bar{X}_2$, avec S étant l'écart-type poolé et n la taille de chaque échantillon, lorsque les échantillons ont tous les deux la même taille et sont extraits de population ayant la même variance)[24]. Le triangle grisé représente l'ensemble des combinaisons estimation/précision qui vont amener à conclure à l'équivalence, avec l'approche de la puissance, lorsqu'on travaille avec des échantillons de taille 50, en acceptant un risque α de 5% et en exigeant une puissance minimale de 80% pour détecter une différence de 20 unités ($|\theta_j = 20|,\ j = 1, 2$). Dans cet exemple, pour toutes les valeurs de $S\sqrt{\frac{2}{n}}$ supérieures à 7.07 aucune estimation de différence de moyennes ne permettra de conclure à l'équivalence (pas même 0) puisque la puissance du test à détecter une différence de 20 unités est inférieure à 80%. Pour toutes les valeurs de $S\sqrt{\frac{2}{n}}$ inférieures à 7.07, on constate que plus notre estimation de $\bar{X}_1 - \bar{X}_2$ est précise (lorsqu'on se déplace du haut vers le bas, sur l'axe des ordonnées), plus l'estimation doit être proche de 0 pour pouvoir conclure à l'équivalence. Comme on peut le voir à travers le triangle hachuré sur la Figure 1, cette propriété peu désirable n'est pas partagée par le TOST, un test d'équivalence que nous allons décrire ci-dessous (Schuirmann, 1987).

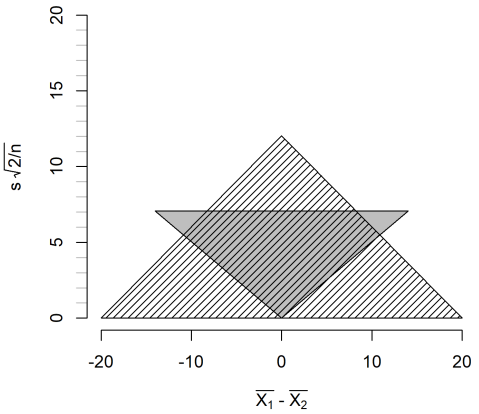


FIGURE 10 – Région d'équivalence pour l'approche de la puissance (zone grisée) et pour le TOST (zone hachurée), pour l'exemple où $|\theta|$=20, n = 50 et $\alpha = .05$

Les tests d'équivalence

Avec les tests d'équivalence, il n'est pas possible de démontrer qu'un effet vaille exactement zéro (Meyners, 2012). Il est par contre possible de montrer que l'effet observé est suffisamment petit pour être jugé non pertinent. Or, cela peut s'avérer précieux dans de nombreuses situations, par exemple pour justifier la décision de regrouper plusieurs groupes de sujets ensemble (Rogers, Howard, & Vessey, 1993), pour contrôler qu'il n'y ait pas de différence trop importante entre les groupes sur base de critères autres que le (ou les) facteur(s) d'intérêts en cas de quasi-expérience (Seaman & Serlin, 1998) ou encore pour falsifier une théorie qui prônerait en faveur d'un effet dépassant une certaine taille (Anderson & Maxwell, 2016 ; Lakens, 2017).

Le point de départ des tests d'équivalence est de définir θ_1 et θ_2, les bornes inférieures et supérieures de la zone d'équivalence, cette dernière contenant l'ensemble des valeurs jugées trop petites pour être susceptibles de nous intéresser. Ces bornes peuvent être exprimées soit dans l'unité des données brutes, soit en terme standardisé, mais doivent être définies avant la récolte des données (Anderson & Maxwell, 2016 ; Lakens, Scheel, & Isager, 2018). Il existe ensuite plusieurs approches pour démontrer que l'effet observé se situe dans la zone d'équivalence (voir Meyners, 2012, par exemple). Parmis celles-ci, une approche très simple est celle du "Two one-sided tests" (Lakens, 2017 ; Schuirmann, 1987), plus communément appelé le TOST [25]. Le principe est de définir deux hypothèses nulles. La première est que l'effet observé est inférieur à la borne inférieure de la zone d'équivalence:

$$H0_1 : \theta < \theta_1, \ avec\ \theta_1 \neq 0$$

La deuxième est que l'effet observé est supérieur à la borne supérieure de la zone d'équivalence:

$$H0_2 : \theta > \theta_2, \ avec\ \theta_2 \neq 0$$

Lorsque les deux hypothèses nulle peuvent être simultanément rejetées, on peut conclure à l'équivalence (Seaman & Serlin, 1998). Cela équivaut, statistiquement parlant, à montrer que l'intervalle de confiance à $(1-2\times\alpha)\%$ est entièrement inclus dans la zone d'équivalence (Lakens, 2017 ; Seaman & Serlin, 1998). Notons qu'il n'est pas nécessaire de reporter les résulats des deux tests unilatéraux, lorsqu'on réalise le TOST: il suffit de reporter les résultats du test

associé à la plus petite valeur de statistique (et par conséquent, à la plus grande p-valeur). En effet, si ce test amène à conclure au rejet de l'hypothèse nulle, le second test amènera automatiquement à la même conclusion (Lakens, Scheel, & Isager, 2018 ; Rogers, Howard, & Vessey, 1993). Cette remarque reste vraie dans le cas particulier où les deux tests sont associés à la même valeur de statistique puisque dans ce cas, les deux tests mèneront à une conclusion identique (Rogers, Howard, & Vessey, 1993). Notons également qu'il n'est pas nécessaire de procéder à une correction du risque alpha dûe à la réalisation simultanée de deux tests. En effet, une erreur de type I (rejeter à tort l'hypothèse nulle) ne peut être commise que si l'hypothèse nulle est vraie. Or, les deux hypothèses nulles testées sont mutuellement exclusives: il n'est pas possible que θ soit simultanément inférieur à θ_1 (ce qui correspond à $H0_1$) et supérieur à θ_2 (ce qui correspond à $H0_2$).

Jusqu'il y a peu, le TOST n'était pas disponible dans la plupart des logiciels, à l'exception de Minitab, ce qui constituait un frein important à son usage. Pour cette raison, Lakens (2016) a créé le package R "TOSTER" et plus récemment encore, ce même package a été implémenté dans Jamovi [26]. Tant dans R que dans Jamovi, le package compare simultanément l'effet observé à l'absence d'effet (cela correspond au test traditionnel) ainsi qu'aux deux bornes de la zone d'équivalence (cela correspond au TOST). Il en découle 4 conclusions distinctes possibles (Lakens, 2017), qui sont illustrées dans la figure 11 dans le contexte de la comparaison de deux moyennes indépendantes:

(1) La différence de moyenne observée diffère significativement des deux bornes d'équivalence, mais pas de 0 (scénario A, Figure 11): dans ce cas, on conclura à l'absence d'un effet au moins aussi grand que les bornes d'équivalence.

(2) La différence de moyenne observée diffère significativement des deux bornes d'équivalence ainsi que de 0 (scénario B, Figure 11): on conclura alors qu'il existe un effet non nul, mais qui ne dépasse pas une certaine taille fixée par les bornes. C'est ce qui arrive typiquement lorsqu'on travaille avec de très grands échantillons, si bien que le test traditionnel est très puissant, même pour détecter des effets très petits (Rogers, Howard, & Vessey, 1993).

(3) La différence de moyenne observée diffère significativement de 0, mais ne diffère pas significativement d'au moins une des deux bornes d'équivalence (scénario C, Figure 11): on conclura alors à la présence d'un effet non nul (Rogers, Howard, & Vessey, 1993).

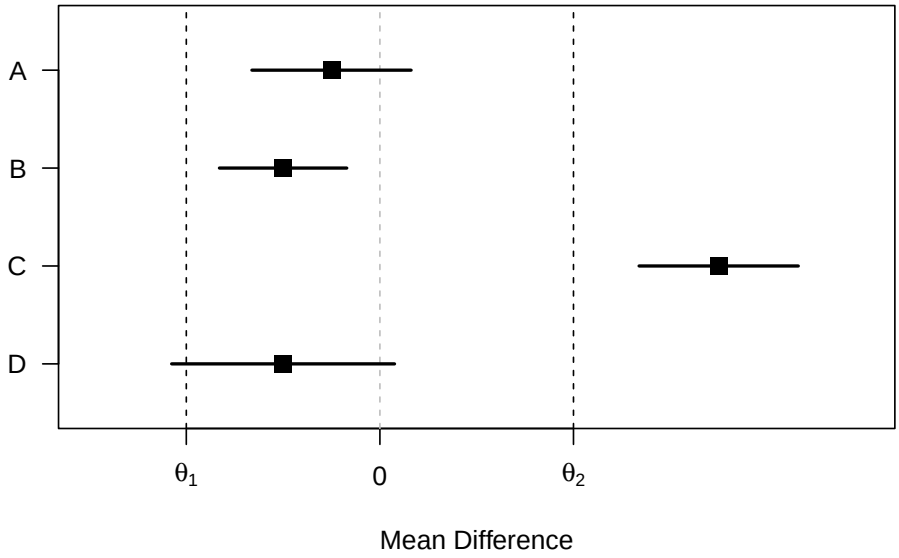


FIGURE 11 – Différence de moyennes ($\bar{X}_1 - \bar{X}_2$) et IC à $1 - 2\alpha\%$ autour de la différence de moyennes ($\bar{X}_1 - \bar{X}_2$) pour 4 scénarios distincts.

(4) La différence de moyenne observée ne diffère significativement ni d'au moins une des deux bornes d'équivalence, ni de 0 (scénario D, Figure 11): c'est ce qui arrive lorsque les données sont si imprécises qu'on ne peut tirer aucune conclusion. Les données semblent compatibles tant avec un effet nul qu'avec un effet supérieur au SESOI.

Définir les bornes de la zone d'équivalence

L'aspect le plus compliqué dans la réalisation du TOST est la définition des bornes d'équivalence. Dans certains cas, il est possible de définir un critère objectif qui permettra de déterminer à partir de quand un effet est jugé pertinent (Lakens, Scheel, & Isager, 2018). Dans ce cas, établir l'équivalence revient à rejeter la présence d'un effet ayant un quelconque intérêt pratique (Rogers, Howard, & Vessey, 1993). Par exemple, Burriss et al. (2015) avaient émis l'hypothèse qu'une augmentation de la rougeur de la peau chez femmes les rendraient plus attractives pour les hommes en période d'ovulation. Or, une telle hypothèse n'est crédible que si le changement facial est visible à l'oeil nu. Dans ce contexte, le SESOI serait la plus petite variation dans la rougeur de la peau qu'il est possible de détecter à l'oeil nu (Lakens, Scheel, & Isager, 2018). Il est

également parfois possible pour des experts de déterminer expérimentalement ce qui constitue un changement important, pour certaines échelles fréquemment utilisées en vue mesurer des construits psychologiques. Button et al. (2015), par exemple, ont interrogé un grand nombre de patients dépressifs quant à leur ressenti subjectif en termes d'amélioration de leur dépression au cours d'un certain laps de temps, et ont comparé leur réponses à la différence de scores obtenus à l'aide du BDI [27] dans ce même laps de temps (Lakens, Scheel, & Isager, 2018). Cela leur a permis de déterminer ce qu'il considérait comme étant une différence pratiquement significative sur l'échelle du BDI. Malheureusement, il n'est pas toujours possible d'établir un critère objectif en vue de définir les bornes d'équivalence. Dans ce cas, il existe diverses stratégies, plus subjectives, en vue d'établir ces bornes. En les utilisant, il faut cependant avoir conscience du fait que la question à laquelle nous répondons varie en fonction de la stratégie utilisée.

Il est possible de déterminer des bornes en s'inspirant de balises existantes, en vue d'exclure la présence d'un effet jugé petit, moyen ou grand par ces balises (Lakens, Scheel, & Isager, 2018). Notons que si cette stratégie est tentante de par sa simplicité, elle doit être utilisée avec prudence. D'abord, un effet ne devrait être qualifié de petit, moyen ou grand qu'en comparaison à d'autres effets connus, et non sur base d'impressions qualitatives (Gignac & Szodorai, 2016). Dit autrement, il est important d'avoir un cadre de référence pour juger de la taille d'un effet. Or, les balises de Cohen (en l'occurence, les balises les plus célèbres et les plus largement utilisées) sont dépourvues de ce cadre de référence, puisqu'elles ont été établies à une époque où très peu de chercheurs se préoccupaient de la taille des effets étudiés (Funder & Ozer, 2019). Depuis Cohen, certains chercheurs ont déployé de gros efforts en vue d'établir de nouvelles normes sur base d'analyses systématiques quantitative de la litérature. Gignac & Szodorai (2016), par exemple, ont établi de nouvelles balises pour interpréter le r de Pearson, en définissant les quartiles d'une distribution de 708 mesures dérivées de méta-analyses issues de la psychologie sociale et de la personnalité. C'est de la sorte qu'ils ont proposé d'interpréter respectivement des mesures de 0.10, 0.20 et 0.30, dans ces domaines de la psychologie, comme représentant des effets relativement petits, typiques et relativement larges. Ces normes ont également été approuvées par Funder & Ozer (2019) [28]. Ensuite, les balises ne prennent pas en compte le contexte de l'étude si bien que statuer sur la taille d'un effet ne fournit pas nécessairement d'information sur sa valeur. Imaginons un antidépresseur B qui permette de réduire très légèrement les symptômes de dépression par rapport à un antidépresseur A déjà présent sur le marché, et qui en outre

coûte moins cher. Même si statistiquement parlant, on observe une très faible taille d'effet, une analyse coût/bénéfice amènerait très vraisemblablement à conclure à la pertinence de cet effet. Pour cela, les balises devraient toujours être utilisées en dernier recours, lorsqu'on ne dispose d'aucune autre information (Gignac & Szodorai, 2016).

La taille des échantillons d'une étude est une information sur laquelle nous pouvons également nous baser en vue de fixer des bornes d'équivalence. Dans le contexte d'une réplication d'étude, cette information peut servir à déterminer si la puissance de l'outil utilisé par le chercheur d'origine était suffisante en vue de tester l'effet ciblé. Une stratégie proposée par Lakens, Scheel, & Isager (2018) consiste à définir comme borne d'équivalence le plus petit effet que l'étude d'origine aurait pu détecter comme étant significative.[29]. L'idée sous-jacente est la suivante: bien qu'idéalement, les chercheurs devraient toujours spécifier ce qu'ils considèrent comme étant le plus petit effet d'intérêt, cette pratique n'est pas encore commune. Heureusement, même lorsqu'un auteur ne statue pas explicitement sur ce qu'il considère comme étant un effet pertinent, la taille des échantillons qu'il utilise aura un impact sur la taille des effets que l'on sera capable de mettre en évidence (en effet, plus les échantillons sont petits, plus l'effet observé doit être grand pour pouvoir être détecté comme étant significatif). Imaginons par exemple qu'un auteur compare les moyennes de deux groupes de 30 sujets à l'aide d'un test t de Student (par facilité, considérons les conditions d'application de ce test comme étant toutes respectées). Avec ces tailles d'échantillon, on conclura au rejet de l'hypothèse nulle si la statistique t de Student vaut au minimum 2.002. Compte tenu de la relation directe entre la statistique t de Student et le d de Cohen (voir chapitre 4), on en déduit qu'on conclura au rejet de l'hypothèse nulle si le d de Cohen est supérieur ou égal à 0.517. Si l'on fixe 0.517 comme borne d'équivalence, et que l'on démontre lors d'une réplication que l'effet observé est significativement inférieur à cette borne, on suspectera que l'étude d'origine n'aurait pu détecter proprement l'effet qu'elle prétendait détecter. Une autre stratégie, proposée par Simonsohn, Nelson, & Simmons (2014) consiste à déterminer la taille d'effet que l'étude d'origine aurait pu détecter avec une puissance de 33%, et à utiliser cette information pour définir la borne d'équivalence. Par exemple, avec une hypothèse bilatérale, un test t de Student aura une puissance de 33% pour détecter un effet de taille 0.399 avec 30 sujets par groupe (à condition que les conditions d'application du test soient toutes respectées). Si lors de la réplication, on obtient un effet significativement inférieur à 0.399, on en déduira que sur le long terme, la probabilité que l'outil d'origine puisse proprement

détecter un effet de cette taille était inférieure à 33%, ce qui remet sa pertinence en cause [30].

Hors du contexte des réplications d'études, on peut également se baser sur la taille des échantillons que l'on est apte à collecter soi-même, en vue de déterminer si l'on dispose d'assez de ressources pour détecter un effet ciblé (Lakens, Scheel, & Isager, 2018). Par exemple, si nous sommes dans l'incapacité de collecter des échantillons de plus de 2000 personnes, il y a certains effets que nous ne pourrons jamais détecter avec une puissance suffisante. Il est possible de déterminer la taille d'effet que nous sommes certains de pouvoir détecter avec suffisamment de puissance (ou autrement dit, dans un pourcentage raisonnable de cas, sur le long terme). *Si l'on utilise cette information pour fixer la (ou les) borne(s) d'équivalence et que l'on conclut effectivement à l'équivalence, la conclussion sera que pour détecter proprement l'effet que nous ciblons, il est indispensable de collecter de plus grands échantillons.*

Comparaison du TOST et du SGPV

Meta-Psychology, 2020, vol 4, MP.2018.933
https://doi.org/10.15626/MP.2018.933
Article type: Original Article

Open data: N/A
Open materials: Yes
Open and reproducible analysis: Yes
Open reviews and editorial process: Yes
Preregistration: N/A

Edited by: S. R. Martin
Reviewed by: J. D. Blume, O. L. Olvera Astivia
Analysis reproduced by: André Kalmendal
All supplementary files can be accessed at OSF:
https://doi.org/10.17605/OSF.IO/ZP3KF

Equivalence Testing and the Second Generation P-Value.

Daniël Lakens
Eindhoven University of Technology, The Netherlands

Marie Delacre
Université Libre de Bruxelles, Belgium

Abstract

To move beyond the limitations of null-hypothesis tests, statistical approaches have been developed where the observed data are compared against a range of values that are equivalent to the absence of a meaningful effect. Specifying a range of values around zero allows researchers to statistically reject the presence of effects large enough to matter, and prevents practically insignificant effects from being interpreted as a statistically significant difference. We compare the behavior of the recently proposed second generation *p*-value (Blume, D'Agostino McGowan, Dupont, & Greevy, 2018) with the more established Two One-Sided Tests (TOST) equivalence testing procedure (Schuirmann, 1987). We show that the two approaches yield almost identical results under optimal conditions. Under suboptimal conditions (e.g., when the confidence interval is wider than the equivalence range, or when confidence intervals are asymmetric) the second generation *p*-value becomes difficult to interpret. The second generation *p*-value is interpretable in a dichotomous manner (i.e., when the SGPV equals 0 or 1 because the confidence intervals lies completely within or outside of the equivalence range), but this dichotomous interpretation does not require calculations. We conclude that equivalence tests yield more consistent *p*-values, distinguish between datasets that yield the same second generation *p*-value, and allow for easier control of Type I and Type II error rates.

Keywords: equivalence testing, second generation p-values, hypothesis testing, TOST, statistical inference

To test predictions researchers predominantly rely on null-hypothesis tests. This statistical approach can be used to examine whether observed data are sufficiently surprising under the null hypothesis to reject an effect that equals exactly zero. Null-hypothesis tests have an important limitation, in that this procedure can only reject the hypothesis that there is no effect, while scientists should also be able to provide statistical support for *equivalence*. When testing for equivalence researchers aim to examine whether an observed effect is too small to be considered meaningful, and therefore is practically equivalent to zero. By specifying a range around the null hypothesis of values that are deemed practically equivalent to the absence of an effect (i.e., 0 +- 0.3) the observed data can be compared against an *equivalence range* and researchers can test if a meaningful effect is absent (Hauck & Anderson, 1984; Kruschke, 2018; Rogers, Howard, & Vessey, 1993; Serlin & Lapsley, 1985; Spiegelhalter, Freedman, & Parmar, 1994; Wellek, 2010; Westlake, 1972).

Second generation *p*-values (SGPV) were recently proposed as a statistic that represents "the proportion of

data-supported hypotheses that are also null hypotheses" (Blume et al., 2018). The researcher specifies an equivalence range around a null hypothesis of values that are considered practically equivalent to the null hypothesis. The SGPV measures the degree to which a set of data-supported parameter values falls within the interval null hypothesis. If the estimation interval falls completely within the equivalence range, the SGPV is 1. If the confidence interval falls completely outside of the equivalence range, the SGPV is 0. Otherwise the SGPV is a value between 0 and 1 that expresses the overlap of data-supported hypotheses and the equivalence range. When calculating the SGPV the set of data-supported parameter values can be represented by a confidence interval (CI), although one could also choose to use credible intervals or Likelihood support intervals (SI). When a confidence interval is used, the SGPV and equivalence tests such as the Two One-Sided Tests (TOST) procedure (Lakens, 2017; Meyners, 2012; Quertemont, 2011; Schuirmann, 1987) appear to have close ties, because both tests compare a confidence interval against an equivalence range. Here, we aim to examine the similarities and differences between the TOST procedure and the SGPV. We limit our analysis to continuous data sampled from a bivariate normal distribution. The TOST procedure also relies on the confidence interval around the effect. In the TOST procedure the data are tested against the lower equivalence bound in the first one-sided test, and against the upper equivalence bound in the second one-sided test (Lakens, Scheel, & Isager, 2018). For an excellent discussion of the strengths and weaknesses of different frequentist equivalence tests, including alternatives to the TOST procedure, see Meyners (2012). If both tests statistically reject an effect as extreme or more extreme than the equivalence bound, you can conclude the observed effect is practically equivalent to zero from a Neyman-Pearson approach to statistical inferences. Because one-sided tests are performed, one can also conclude equivalence by checking whether the 1-2×α confidence interval (e.g., when the alpha level is 0.05, a 90% CI) falls completely within the equivalence bounds. Because both equivalence tests as the SGPV are based on whether and how much a confidence interval overlaps with equivalence bounds, it seems worthwhile to compare the behavior of the newly proposed SGPV to equivalence tests to examine the unique contribution of the SGPV to the statistical toolbox.

The relationship between *p*-values from TOST and SGPV when confidence intervals are symmetrical

The second generation *p*-value (SGPV) is calculated as:

$$p_\delta = \frac{|I \cap H_0|}{|I|} \times \max\left\{\frac{|I|}{2\,|H_0|}, 1\right\}$$

where I is the interval based on the data (e.g., a 95% confidence interval) and H_0 is the equivalence range. The first term of this formula implies that the second generation *p*-value is the width of the confidence interval that overlaps with the equivalence range, divided by the total width of the confidence interval. The second term is a "small sample correction" (which will be discussed later) that comes into play whenever the confidence interval is more than twice as wide as the equivalence range. To examine the relation between the TOST *p*-value and the SGPV we can calculate both statistics across a range of observed effect sizes. Building on the example by Blume et al. (2018), in Figure 1 *p*-values are plotted for the TOST procedure and the SGPV. The statistics are calculated for hypothetical one-sample *t*-tests for observed means ranging from 140 to 150 (on the x-axis). The equivalence range is set to 145 +- 2 (i.e., an equivalence range from 143 to 147), the observed standard deviation is assumed to be 2, and the sample size is 30. For example, for the left-most point in Figure 1 the SGPV and the TOST *p*-value is calculated for a hypothetical study with a sample size of 30, an observed standard deviation of 2, and an observed mean of 140, where the *p*-value for the equivalence test is 1, and the SGPV is 0.

Our conclusions about the relationship between TOST *p*-values and SGPV hold for second generation *p*-values calculated from confidence intervals, and assuming data is sampled from a bivariate normal distribution. Readers can explore the relationship between TOST *p*-values and SGPV for themselves in an online Shiny app: http://shiny.ieis.tue.nl/TOST_vs_SGPV/.

The SGPV treats the equivalence range as the null-hypothesis, while the TOST procedure treats the values outside of the equivalence range as the null-hypothesis. For ease of comparison we can plot 1-SGPV (see Figure 2) to make the values more easily comparable. We see that the *p*-value from the TOST procedure and the SGPV follow each other closely. When we discuss the relationship between the *p*-values from TOST and the SGPV, we focus on their correspondence at three values, namely where the TOST $p = 0.025$ and SGPV is 1, where the TOST $p = 0.5$ and SGPV $= 0.5$, and where the TOST $p = 0.975$ and SGPV $= 1$. These three values are important for the SGPV because they indicate the values at which the SGPV indicates the data should be interpreted as compatible with the null hypothesis (SGPV =

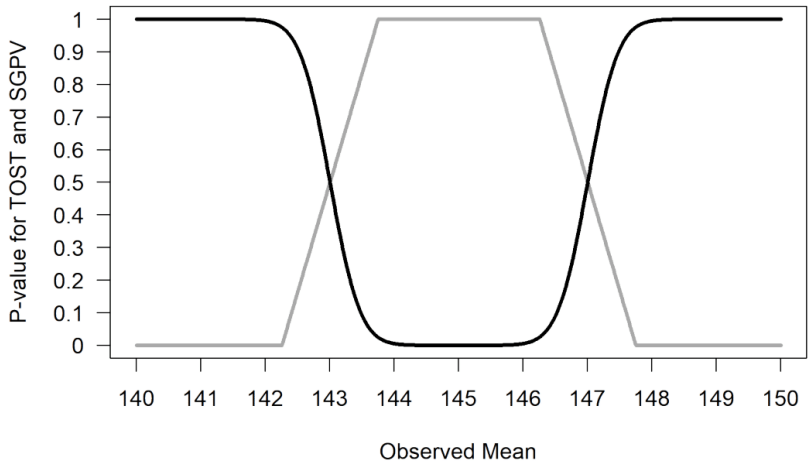


Figure 1. Comparison of *p*-values from TOST (black line) and SGPV (grey line) across a range of observed sample means (x-axis) tested against a mean of 145 in a one-sample *t*-test with a sample size of 30 and a standard deviation of 2, illustrating that when the TOST *p*-value = 0.5, the SGPV = 0.5, when the TOST *p*-value is 0.975, 1-SGPV = 1, and when the TOST *p*-value = 0.025, 1-SGPV = 0.

1), or with the alternative hypothesis (SGPV = 0), or when the data are strictly inconclusive (SGPV = 0.5).

These three points of overlap are indicated by the horizontal dotted lines in Figure 2 at TOST *p*-values of 0.975, 0.5, and 0.025.

When the observed sample mean is 145, the sample size is 30, and the standard deviation is 2, and we are testing against equivalence bounds of 143 and 147 using the TOST procedure for a one-sample *t*-test, the equivalence test is significant, $t(29) = 5.48$, $p < .001$. Because the 95% CI falls completely within the equivalence bounds, the SGPV is 1 (see Figure 1). On the other hand, when the observed mean is 140, the equivalence test is not significant (the observed mean is far outside the equivalence range of 143 to 147), $t(29) = -8.22$, $p = 1$ (or more accurately, $p > .999$ as *p*-values are bounded between 0 and 1). Because the 95% CI falls completely outside the equivalence bounds, the SGPV is 0 (see Figure 1).

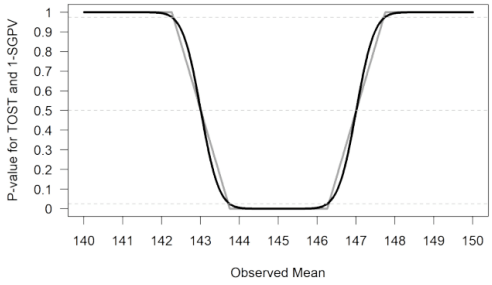


Figure 2. Comparison of *p*-values from TOST (black line) and 1-SGPV (grey line) across a range of observed sample means (x-axis) tested against a mean of 145 in a one-sample *t*-test with a sample size of 30 and a standard deviation of 2.

SGPV as a uniform measure of overlap

It is clear the SGPV and the *p*-value from TOST are closely related. When confidence intervals are symmetric we can think of the SGPV as a straight line that is directly related to the *p*-value from an equivalence test for three values. When the TOST *p*-value is 0.5, the SGPV is also 0.5 (note that the reverse is not true). The

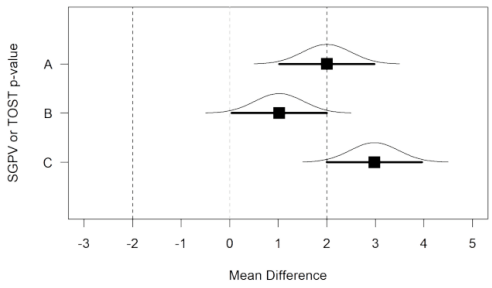


Figure 3. Means, normal distribution, and 95% CI for three example datasets that illustrate the relationship between *p*-values from TOST and SGPV.

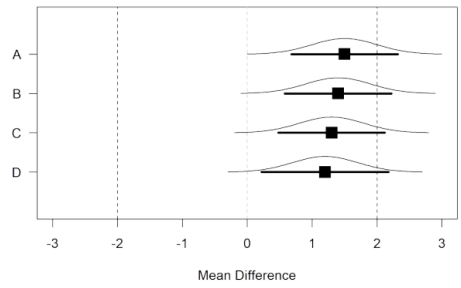


Figure 4. Means, normal distribution, and 95% CI for samples where the observed population mean is 1.5, 1.4, 1.3, and 1.2.

SGPV is 50% when the observed mean falls exactly on the lower or upper equivalence bound, because 50% of the symmetrical confidence interval overlaps with the equivalence range. When the observed mean equals the equivalence bound, the difference between the mean in the data and the equivalence bound is 0, the *t*-value for the equivalence test is also 0, and thus the *p*-value is 0.5 (situation A, Figure 3).

Two other points always have to overlap. When the 95% CI falls completely inside the equivalence region, and one endpoint of the confidence interval is exactly equal to one of the equivalence bounds (see situation B in Figure 3) the TOST *p*-value (which relies on a one-sided test) is always 0.025, and the SGPV is 1. Note that when sample sizes are small or equivalence bounds are narrow, small *p*-values for the TOST or a SGPV = 1 might not be observed in practice if too few observations are collected. The third point where the SGPV and the *p*-value from the TOST procedure should overlap is where the 95% CI falls completely outside of the equivalence range, but one endpoint of the confidence interval is equal to the equivalence bound (see situation C in Figure 3), when the *p*-value will always be 0.975, and the SGPV is 0. Note that this situation is in essence a minimum-effect test (Murphy, Myors, & Wolach, 2014). The goal of a minimum-effect is not just to reject a difference of zero, but to reject the smallest effect size of interest (i.e., the equivalence bounds). An equivalence test and minimum effect test against the same equivalence bound are complementary, and when a TOST *p*-value is larger than 0.975, the *p*-value for the minimum effect test is smaller than 0.05 (and therefore the minimum effect test provides no additional information that can not be derived from the *p*-value from the equivalence test). The SGPV summarizes the information from an equivalence test (and the complementary minimum-effect test). These can be two relevant questions to ask, although it often makes sense to combine an equivalence test and a null-hypothesis test instead (Lakens et al., 2018).

For example, in Figure 4 we have plotted four SGPV's. From A to D the SGPV is 0.76, 0.81, 0.86, and 0.91. The difference in the percentage of overlap between A and B (-0.05) is identical to the difference in the percentage of overlap between C and D as the mean gets 0.1 closer to the test value (-0.05). As the observed mean in a one-sample *t*-test lies closer to the test value, from situation A to D, the difference in the overlap changes uniformly. As we move the observed mean closer to the test value in steps of 0.1 across A to D the *p*-value calculated for normally distributed data are not uniformly distributed. The probability of observing data more extreme than the upper bound of 2 is (from A to D) 0.16, 0.12, 0.08, and 0.05. As we can see, the difference between A and B (0.04) is not the same as the difference between C and D (0.03). Indeed, the difference in *p*-values is the largest as you start at $p = 0.5$ (when the observed mean falls on the test value), which is why the line in Figure 1 is the steepest at $p = 0.5$. Note that where the SGPV reaches 1 or 0, *p*-values closely approximate 0 and 1, but never reach these values.

When different *p*-values for equivalence tests yield the same SGPV

There are three situations where *p*-values for TOST differentiate between observed results, while the SGPV does not differentiate. The first two situations were discussed before and can be seen in Figure 1. When

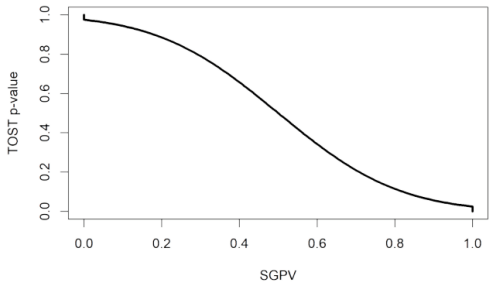


Figure 5. The relationship between *p*-values from the TOST procedure and the SGPV for the same scenario as in Figure 1.

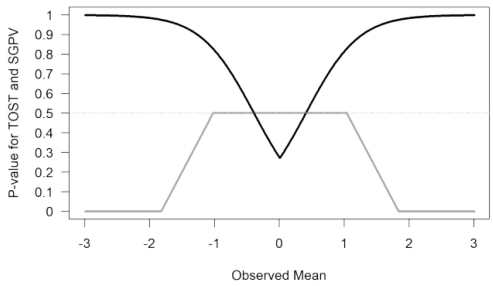


Figure 6. Comparison of *p*-values from TOST (black line) and SGPV (grey line) across a range of observed sample means (x-axis). Because the sample size is small (n = 10) and with a standard deviation of 2 the CI is more than twice as wide as the equivalence range (set to -0.4 to 0.4), the SGPV is set to 0.5 (horizontal light-grey line) across a range of observed means.

the SGPV is either 0 or 1, *p*-values from the equivalence test fall between 0.975 and 1 or between 0 and 0.025. Where the SGPV is 1 as long as the confidence interval falls completely within the equivalence bounds, the *p*-value for the TOST continues to differentiate between results as a function of how far the confidence interval lies within the equivalence bounds (the further the confidence interval is from both bounds, the lower the *p*-value). The easiest way to see this is by plotting the SGPV against the *p*-value from the TOST procedure. The situations where the *p*-values from the TOST procedure continue to differentiate based on how extreme the results are, but the SGPV is a fixed value are indicated by the parts of the curve where there are vertical straight lines at second generation *p*-values of 0 and 1.

A third situation in which the SGPV remains stable across a range of observed effects, while the TOST *p*-value continues to differentiate, is whenever the CI is wider than the equivalence range, and the CI overlaps with the upper *and* lower equivalence bound. When the confidence interval is more than twice as wide as the equivalence range the SGPV is set to 0.5. Blume et al. (2018) call this the "small sample correction factor". However, it is not a correction in the typical sense of the word, since the SGPV is not adjusted to any "correct" value. When the normal calculation would be "misleading" (i.e., the SGPV would be small, which normally would suggest support for the alternative hypothesis, but at the same time all values in the equivalence range are supported), the SGPV is set to 0.5 which according to Blume and colleagues signals that the SGPV is "uninformative". Note that the CI can be twice as wide as the equivalence range whenever the sample size is small (and the confidence interval width is large) *or* when then equivalence range is narrow. It is therefore not so much a "small sample correction" as it is an exception to the typical calculation of the SGPV whenever the ratio of the confidence interval width to the equivalence range exceeds 2:1 and the CI overlaps with the upper and lower bounds.

We can examine this situation by calculating the SGPV and performing the TOST for a situation where sample sizes are small and the equivalence range is narrow, such that the CI is more than twice as large as the equivalence range (see Figure 6). When the two statistics are plotted against each other we can see where the SGPV is the same while the TOST *p*-value still differentiates different observed means (indicated by straight lines in the curve, see Figure 7). We see the SGPV is 0.5 for a range of observed means where the *p*-value from the equivalence test still varies. It should be noted that in these calculations the *p*-values for the TOST procedure are *never* smaller than 0.05 (i.e., they do not get below 0.05 on the y-axis). In other words, we cannot conclude equivalence based on any of the observed means. This happens because we are examining a scenario where the 90% CI is so wide that it never falls completely within the two equivalence bounds.

As Lakens (2017) notes: "in small samples (where CIs are wide), a study might have no statistical power (i.e., the CI will always be so wide that it is necessarily wider than the equivalence bounds)." None of the *p*-values based on the TOST procedure are below 0.05, and thus, in the long run we have 0% power.

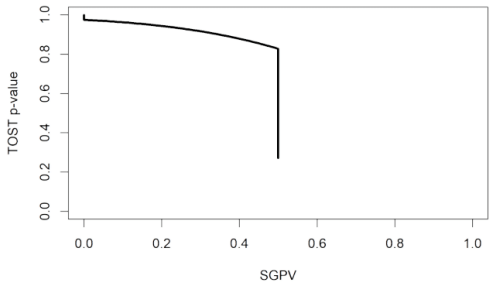


Figure 7. The relationship between *p*-values from the TOST procedure and the SGPV for the same scenario as in Figure 6.

The *p*-value from the TOST procedure still differentiates observed means, while the SGPV does not, when the CI is wider than the equivalence range (so the precision is low) and overlaps with the upper and lower equivalence bound, but the CI is *not* twice as wide as the equivalence range. In the example below, we see that the CI is only 1.79 times as wide as the equivalence bounds, but the CI overlaps with the lower and upper equivalence bounds (Figure 8). This means the SGPV is not set to 0.5, but it is constant across a range of observed means, while the TOST *p*-value is not constant across this range.

If the observed mean would be somewhat closer to 0, or further away from 0, the SGPV remains constant (the CI width does not change, and it completely overlaps with the equivalence range) while the *p*-value for the TOST procedure does vary. We can see this in Figure 9 below. The SGPV is not set to 0.5, but is slightly higher than 0.5 across a range of means. How high the SGPV will be for a CI that is not twice as wide as the equivalence range, but overlaps with the lower and upper equivalence bounds, depends on the width of the CI and the equivalence range.

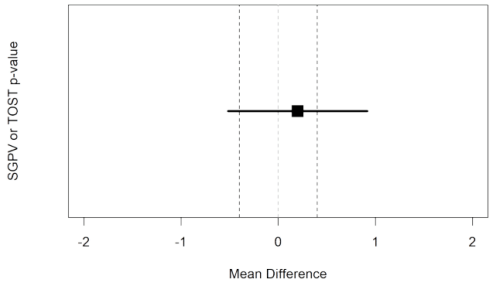


Figure 8. Example of a 95% CI that overlaps with the lower and upper equivalence bound (indicated by the vertical dotted lines).

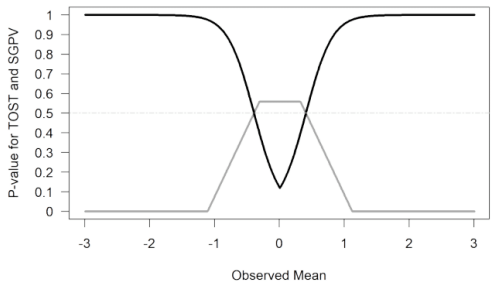


Figure 9. Comparison of *p*-values from TOST (black line) and SGPV (grey line) across a range of observed sample means (x-axis). The sample size is small (n = 10), but because the sd is half as big as in Figure 7 (1 instead of 2) the CI is less than twice as wide as the equivalence range (set to -0.4 to 0.4). The SGPV is not set to 0.5 (horizontal light grey line) but reaches a maximum slightly above 0.5 across a range of observed means.

If we once more plot the two statistics against each other we see the SGPV is 0.56 for a range of observed means where the *p*-value from the equivalence test still varies, as indicated by the straight section of the line (Figure 10).

To conclude this section, there are situations where the *p*-value from the TOST procedure continues to differentiate, while the SGPV does not. Therefore, interpreted as a continuous statistic, the SGPV is more limited than the *p*-value from the TOST procedure.

The relation between equivalence tests and SGPV for asymmetrical confidence intervals around correlations

So far we have only looked at the relation between equivalence tests and the SGPV when confidence intervals are symmetric (e.g., for confidence intervals around mean differences). For correlations, which are bound between -1 and 1, confidence intervals are only symmetric for a correlation of exactly 0. The confidence interval for a correlation becomes increasingly asymmetric

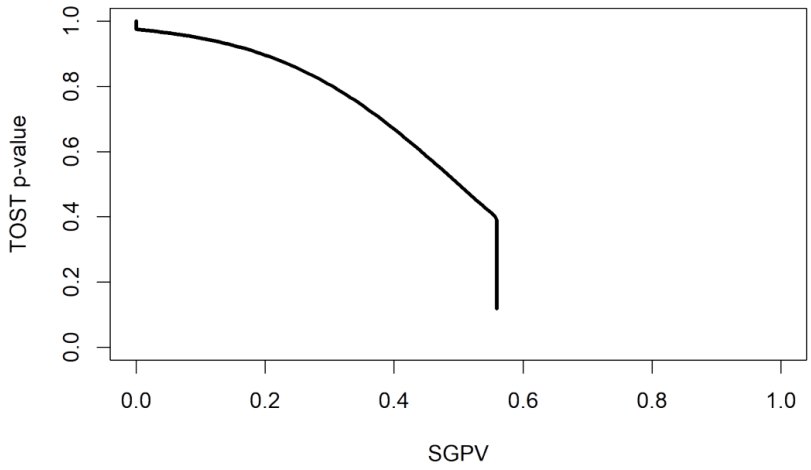


Figure 10. The relationship between *p*-values from the TOST procedure and the SGPV for the same scenario as in Figure 9.

as the observed correlation nears -1 or 1. For example, with ten observations, an observed correlation of 0 has a symmetric 95% confidence interval ranging from -0.63 to 0.63, while and observed correlation of 0.7 has an asymmetric 95% confidence interval ranging from 0.13 to 0.92. Note that calculating confidence intervals for a correlation involves a Fisher's z-transformation, which transforms values such that they are approximately normally z-distributed, which allows one to compute symmetric confidence intervals. These confidence intervals are then retransformed into a correlation, where the confidence intervals are asymmetric if the correlation is not exactly zero.

The effect of asymmetric confidence intervals around correlations is most noticeable at smaller sample sizes. In Figure 11 we plot the *p*-values from equivalence tests and the SGPV (again plotted as 1-SGPV for ease of comparison) for correlations. The sample size is 30 pairs of observations, and the lower and upper equivalence bounds are set to -0.45 and 0.45, with an alpha of 0.05. As the observed correlation in the sample moves from -.99 to 0 the *p*-value from the equivalence test becomes smaller, as does 1-SGPV. The pattern is quite similar to that in Figure 2. The *p*-value for the TOST procedure and 1-SGPV are still related as discussed above, with TOST *p*-values of 0.975 and 0.025 corresponding to a 1-SGPV of 1 and 0, respectively. There are two important differences, however. First of all, the SGPV is no longer a straight line, but a curve, due to the asymmetry in the 95% CI. Second, and most importantly, the *p*-value for the equivalence test and the SGPV do no longer overlap at $p = 0.5$.

The reason that the equivalence test and SGPV no longer overlap is due to asymmetric confidence intervals. If the observed correlation falls exactly on the equivalence bound the *p*-value for the equivalence test is 0.5. In the equivalence test for correlations the *p*-value is computed based on a z-transformation which better controls error rates (Goertzen & Cribbie, 2010). This transformation is computed as follows, where r is the observed correlation and ρ is the theoretical correlation under the null:

$$z = \frac{\frac{\log(\frac{1+r}{1-r})}{2} - \frac{\log(\frac{1+\rho}{1-\rho})}{2}}{\sqrt{\frac{1}{n-3}}}$$

Because the z-distribution is symmetric, the probability of observing the observed or more extreme z-score, assuming the equivalence bound is the true effect size, is 50%. However, because the r distribution is not symmetric, this does not mean that there is always a 50% probability of observing a correlation smaller or larger

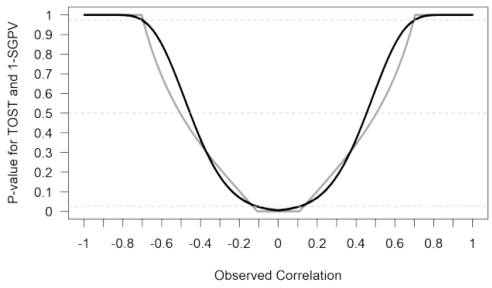


Figure 11. Comparison of *p*-values from TOST (black line) and 1-SGPV (grey curve) across a range of observed sample correlations (x-axis) tested against equivalence bounds of r = -0.45 and r = 0.45 with n = 30 and an alpha of 0.05.

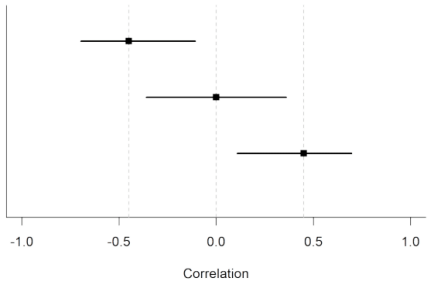


Figure 12. Three 95% confidence intervals for observed effect sizes of r = -0.45, r = 0, and r = 0.45 for n = 30. Only the confidence interval for r = 0 is symmetric.

than the true correlation. As can be seen in Figure 12, the proportion of the confidence interval that overlaps with the equivalence range is larger than 50% when the observed correlations are r = -.45 and r = .45, meaning that the two second generation p-values associated with these correlations are larger than 50%. Because the confidence intervals are asymmetric around the observed effect size of 0.45 (ranging from 0.11 to 0.70) according to Blume et al. (2018) 58.11% of the data-supported hypotheses are null hypotheses, and therefore 58.11% of the data-supported hypotheses are compatible with the null premise.

The further away from 0, the larger the SGPV when the observed mean falls on the equivalence bound. The SGPV is the proportion of values in a 95% confidence interval that overlap with the equivalence range, but not the probability that these values will be observed. In the most extreme case (i.e., a sample size of 4, and equivalence bounds set to r = -0.99 and 0.99, with a true correlation of 0.99) 97.60% of the confidence interval overlaps with the equivalence range, even though in the long run only 36% of the correlations observed in the future will fall in this range.

It should be noted that in larger sample sizes the SGPV is closer to 0.5 whenever the observed correlation falls on the equivalence bound, but this extreme example nevertheless clearly illustrates the difference between question the SGPV answers, and the question a *p*-value answers. The conclusion of this section on asymmetric confidence intervals is that a SGPV of 1 or 0 can still be interpreted as a $p < 0.025$ or $p > 0.975$ in an equivalence test, since the SGPV and *p*-value for the TOST procedure are always directly related at the values $p = 0.025$ and $p = 0.975$. Although Blume et al. (2018) state that "the degree of overlap conveys how compatible the data are with the null premise" this definition of what the SGPV provides does not hold for asymmetric confidence intervals. Although a SGPV of 1 or 0 can be directly interpreted, a SGPV between 0 and 1 is not interpretable as "compatibility with the null hypothesis" under the assumption of a bivariate normal distribution, and the generalizability of this statement needs to be examined beyond normal bivariate distributions. Indeed, Blume and colleagues write in the supplemental material that "The magnitude of an inconclusive second-generation *p*-value can vary slightly when the effect size scale is transformed. However definitive findings, i.e. a *p*-value of 0 or 1 are *not* affected by the scale changes."

What are the Relative Strengths and Weaknesses of Equivalence Testing and the SGPV?

When introducing a new statistical method, it is important to compare it to existing approaches and specify its relative strengths and weaknesses. Here, we aimed to compare the SGPV against equivalence tests based on the TOST procedure. First of all, even though a SGPV of 1 or 0 has a clear interpretation (we can reject effects outside or inside the equivalence range), intermediate values are not as easy to interpret (especially for effects that have asymmetric confidence intervals). In one sense, they are what they are (the proportion of overlap), but it can be unclear what this number tells us about the data we have collected. This is not too problematic, since the main use of the SGPV (e.g., in all examples provided by Blume and colleagues) seems to be to examine whether the SGPV is 0, 1, or inconclusive.

As already mentioned, this interpretation of a SGPV is very similar to the Neyman-Pearson interpretation of an equivalence test and a minimum effect tests (which are complementary). The difference is that where a SGPV of 1 can be interpreted as $p < .025$, equivalence tests provide exact p-values, and they continue to differentiate between for example $p = 0.024$ and $p = 0.002$. Whether this is desirable depends on the perspective that is used. From a Neyman-Pearson perspective on statistical inferences the main conclusion is based on whether or not $p < \alpha$, and thus an equivalence test and SGPV can be performed by simply checking whether the confidence interval falls within the equivalence range, just as a null-hypothesis test can be performed by checking whether the confidence interval contains zero or not. At the same time, it is recommended to report exact p-values (American Psychological Association, 2010), and exact p-values might provide information of interest to readers about how precisely how surprising the data, or more extreme data, is under the null model. Some researchers might be interested in combining an equivalence test with a null-hypothesis significance test. This allows a researcher to ask whether there is an effect that is statistically different from zero, and whether effect sizes that are considered meaningful can be rejected. Equivalence tests combined with null-hypothesis tests classify results into four possible categories, and for example allow researchers to conclude an effect is significant *and* equivalent (i.e., statistically different from zero, but also too small to be considered meaningful; see Lakens et al., 2018).

An important issue when calculating the SGPV is its reliance on the "small sample correction", where the SGPV is set to 0.5 whenever the ratio of the confidence interval width to the equivalence range exceeds 2:1 and the CI overlaps with the upper and lower bounds. This exception to the normal calculation of the SGPV is introduced to prevent misleading values. Without this correction it is possible that a confidence interval is extremely wide, and an equivalence range is extremely narrow, which without the correction would lead to a very low value for the SGPV. Blume et al. (2018) suggest that under such a scenario "the data favor alternative hypotheses", even when a better interpretation would be that there is not enough data to accurately estimate the true effect compared to the width of the equivalence range. Although it is necessary to set the SGPV to 0.5 whenever the ratio of the confidence interval width to the equivalence range exceeds 2:1, it leads to a range of situations where the SGPV is set to 0.5, while the p-value from the TOST procedure continues to differentiate (see for example Figure 6). An important benefit of equivalence tests is that it does not need such a correction to prevent misleading results.

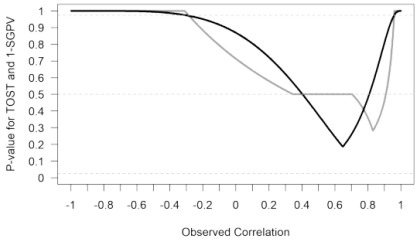


Figure 13. Comparison of p-values from TOST (black line) and 1-SGPV (grey curve) across a range of observed sample correlations (x-axis) tested against equivalence bounds of $r = 0.4$ and $r = 0.8$ with n = 10 and an alpha of 0.05.

As a more extreme example of the peculiar behavior of the "small sample correction" as currently implemented in the calculation of the SGPV, see Figure 13. In this figure observed correlations (from a sample size of 10) from -.99 to .99 are tested against an equivalence range from r = 0.4 to r = 0.8. We can see the SGPV has a peculiar shape because it is set to 0.5 for certain observed correlations, even though there is no risk of a "misleading" SGPV in this range. This example suggests that the current implementation of the "small sample correction" could be improved. If, on the other hand, the SGPV is mainly meant to be interpreted when it is 0 or 1, it might be preferable to simply never apply the "small sample correction".

Blume et al. (2018) claim that when using the SGPV "Adjustments for multiple comparisons are obviated" (p. 15). However, this is not correct. Given the direct relationship between TOST and SGPV highlighted in this manuscript (where a TOST $p = 0.025$ equals SGPV = 1, as long as the SGPV is calculated based on confidence intervals, and assuming data are sampled from a continuous bivariate normal distribution), not correcting for multiple comparisons will inflate the probability of concluding the absence of a meaningful effect based on the SGPV in exactly the same way as it will for equivalence tests. Whenever statistical tests are interpreted as support for a hypothesis (e.g., SPGV = 0 or SGPV = 1), it is possible to do so erroneously, and if researchers want to control error rates, they need to correct for multiple comparisons.

Conclusion

We believe that our explanation of the similarities between the TOST procedure and the SGPV provides context to interpret the contribution of second generation *p*-values to the statistical toolbox. The novelty of the SGPV can be limited when confidence intervals are asymmetrical or wider than the equivalence range. There are strong similarities with *p*-values from the TOST procedure, and in all situations where the statistics yield different results, the behavior of the *p*-value from the TOST procedure is more consistent. We hope this overview of the relationship between the SGPV and equivalence tests will help researchers to make an informed decision about which statistical approach provides the best answer to their question. Our comparisons show that when proposing alternatives to null-hypothesis tests, it is important to compare new proposals to already existing procedures. We believe equivalence tests achieve the goals of the second generation *p*-value while allowing users to easily control error rates, and while yielding more consistent statistical outcomes.

Authors Note

All code associated with this article, including the reproducible manuscript, is available from https://github.com/Lakens/TOST_vs_SGPV and https://osf.io/8crkg/. The preprint can be found at https://psyarxiv.com/7k6ay/.

Correspondence concerning this article should be addressed to Daniël Lakens, Den Dolech 1, IPO 1.33, 5600 MB, Eindhoven, The Netherlands. E-mail: D.Lakens@tue.nl

Open Science Practices

This article earned the Open Materials badge for making the materials openly available. It has been verified that the analysis reproduced the results presented in the article. The entire editorial process, including the open reviews, are published in the online supplement.

Conflict of Interest and Funding

No conflict of interest and no external funding. This work was supported by the NetherlandsOrganization for Scientific Research (NWO) VIDI grant 452-17-013.

Author Contributions

DL conceptualized the idea, both authors wrote and revised this manuscript.

References

American Psychological Association (Ed.). (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.

Blume, J. D., D'Agostino McGowan, L., Dupont, W. D., & Greevy, R. A. (2018). Second-generation p-values: Improved rigor, reproducibility, & transparency in statistical analyses. *PLOS ONE*, *13*(3), e0188299. doi:10.1371/journal.pone.0188299

Goertzen, J. R., & Cribbie, R. A. (2010). Detecting a lack of association: An equivalence testing approach. *British Journal of Mathematical and Statistical Psychology*, *63*(3), 527–537. doi:10.1348/000711009X475853

Hauck, D. W. W., & Anderson, S. (1984). A new statistical procedure for testing equivalence in two-group comparative bioavailability trials. *Journal of Pharmacokinetics and Biopharmaceutics*, *12*(1), 83–91. doi:10.1007/BF01063612

Kruschke, J. K. (2018). Rejecting or Accepting Parameter Values in Bayesian Estimation. *Advances in Methods and Practices in Psychological Science*, *1*(2), 270–280. doi:10.1177/2515245918771304

Lakens, D. (2017). Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses. *Social Psychological and Personality Science*, *8*(4), 355–362. doi:10.1177/1948550617697177

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269. doi:10.1177/2515245918770963

Meyners, M. (2012). Equivalence tests A review. *Food Quality and Preference*, *26*(2), 231–245. doi:10.1016/j.foodqual.2012.05.003

Murphy, K. R., Myors, B., & Wolach, A. H. (2014). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests* (Fourth edition.). New York: Routledge, Taylor & Francis Group.

Quertemont, E. (2011). How to Statistically Show the Absence of an Effect. *Psychologica Belgica*, *51*(2), 109–127. doi:10.5334/pb-51-2-109

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin*, *113*(3), 553–565. doi:http://dx.doi.org/10.1037/0033-2909.113.3.553

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, *15*(6), 657–680. doi:https://doi.org/10.1007/BF01068419

Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist*, *40*(1), 73–83. doi:http://dx.doi.org/10.1037/0003-066X.40.1.73

Spiegelhalter, D. J., Freedman, L. S., & Parmar, M. K. (1994). Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 357–416. doi:10.2307/2983527

Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority* (2nd ed.). Boca Raton: CRC Press.

Westlake, W. J. (1972). Use of Confidence Intervals in Analysis of Comparative Bioavailability Trials. *Journal of Pharmaceutical Sciences*, *61*(8), 1340–1341. doi:10.1002/JPS.2600610845

Notes de fin de chapitre

[20]Nous observons en réalité des proportions qui varient de 94.8% à 95%, à cause du hasard d'échantillonnage, mais sur le long terme, lorsque le nombre d'itérations tend vers l'infini, toutes les proportions de non rejet de l'hypothèse nulle quand l'hypothèse nulle est vraie vont tendre vers $(1 - \alpha)\%$.

[21]On parle d'estimation et non de mesure, car la puissance du test dépend de σ, l'écart-type de la population, qu'on ne connait pas et devra donc estimer sur base de S, l'écart-type de l'échantillon (Schuirmann,1987).

[22]Avec 100 sujets par groupe, on estime la puissance du test à 80% lorsque l'estimation d de Cohen vaut .3981. Par conséquent, un test sera susceptible de conclure à l'équivalence si les bornes de la zone d'équivalence, exprimée en mesure standardisée d de Cohen, sont supérieures ou égales à .3981. Lorsqu'on fixe les bornes aux différences de moyennes $\pm.3$, cela n'est possible que si S est inférieur ou égal à .7535. En effet, $d = \frac{\theta}{S} \leftrightarrow .3981 = \frac{.3}{S} \leftrightarrow S = \frac{.3}{.3981} = .7535$. Or, avec 100 sujets par groupe, aucune estimation S ne sera inférieure ou égale à .7535 lorsque σ vaut 1.

[23]En comparant les Tables 1 et 2, on constate qu'avec 200 sujets par groupes, les proportions d'itérations de chaque scénario qui amènent à conclure à l'équivalence, dans la Table 2, sont inférieures aux proportions d'itérations de chaque scénarios qui amènent à ne pas rejeter l'hypothèse nulle, dans la Table 1. Plus les échantillons sont grands, moins on observera d'écart entre les 2 tables car la proportion d'itérations qui ne pourront amener à conclure à l'équivalence en raison d'un manque de puissance diminuera. En effet, plus la taille des échantillons sera grande, plus la valeur maximale de S permettant d'assurer la puissance des 80% sera élevée. Par exemple, avec 200 sujets par groupes, la valeur maximale autorisée pour S sera de $\frac{.3}{.2808} = 1.07$. Avec 300 sujets par groupes, la valeur maximale autorisée pour S sera de $\frac{.3}{.2291} = 1.31$. Par conséquent, plus les tailles d'échantillons seront grandes, moins il sera probable que S dépasse le seuil autorisé.

[24]Par facilité, à l'instar de Schuirman (1987), on envisage le cas où les échantillons sont de même taille et que l'on suppose que la condition d'homogénéité des variances est respectée. Notons cependant que d'après Schuirman, ce raisonnement peut être généralisé aux scénarios où les deux échantillons n'ont pas la même taille et sont extraits de population n'ayant pas la même variance.

[25]Il existe des alternatives au TOST qui sont très légèrement plus puissantes, mais le gain marginal en termes de puissance est contrebalancé par un niveau de complexité beaucoup plus élevé (Meyners, 2012).

[26]Jamovi est un logiciel clic-bouton entièrement gratuit qui gagne en popularité et qui présente, parmi ses nombreux avantages, le fait d'être particulièrement convivial. Dans la mesure où la plupart des chercheurs sont plus enclins à utiliser des procédures si elles sont implémentées dans ce type de logiciel (Fraas & Newman, 2000), cela constitue une excellente nouvelle pour le devenir du TOST dans la recherche en psychologie.

[27]Le BDI (Beck Depression Inventory) est une échelle auto-rapportée évaluant les symptômes cognitifs courants

de la dépression. Cette échelle est constituée de 21 items évalués à l'aide des échelles de Likert allant de 0 à 3, ce qui donne un score total compris entre 0 et 63 qui sera d'autant plus élevé que la dépression sera sévère (Button et al., 2015).

[28]Funder & Ozer (2019) ont relevé plusieurs enquêtes ayant calculé un r de Pearson moyen de .21 sur base d'effets publiés dans la litérature en psychologie sociale et en psychologie de la personnalité. Par ailleurs, ils rappellent qu'en raison du biais de publication, un chercheur obtenant un r de Pearson de .21 dans une nouvelle étude peut être assuré d'avoir détecté un effet plus grand que généralement trouvé

[29]Dans le contexte d'une réplication, il nous semble souvent plus logique de réaliser le test d'équivalence en unilatéral, puisque l'étude d'origine précisera généralement le sens de l'effet observé. C'est pourquoi nous parlons ici de bornd d'équivalence au singulier

[30]Bien entendu, la valeur 33% a une dimension arbitraire, comme chaque fois que l'on fixe une valeur par défaut.

Discussion générale et conclusion

A travers cette thèse, nos objectifs de départ étaient :

(1) D'identifier des manquements dans les pratiques actuelles des chercheurs, via des analyses d'articles publiés dans des revues de psychologie ;

(2) De réaliser des simulations, en vue de montrer l'impact de ces pratiques ;

(3) De proposer des recommandations pour les améliorer.

Nous nous sommes d'abord focalisés sur l'usage de mesures communément utilisées lorsqu'on compare les moyennes de deux ou plusieurs groupes, et qui reposent sur des conditions d'application paramétriques: le test t de Student, l'ANOVA F de Fisher et le d de Cohen.

Le test t de Student (chapitre 2) est un test de comparaison de moyennes qui repose théoriquement sur les conditions que ... Or, ces conditions sont rarément rencontrées en situation réelles. Après avoir expliqué théoriquement en quoi ces conditions sont peu réalistes, nous avons réalisé des simulations Monte Carlo en vue de montrer les conséquences réelles de leurs violations et de comparer les résultats du test t de Student à ceux du test t de Welch sous diverses conditions. Il est essentiellement ressorti de ces simulations que conformément à nos attentes théoriques, sur le long-terme, utiliser le test t de Student en cas de violation de la condition d'homogénéité des variances est bien plus susceptible d'altérer les proportions de faux positifs et de faux négatifs que l'utiliser en cas de violation de la condition de normalité des résidus *réfléchir pour être sûre que je peux bien parler de faux positifs et faux négatifs plutôt que taux d'erreur de type I et II car c'est pas tt à fait la même chose*. Si le test t de Student est suffisamment robuste à l'hétérogénéité des variances lorsque les deux échantillons extraits sont de même taille, ce n'est pas le cas lorsque les deux échantillons sont de taille différence: lorsque le plus petit groupe est extrait de la population ayant la plus grande variance, on observe une inflation du taux d'erreur de type I et la courbe de puissance théorique sous-estime la puissance du test. Au contraire, lorsque le plus petit groupe est extrait de la population ayant la plus petite variance, on observe une diminution du taux d'erreur de type I et une sous-estimation de la puissance théorique du test. Conscients de ces limites, on conseille souvent aux chercheurs de réaliser d'abord le test de Levene (un test d'homogénéité des variances) et de remplacer les tests t de Student par le test t de Welch lorsque le test de Levene amène

à rejeter l'hypothèse d'égalité des variances. Contrairement à cette recommandation, nous prônons en faveur de l'usage systématique du test t de Welch, au moins lorsque les échantillons sont de taille différence. En effet, il est très difficile (voire impossible) de montrer efficacement le respect de la condition d'homogénéité des variances, cet état étant par ailleurs plus souvent l'exception que le norme, et même dans le cas où cette condition est respectée, les écarts de puissance entre le test t de Welch et celui de Student sont très faibles.

Lorsque l'on compare les moyennes de plus de deux groupes, on utilise communément l'ANOVA F de Fisher qui est, statiquement parlant, une généralisation du test t de Student et qui repose sur les mêmes conditions (chapitre 3). A travers nos simulations, nous avons révélé que lorsqu'on compare plus de deux groupes indépendants, l'ANOVA F de Fisher est perturbée par l'hétérogénéité des variances, même lorsque tous les échantillons sont de même tailles, ce qui nous a amené à recommander de privilégier systématiquement l'ANOVA de Welch plutôt que l'ANOVA de Fisher.

Ecrire un petit paragraphe sur le d de Cohen.

Ecrire sur les tests d'équivalence. *Nous nous sommes ensuite intéressés à la définition courante des hypothèses nulles comme hypothèse d'absence d'effet ainsi qu'à un outil qui permet, au contraire, d'étudier des hyopthèses nulles de présence d'effet (voir le chapitre 5).*

Apports théoriques et appliqués

Apports théoriques

Afin d'aider les chercheurs à prendre conscience des conséquences néfastes de l'usage récurrent des tests t de Student, F de Fisher et du d de Cohen, conformément à ce que nous avions annoncé en introduction, nous avons fourni des exemples concrets et issus de la psychologie expliquant en quoi la condition d'homoscédasticité n'était pas réaliste et les conséquences réelles de ces violations.

[Un consultant doit pouvoir parler de langage des psycs, 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 plsu à mˆmee 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))].

Beaucoup avant nous ont tenté de faire passer ce message. L'idée de remplacer le test t de Student par le test de Welch n'est pas récente. Mais les moyens informatiques actuels nous ont permis la réalisation de simulations intensives, avec 1,000,000 d'itérations pour un très grand nombre de scénarios qui varient en fonction du niveau d'hétéroscédasticité et du degré d'écart aux conditions d'application. En clair, nous avons remis au goût du jour une problématique ancienne, mais avec des simulations bcp plus poussées car pas possible avant.

De plus, nous avons tout mis en oeuvre pour rendre notre recherche la plus tranparente possible et pour assurer une grande accessibilité pour tous les chercheurs en psychologie. - Premièrement, nous avons systématiquement diffusé des preprint, chaque fois que nous avons soumis un article. *Cette pratique a eu des retombées positives qui ont dépassé nos attentes, grâce à la collaborations d'éminents chercheurs (exemple du dernier article qui a été très longuement commenté par Cumming),*
- Deuxièmement, nous avons systématiquement privilégié des revues Open Access (*International Review of Social Psychology*, *British Journal of Mathematical and Statistical Science*, *AMPPS*).
- Troisièmement, nous avons utilisé Github.
Nous semblons avoir relevé le défi de la diffusion et de l'impact théorique puisque nos articles ont un taux de citation élevé. Notre article sur le test t de Welch (chapitre 2) a à ce jour été cité 387 fois. Notre article sur l'ANOVA de Welch (chapitre 3) a déjà été cité 61 fois et le preprint sur les mesures de taille d'effet a été source d'intérêt pour plusieurs chercheurs.

Bien que nos trois premiers articles soulignent des questions importantes, toutes les recommandations que nous y faisons permettent d'améliorer la fiabilité des résultats en cas de rejet de l'hypothèse nulle, mais n'apportent aucune solution pour démontrer l'*absence* d'effet. Contrairement au TOST. Ce test commence à gagner en popularité, notamment grâce aux travaux de Lakens. Même si ce test est simple et efficace, il est important de continuer à le comparer à de nouvelles techniques existantes. C'est ce que nous avons fait à travers notre article comparant le TOST et le SGPV *relire article pour écrire un petit paragraphe là-dessus).*

A nouveau, nous avons attaché beaucoup d'importance à la transparence et à l'accessibilité de ce travail. Nous avons cette fois encore réalisé un preprint et régulièrement mis à jour un lien Github permettant d'appréhender l'évolution de l'étude. L'article a été soumis dans la revue "MetaPsychology". Dans cette revue, même le processus re Reviewing est transparent!

Apports appliqués

Bien qu'il est très important d'aider les chercheurs à comprendre, théoriquement, en quoi les pratiques actuelles sont problématiques, il nous semble plus important encore de leur fournir des outils, afin de leur permettre de modifier ces pratiques. Or, un article méthodologique à lui seul suffit rarement à cela (d'après Mills, Abdulla, & Cribbie (2010), les chercheurs appliqués citent très peu les articles méthodologiques dans leur référence pour justifier leurs choix, ce qui pourrait être un signal du fait qu'ils basent peu leurs décisions sur ces articles).

Dans l'autre sens, on constate que les articles méthodos sont généralement peu cités, et ils le sont encore 3 fois moins par les chercheurs appliqués que par les autres méthodologistes (Mills, Abdulla, & Cribbie, 2010, p. 56). On est en droit de questionner l'impact réel des publications méthodologiques, pour 2 raisons, d'après Mills, Abdulla, & Cribbie (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).

Par contre, les articles méthodos peuvent servir à la créations d'outils/logiciels qui seront très utiles aux chercheurs. *Retravailler cette partie sur l'importance des simulations et des logiciels mordernes 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.

(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))

[Anecdote, pour quand je parlerai des logiciels et de leur intérêt: 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 prémacher 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: càd comprendre la distribution d'échantillonnage, dont pratiquement tt découle.]

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.

D'après Thompson (1999a, cité par Fraas & Newman (2000)), les chercheurs continuent à utiliser la nil nul hypothesis pour 2 raisons: (1) la plupart des logiciels partent du postulat que c'est l'hypothèse nulle qu'utilisent les chercheurs et ne donnent pas la possibilité de faire autre chose (2) les non nil-nul hypotheses incluent un niveau de complexité pas toujours possible à atteindre dans bcp de designs. Fraas et Newman (2001) admettent que les chercheurs sont probablement plus enclins à utiliser des procédures si elles sont implémentées dans des logiciels "user friendly".

–> Concernant la raison (1), ce n'est plus tellement vrai en 2021. Jamovi, par exemple, contient un package "TOSTER" qui permet de faire des tests d'effets minimaux ET des tests d'équivalence. Il est très important que des logiciels le fassent, car comme disaient Fraas

& Newman (2000), “unless researchers are able to test non-nil null hypotheses with readily available computer software, they may continue to exclusively use nil null hypothesis” (p.4).

Au final, nos articles ont souvent été utiles pour ce genre de choses:

— Pour le Welch, vérifier mais je crois que Daniel s’en était aussi servi pour le package TOSTER (ou en tout cas en parlait). Nous avons créés des outils pour aider les autres chercheurs (packages R et Shiny App). De plus, nos articles ont été utilisés par d’autres chercheurs qui ont eux-mêmes créés des outils utiles. Par exemple, notre dernier article sur l’ES a inspiré plusieurs chercheurs :

— Aardon Caldwell s’est appuyé sur nos équations pour implémenter le g* de Hedges dans sa mise à jour du Package TOSTER (sur Jamovi)

— Intrajeel Patil (twitter) et son collaborateur se sont servis de notre preprint pour corriger une erreur dans un package R qu’ils ont créés Ça met en lumière ce à quoi à mon avis les articles méthodos doivent servir : à inspirer la création de packages/fonctions dans des logiciels user friendly et communément utilisés pour que les chercheurs sachent concrètement comment agir ! (citer la source qui dit que bien souvent, les chercheurs ne vont utiliser des techniques que si elles sont implémentées dans ce type de logiciel) Pour continuer à mettre à jour les logiciels, il est important de continuer à éprouver les méthodes existantes, et à les comparer aux nouvelles méthodes existantes. Comme la SGPV proposé par Blume était une nouvelle statistique qui semblait remplir les mêmes objectifs que le TOST, il était tout naturel de comparer ces deux techniques de la manière la plus détaillée possible. Peut-être que de nouvelles techniques avec un meilleur ratio coût/bénéfice seront mises en lumière et viendront dès lors naturellement remplacer les techniques existantes (c’est la magie de la science : un questionnement critique et des mises à jours régulières). Par exemple, on a pu montrer effectivement que conformément à ce que certains auteurs suggéraient mais sans que ça soit entendu dans le mnode de la psychologie (car pas appliqué à une compréhension claire du fait que l’hétéroscédatisticité était presque inhérente au domaine, par exemple), il vaut mieux utiliser le test de Welch tt le tps.

Je peux aussi signaler que confirmer une hypothèse nulle, ce qui est important car il y a pas mal de domaines dans la psycho ou le but est de montrer les similitudes, surtout quand c’est

mêlé à des questions sociétales (est-ce que les couples homosexuels sont aussi performants que les couples hétérosexuels ?)

Limites

En recherche, on apprend constamment de nos erreurs. Pour chaque article, j'ai pu identifier des éléments que je ne reproduirais plus à l'identique avec dû recul. Par exemple, dans l'article sur le test de Welch (le premier) j'aurais dû fragmenter. Commencer par faire des simultions avec des distributions identiques dans tous les groupes (car ça permet de moins bien visualiser l'effet de compensation, ex . quand asymétrie positive et négative en mm tps). J'ai également identifié des erreurs dans certains articles

Concernant le premier article sur le test t de Welch (vérifier mais je crois que ça s'applique aussi au 2ème article en fait ces erreurs) Nous spécifions à plusieurs reprises que le test t de Yuen contrôle moins bien le taux d'erreur de type I que le test t de Welch:
- p.14: *"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"*;
- p.15: *"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"*.

Ceci n'est pas exact d'un point de vue purement statistique. A 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 faibles 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 population étaient identiques. Lorsque la distribution d'une population est parfaitement symétrique, la moyenne et la moyenne trimmée seront identiques. Au contraire, lorsque la distribution d'une population est asymétrique, la moyenne et la moyenne trimmée diffèreront (la moyenne trimmée sera plus proche du mode de la distribution et donc, représentera mieux cette dernière). Notons malgré tout que d'un point de vue méthodologique, nous avons déjà relevé que la plupart du temps, les chercheurs définissent l'absence de différence

entre les moyennes comme hypothèse nulle et nos simulations démontrent que dans ce contexte, le test de Yuen n'est pas approprié. En conclusion, le test de Yuen ne devrait être utilisé que par des chercheurs ayant pleinement conscience du fait que les tests t de Student et de Welch ne reposent pas sur la même hypothèse que le test t de Yuen.

Revoir si je parle du fait que le kurtosis impacte la puissance du test de Welch dans l'article en tant que tel (je le fais en tout cas dans les annexes). Expliquer que j'avais fait une erreur en confondant kurtosis et sd (j'ai cru à tort que la mesure de dispersion de la double expo était le sd alors qu'en fait non). Ca m'a fait prendre conscience qu'il est SUPER important de toujours demander, dans les simulations, le calcul des descriptives, afin de vérifier que tout s'est bien passé (si la variance moyenne n'est pas égale à la variable théorique, par exemple, en tt cas qd la condition de normalité est ok, c'est qu'il y a eu un couac). En plus de permettre un contrôle des erreurs, ça peut être utile comme aide à l'interprétation. A partir de l'article 2, je l'ai systématiquement fait. Et pour me "rattraper", j'ai expliqué en détail la différence entre le kurtosis et le SD dans le 2ème article.

Commentaires divers

Dans les deux articles sur le Welch: (même si les pages relevées concerne le test t, c vrai aussi pour le suivant): - p.9: nous décrivons 3 arguments en défaveur de l'usage du test de Levene. En troisième argument, nous mentionnons le manque de puissance du test de Levenne. Nous ne mentionnons cependant pas le fait qu'utiliser le test t de Student lorsque le test de Levene est non significatif revient à confondre le non rejet de l'hypothèse d'égalité des variances avec l'acceptation de l'hypothèse d'égalité des variances. Au sein du chapitre 5 sur les tests d'équivalence, il est démontré par simulation que même lorsqu'on s'assure d'avoir une puissance suffisante pour détecter une différence attendue, la stratégie qui consiste à interpréter le non rejet de l'hypothèse nulle comme un soutien en faveur de l'hypothèse nulle n'est pas appropriée. - p.12: nous mentionnons ceci : *"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).* Avec du recul, cette phrase peut porter à confusion. Par "variances" il faut comprendre "*sample* variances" ou "variances* estimates*". Nous ne sommes donc* pas* en train de dire que les deux statistiques, ainsi que les degrés de liberté et p-valeurs

qui leur sont associées seront identiques lorsque la condition d'homogénéité des variances sera respectée au niveau de la population, mais bien lorsque les estimations de chaque variance de population seront identiques.

Limites dans le chapitre 4, on compare essentiellement les estimateurs sur base de leurs propriétés inférentielles. Nous avons tenté de prendre la dimension interprétative en compte, mais c'est parfois très compliqué. Cette dimension est d'ailleurs rarement prises en compte par les chercheurs. On constate que même si les mesures de taille d'effet sont de plus en plus fréquemment reportées, elles ne sont que rarement interprétées et incluses dans les discussions (Funder & Ozer, 2019 ; Thompson & Snyder, 1997) par les chercheurs. Dans un tel contexte, il est particulièrement important d'ouvrir les débats sur cette question.

Perspectives futures

— Je ne travaille pas avec des distributions discrètes.
— On n'a pe pas assez exploré le bootstrapping (on s'est contenté de croire l'avis de machin qui dit que ça marche pas bien) mais on pourrait requestioner cela et dans les recherches futures le ré-investiguer la comparaison du test de welch classique avec sa version boostrappé.

Take home message : Ce qu'il faut retenir de ma thèse : les recommandations en qlq points. Parler ici du fait que j'ai créé des packages et les indiquer : donner le lien github vers ces packages (même si déjà donné ailleurs).

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.

Algina, J., Keselman, H. J., & Penfield, R. D. (2006). Confidence intervals for an effect size when variances are not equal. *Journal of Modern Applied Statistical Methods*, *5*(1), 2.

Altman, D. G. (2005). Why we need confidence intervals. *World journal of surgery*, *29*(5), 554-556.

Andersen, M. B., McCullagh, P., & Wilson, G. J. (2007). But what do the numbers really tell us?: Arbitrary metrics and effect size reporting in sport psychology research. *Journal of sport and exercise psychology*, *29*(5), 664-672.

Anderson, S. F., & Maxwell, S. E. (2016). There's more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods*, *21*(1), 1.

Association, A. P. (2010). *Publication manual of the American Psychological Association [APA] (6 Ed.)* (American Psychological Association). Washington, DC : (s.n.).

Bakker, M., Van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*(6), 543-554.

Balluerka, N., Gómez, J., & Hidalgo, D. (2005). The controversy over null hypothesis significance testing revisited. *Methodology*, *1*(2), 55-70.

Blume, J. D., D'Agostino McGowan, L., Dupont, W. D., & Greevy Jr, R. A. (2018). Second-generation p-values: Improved rigor, reproducibility, & transparency in statistical analyses. *PLoS One*, *13*(3), e0188299.

Burriss, R. P., Troscianko, J., Lovell, P. G., Fulford, A. J., Stevens, M., Quigley, R., ... Rowland, H. M. (2015). Changes in women's facial skin color over the ovulatory cycle are not detectable by the human visual system. *PLoS One*, *10*(7), e0130093.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature reviews neuroscience*, *14*(5), 365-376.

Button, K. S., Kounali, D., Thomas, L., Wiles, N. J., Peters, T. J., Welton, N. J., ... Lewis, G. (2015). Minimal clinically important difference on the Beck Depression Inventory-II according to the patient's perspective. *Psychological medicine*, *45*(15), 3269-3279.

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.

Coe, R. (2002). It's the effect size, stupid: What effect size is and why it is important.

Cohen, J. (1965). Some statistical issues in psychological research. *Handbook of clinical psychology*, 95-121.

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.

Cumming, G. (2013). Cohen's d needs to be readily interpretable: Comment on Shieh (2013). *Behavior Research Methods*, *45*(4), 968-971.

Cumming, G., Fidler, F., Kalinowski, P., & Lai, J. (2012). The statistical recommendations of the American Psychological Association Publication Manual: Effect sizes, confidence intervals, and meta-analysis. *Australian Journal of Psychology*, *64*(3), 138-146.

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).

Delacre, M., Lakens, D., & 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).

Delacre, M., Leys, C., Mora, Y. L., & Lakens, D. (2019). Taking parametric assumptions seriously: Arguments for the use of Welch's F-test instead of the classical F-test in one-way ANOVA. *International Review of Social Psychology*, *32*(1).

Duran, R. P., Eisenhart, M. A., Erickson, F. D., Grant, C. A., Green, J. L., Hedges, L. V., & Schneider, B. L. (2006). Standards for reporting on empirical social science research in AERA publications: American Educational Research Association. *Educational Researcher*, *35*(6), 33-40.

Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. (S.l.) : Cambridge university press.

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.

Finch, S., Cumming, G., & Thomason, N. (2001). Reporting of statistical inference in the Journal of Applied Psychology: Little evidence of reform. *Educational and Psychological Measurement*.

Fraas, J. W., & Newman, I. (2000). Testing for Statistical and Practical Significance: A Suggested Technique Using a Randomization Test.

Funder, D. C., Levine, J. M., Mackie, D. M., Morf, C. C., Sansone, C., Vazire, S., & West, S. G. (2014). Improving the dependability of research in personality and social psychology: Recommendations for research and educational practice. *Personality and Social Psychology Review*, *18*(1), 3-12.

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156-168.

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and individual differences*, *102*, 74-78.

Glass, G. V., McGav, B., & Smith, M. L. (1981). *Meta-analysis in Social Research* (Sage). Beverly Hills, CA : (s.n.).

Glass, Gene 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.

Golinski, C., & Cribbie, R. A. (2009). The expanding role of quantitative methodologists in advancing psychology. *Canadian Psychology/Psychologie canadienne*, *50*(2), 83.

Goulet-Pelletier, J.-C., & Cousineau, D. (2018). A review of effect sizes and their confidence intervals, Part I: The Cohen'sd family. *The Quantitative Methods for Psychology*, *14*(4), 242-265.

Grissom, R. J. (2000). Heterogeneity of variance in clinical data. *Journal of consulting and clinical psychology*, *68*(1), 155.

Grissom, R. J., & Kim, J. J. (2001). Review of assumptions and problems in the appropriate conceptualization of effect size. *Psychological Methods*, *6*(2), 135.

Grissom, R. J., & Kim, J. J. (2005). *Effect sizes for research: A broad practical approach.* (S.l.) : Lawrence Erlbaum Associates Publishers.

Haslam, S. A., & McGarty, C. (2014). *Research methods and statistics in psychology.* (S.l.) : Sage.

Hedges, L. V., & Olkin, I. (1985). *Statistical Methods for Meta-Analysis* (Academic Press). Cambridge, Massachusetts : (s.n.).

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:

Huynh, C.-L. (1989). A Unified Approach to the Estimation of Effect Size in Meta-Analysis.

Judd, C. M., McClelland, G. H., & Ryan, C. S. (2011). *Data analysis: A model comparison approach*. (S.l.) : Routledge.

Kelley, K. (2005). The effects of nonnormal distributions on confidence intervals around the standardized mean difference: Bootstrap and parametric confidence intervals. *Educational and Psychological Measurement*, *65*(1), 51-69.

Keselman, H. J., Algina, J., Lix, L. M., Wilcox, R. R., & Deering, K. N. (2008). A generally robust approach for testing hypotheses and setting confidence intervals for effect sizes. *Psychological Methods*, *13*(2), 110.

Keselman, Harvey 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.

Kulinskaya, E., & Staudte, R. G. (2007). Confidence intervals for the standardized effect arising in the comparison of two normal populations. *Statistics in medicine*, *26*(14), 2853-2871.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in psychology*, *4*, 863.

Lakens, D. (2016, 9 décembre). The 20% Statistician: TOST equivalence testing R package (TOSTER) and spreadsheet. The 20% Statistician. Repéré à http://daniellakens.blogspot.com/2016/12/tost-equivalence-testing-r-package.html

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social psychological and personality science*, *8*(4), 355-362.

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259-269.

Meyners, M. (2012). Equivalence tests–A review. *Food quality and preference*, *26*(2), 231-245.

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?

Newman, I., Fraas, J. W., & Herbert, A. (2001). Testing Non-Nil Null Hypotheses with t Tests of Group Means: A Monte Carlo Study.

Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological methods*, *5*(2), 241.

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.

Pek, J., & Flora, D. B. (2018). Reporting effect sizes in original psychological research: A discussion and tutorial. *Psychological methods*, *23*(2), 208.

Peng, C.-Y. J., & Chen, L.-T. (2014). Beyond Cohen's d: Alternative effect size measures for between-subject designs. *The Journal of Experimental Education*, *82*(1), 22-50.

Peng, C.-Y. J., Chen, L.-T., Chiang, H.-M., & Chiang, Y.-C. (2013). The impact of APA and AERA guidelines on effect size reporting. *Educational Psychology Review*, *25*(2), 157-209.

Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological bulletin*, *112*(1), 160.

Quertemont, E. (2011). How to statistically show the absence of an effect. *Psychologica Belgica*, *51*(2), 109-127.

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.

Raviv, E. (2014, 2 juin). Bias vs. Consistency. Repéré à https://eranraviv.com/bias-vs-consistency/

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological bulletin*, *113*(3), 553.

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.

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of pharmacokinetics and biopharmaceutics*, *15*(6), 657-680.

Seaman, M. A., & Serlin, R. C. (1998). Equivalence confidence intervals for two-group comparisons of means. *Psychological methods*, *3*(4), 403.

Shieh, G. (2013). Confidence intervals and sample size calculations for the standardized mean difference effect size between two normal populations under heteroscedasticity. *Behavior research methods*, *45*(4), 955-967.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: a key to the file-drawer. *Journal of experimental psychology: General*, *143*(2), 534.

Stout, D. E., & Ruble, T. L. (1995). Assessing the practical significance of empirical results in accounting education research: The use of effect size information. *Journal of Accounting Education*, *13*(3), 281-298.

Sullivan, G. M., & Feinn, R. (2012). Using effect size—or why the P value is not enough. *Journal of graduate medical education*, *4*(3), 279-282.

Thompson, B., & Snyder, P. A. (1997). Statistical significance testing practices in the Journal of Experimental Education. *The Journal of Experimental Education*, *66*(1), 75-83.

Wackerly, D. D., Mendenhall, W., & Scheaffer, R. L. (2008). *Mathematical Statistics with Applications (7th Edition)*. Belmont, USA : Brooks/Cole, Cengage Learning.

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, *29*(3), 350-362.

Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist*, *53*(3), 300.

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American psychologist*, *54*(8), 594.

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.

Annexes

Annexe A: erratum de l'article "Why psychologists Should by Default Use Welch's *t*-test Instead of Student's *t*-test" (Chapitre 2)

Mise en forme et Notations

Les lettres utilisées pour décrire les statistiques (test-t ou test-F) doivent toujours être inscrites en *italique*. Or, cela a été omis à plusieurs reprises dans l'article. Par exemple, il aurait fallu écrire:
- p.9: "... as the Mann-Whitney U-test..." au lieu de "... as the Mann-Whitney U-test...";
- p.9: "F-ratio test" au lieu de "F-ratio test".

Certaines notations mathématiques auraient également dû être indiquées en italique. Par exemple, à la p.9, il aurait fallu écrire:
- "x_{ij}" au lieu de "$\mathrm{x_{ij}}$";
- $|x_{ij} - \hat{\theta}_j|$ au lieu de $|\mathrm{x_{ij}} - \hat{\theta}_\mathrm{j}|$.

Par ailleurs, il est très important d'être consistant dans le choix des notations mathématiques, afin d'éviter d'embrouiller le lecteur. Or, nous n'avons pas toujours respecté cela. Par exemple, nous avons utilisé plusieurs notations différentes pour décrire l'écart-type et la variance. Par exemple:
- p.9: nous utilisons respectivement SD1 et SD2 pour décrire l'écart-type de chaque groupe;
- p.11 (équation 1): nous utilisons respectivement S_1^2 et S_2^2 pour décrire la variance de chaque groupe;
- p.12 (équation 4): nous utilisons respectivement s_1^2 et s_2^2 (lettres minuscules) pour décrire la variance de chaque groupe.

C'est d'autant plus problématique qu'il y a parfois même des inconsistances entre les notations utilisées dans les formules et celles utilisées dans les légendes des formules. Par exemple, nous spécifions p.11 que dans l'équation 1, s_1^2 et s_2^2 (lettres minuscules) représentent les estimations de variance de chaque groupe indépendant, alors qu'en réalité, les estimations des variances sont représentées par S_1^2 et S_2^2 (lettres majuscules) dans l'équation 1.

Faute(s) de frappe

— p.13: “see ~~v~~ **Figure** 2a”.

Annexe B: erratum de l'article "Taking parametric assumptions very seriously: Arguments for the Use of Welch's *F*-test instead of the Classical *F*-test in One-Way ANOVA" (Chapitre 3)

Mise en forme et Notations

Une légende est manquante pour certaines notations mathématiques. Par exemple, en ce qui concerne l'équation (1), bien que n_j, k et s_j^2 aient été correctement définis, les définitions pour $\bar{x_j}$, $\bar{x_{..}}$ et N ne sont données que plus tard, en référence à d'autres équations. Cela peut rendre la lecture de l'article plus compliquée pour certaines personnes non familières avec ces notations.

Par ailleurs, comme dans l'article précédent sur le test t de Welch, on constate certaines incohérences en termes de notation. Par exemple, si la moyenne de chaque groupe est définie par $\bar{x_j}$ dans l'équation (1), elle est définie par $\bar{X}_j$ dans l'équation (7).

Enfin, 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 apparaisse 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 récemment.

Faute(s) de frappe

— p.18: "Although it is important to make sure ~~test~~ **that** assumptions are met";

— p.19: "... we think that a ~~first~~ realistic first step towards progress would be to get researchers...";

— p.20: "Based on mathematical explanations and Monte~~o~~ Carlo simulations";

— p.21: "~~With~~ **w**ith $N = \ldots$";

— p.21: "~~Where~~ **w**here ~~$\bar{x_j}$~~ $\bar{x_j}$ and s_j^2 are respectively the group mean and the group variance. . . ";

— p.22: ". . . negative pairings (the group with the ~~smallest~~ **largest** sample size is extracted from the population with the smallest SD);

— p.22: "the type I error rate of all test**s**";

— p.24: ". . . which is either more liberal or more conservative, depending on the SDs and ~~SD~~ **sample sizes** pairing";

Annexe C: échanges avec Geoff Cumming, en vue d'améliorer l'article non publié "Why Hedges' g_s^* based on the non-pooled standard deviation should be reported with Welch's t-test"

Le 4 juin 2021, suite à la soumission d'un preprint de notre article sur les tailles d'effet, nous avons eu le plaisir de recevoir un email de la part de Geoff Cumming. Cet email est présenté à la page suivante.

En pièce jointe de cet email, figurait un feedback long et détaillé de notre article. Celui-ci est entièrement retranscris juste après le mail. Le texte en bleu qui est inséré dans ce feedback correspond aux réponses que nous lui avons fournies, et le texte en brun correspond aux réactions de Cumming à nos réponses.

Pour finir, cet échange a donné lieu à un blog post disponible à l'adresse suivante: https://thenewstatistics.com/itns/2021/06/17/which-standardised-effect-size-measure-is-best-when-variances-are-unequal/.

Gmail **Marie Delacre <mdelacre1@gmail.com>**

your fine preprint

Geoff Cumming <g.cumming@latrobe.edu.au> 4 juin 2021 à 06:29
À : "marie.delacre@ulb.ac.be" <marie.delacre@ulb.ac.be>
Cc : "d.lakens@tue.nl" <d.lakens@tue.nl>, "Calin-Jageman, Robert" <rcalinjageman@dom.edu>

Dear Dr Delacre and colleagues,

I was delighted to see your fine preprint "Why Hedges' g_s* based on the non-pooled standard deviation should be reported with Welch's *t*-test". It strikes me as important and compelling, as well, I'm sure, as requiring a vast amount of work.

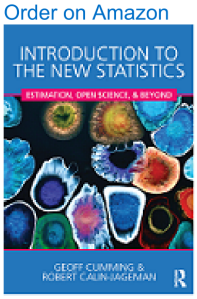
I plan to blog about it, but first I'd like, if I may, to ask some questions and offer comments, as attached. I'd greatly appreciate any comments you might care to make in reply.

Of course, I'm more than happy to discuss any of these issues--it's all fascinating as well as essential stuff!

With thanks, and best regards,

Geoff

Geoff Cumming, DPhil, Emeritus Professor,
School of Psychology and Public Health, La Trobe University, Melbourne Campus, Victoria, Australia 3086
Email: g.cumming@latrobe.edu.au
Intro textbook: Introduction to The New Statistics: Estimation, Open Science, and Beyond
www.thenewstatistics.com
First book: Understanding The New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis
ww.thenewstatistics.com
Own page: http://www.latrobe.edu.au/she/contact-us/staff/profile?uname=GDCumming
ESCI (Exploratory Software for Confidence Intervals): www.thenewstatistics.com
Introduction to the New Statistics is the first statistics textbook to focus on Open Science and the New Statistics.
Instructors can obtain a free desk copy at https://www.routledge.com/resources/deskcopy.
Order on Amazon



Delacre comments 4 Jun 21.docx
65K

Why Hedges' g_s^* based on the non-pooled standard deviation should be reported with Welch's t-test

https://psyarxiv.com/tu6mp/

Comments by Geoff Cumming

g.cumming@latrobe.edu.au

4 June 2021

Subscript s

I'm wondering why you use subscript s for all eight ES measures. Yes, Cohen's original term for the estimate was d_s, but d became the standard usage. Perhaps you use the subscript s to indicate that the standardiser is an SD estimated from data? By contrast d_δ, for example, would indicate that a population value is available to use as standardiser. But in your paper there are no such cases, so you could simplify everything by simply omitting all those subscript s's?

Also, that role for the subscript rules out use of it to distinguish, for example, between d_p for pooled s, and d_C for Control group SD as standardiser. However, I know there are no well-established conventions for such subscripts, and you need to use 'Shieh' as a label for those ES measures, so perhaps using name labels (Cohen's, etc) for all measures—as you do—and dropping all the s subscripts would be simplest.

I used the subscripts to indicate that the ES measure is estimated based on a sample, but I agree that it is maybe not necessary, so I will remove all subscripts and use labels instead.

Typos?

Thanks for pointing all these typos out!

p. 7, l. 139: Hedges' d_s^* should be Cohen's d_s^*? That's right, it should be Cohen's d^*

p. 8, l. 152: Use σ rather than σ_{pooled} (twice) because here we're assuming homogeneity of

variance, with $\sigma = \sigma_1 = \sigma_2$ as the common population SD? (Also Table 1, line 1 of Note.) Ok!

p. 9, l. 3 of footnote: Not 52 but .52? (By my calcs using $df = 3$, $N = 5$, and the approximate debiasing formula the value should be .529?) Yes, it's a typo, the right number is .52 instead of 52 (see the R console below).

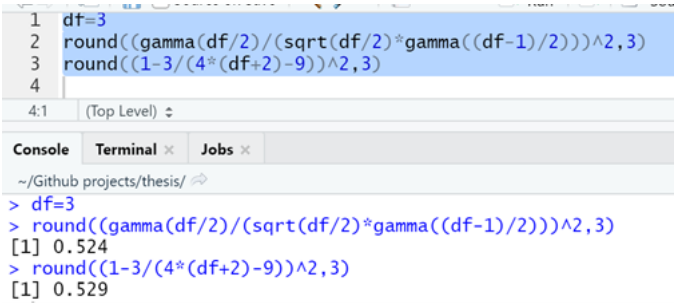
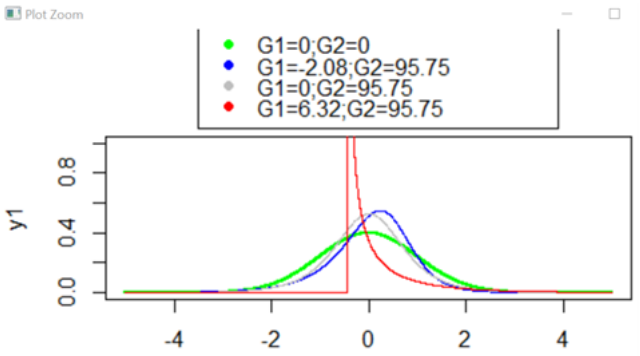


Table 3: Cohen's g_s^* should be Hedges' g_s^*? Also on p. 20, l. 275 (without *), and p. 30, l. 485. Also true.

Non-normality

It's great that you explore non-normality, so that you are investigating the robustness of the various ES estimators. I'd love to see a figure showing the three example distributions you choose to study, alongside a normal curve. Then I could have some intuitions about how extreme they are, especially in comparison with distributions that I might see in data I'm analysing. Thanks.

In the plot below, here is what the four distributions look like when $\mu = 0$ and $\sigma = 1$.



It is a nice suggestion to add a Figure in the manuscript to show the compared distributions, I will take it into account.

Additional comment: in our simulations, we used only continuous distributions. I think it would be interesting to test the properties of estimators when distributions are discrete (e.g. Likert

scales) but I have no idea how to generate such distributions in a realistic way. If you have any suggestion, I would be more than happy to think about it next time!

Yes, it would be valuable to have investigations with discrete distributions, but I can't think of good starting points. We'd hope for better than the normal approximation to the binomial. Maybe start with the investigations that have been made of that, especially with small N (Likert, as you mention).

Bias

I take your point (p. 23) about the value of focusing on relative, rather than absolute, bias (and variance) when there is a change in measurement units. But that doesn't happen in your investigations.

Well, I wrote about a "change of unit" because we use a different standardizer for each estimator. As a consequence, each estimator computation results in a different value in order to describe the same amount of difference. I will clarify in the text.

The one other reason you state for considering relative bias is a reference, on p. 5, bottom, that Tables 1 and 2 illustrate that "bias is directly related to the population effect size". That's true for Cohen's d in Table 1, and for d and d^* in the top two rows of Table 2. (I'm omitting Shieh from my comments—see below for reasons.) The debiasing that gives Hedges' g, etc, is proportional, and results in zero bias for any population effect size. I think, therefore, that neither of the reasons you mention for using relative bias applies for any of the g family measures in Figures 2-9 (omitting Shieh, as I will do unless specifically mentioned), and so I'm wondering whether seeing means (i. e., averaged over δ values) of relative or absolute bias would be more informative.

I agree that under the normality assumption, there is no bias for any unbiased estimator, of course. However, when the normality assumption is not met, biases occur. In Supplemental Material 1, we mention that when samples are extracted from heavy-tailed symmetric distributions, this bias depends on the same parameters as the bias of biased estimators when the normality assumption is met. It's true that we mainly based our decision to compute the relative bias on the comparison between Cohen's d (Hedges' g) and Shieh's d (Shieh's g). For

example, when designs are balanced and population variances are equal across groups, the bias of Shieh's g is always approximately twice smaller as the bias of Hedges' g, and its variance is always approximately four times smaller as the bias of Hedges' g. It gives the illusion that under this specific configuration, Shieh's g is less biased and variable as Hedges' g, but it is only due to what we called "a change of unit".

Note: I explain below why I think it's important to consider Shieh's d in our comparisons (even if I agree with the fact that interpreting this measure is very hard).

I guess I'd like to see the bias for each δ value. (On p. 24, line 2, you give a link to an illustration of raw bias and variance, but I can't find any tables or graphs of values at that site.)

The graphs for raw estimators of goodness are available on my Github account: https://github.com/mdelacre/Effect-sizes/.
You will find them in the following folder: "Scripts outputs/Quality of ES measures/Graphs/Unbiased estimators/Raw estimators of goodness/". The fact that you haven't found it probably means that I should reorganize the folder, in order to make it clearer, to indicate more clearly the Figures position in the folder or to add a read me file!

Thanks for the link for raw results. https://github.com/mdelacre/Effect-sizes/ That's different from the link in the preprint. [En effet, nous avions collé un mauvais lien par erreur dans le preprint].

You say that the main purpose of the figures is to allow comparisons between the different ES measures, rather than absolute values for bias (and variance). Even so, knowing when bias is likely to be absolutely small or large can inform judgments about the different ES measures in various contexts. In Figures 2 and 3, we have $\sigma_1 = \sigma_2$ and four versions of g, all debiased. If I'm following, the relative bias values pictured in the top rows of both figures are averaged over $\delta = 1$, 2, 3, and 4. (I'm assuming $\sigma_1 = \sigma_2 = 1$; correct?) Right! For the non-normal distributions, average relative bias in the best cases is around 1-2%, so not bad. But up to around 5-10% in some other cases, a bit more concerning.

Absolutely right. However, as mentioned in the manuscript, "we chose very extreme conditions, and we know that none of the parametric measures of ES will be robust against such extreme

conditions". It is therefore not surprising that *all* estimators are very large under some conditions. We could insist more on the fact that the parametric estimators are robust only under moderate deviations from the normality assumptions.

It's interesting to see that almost all the relative bias values plotted in Figures 2-9 are positive, so in nearly all cases the ES estimates are on average too high? If so, this may be worth noting? (Apologies if I missed your comment on this.)

You're perfectly right. There is a mathematical reason for that (at least for symmetric distributions; it is explained in Supplemental Material 1, in the "Preliminary note"; see below). We did not include this note in the main manuscript because we were afraid that it would be too technical for many psychologists, and we did not want to complicate even more a paper that already contains many complex concepts, but maybe we could mention this in a footnote if you think it's missing.

Preliminary note

For all previously mentioned estimators (Cohen's d_s, Glass's d_s, Cohen's d_s^* and Shieh's d_s), the theoretical expectation is computed by multiplying the population effect size (respectively δ_{Cohen}, δ_{Glass}, δ^*_{Cohen} and δ_{Shieh}) by the following multiplier coefficient:

$$\gamma = \frac{\sqrt{\frac{df}{2}} \times \Gamma\frac{df-1}{2}}{\Gamma\frac{df}{2}} \quad (1)$$

where df are the degrees of freedom (see the main article). γ is *always* positive, meaning that when the population effect size is not zero, all estimators will overestimate the population effect size. Moreover, its limit tends to 1 when the degrees of freedom (df) tend to infinity, meaning that the larger the degrees of freedom, the lower the bias.

While we focus on the theoretical bias of biased estimators when the normality assumption is met, it is interesting to notice that our main conclusions seem to generalize to :

- biased estimators when samples are extracted from symmetric distributions;
- unbiased estimators when samples are extracted from heavy-tailed symmetric distributions.

Variance

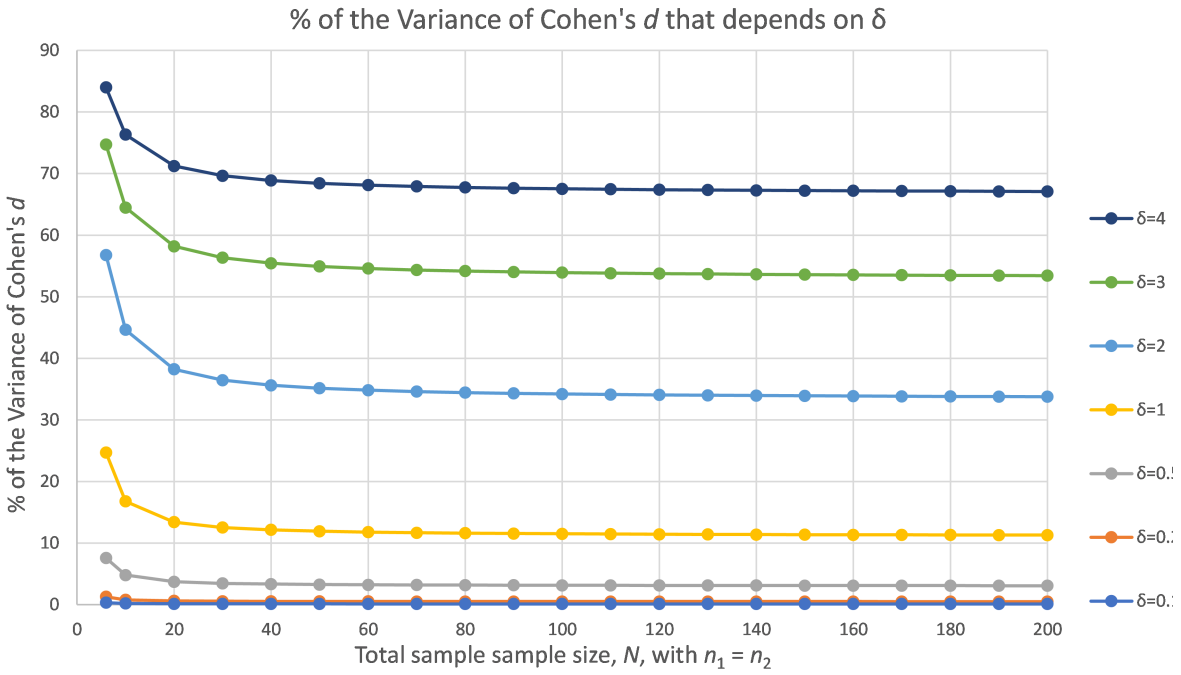
For bias, the goal is zero, an unbiased estimate. For relative variance, the smaller the better, but we don't expect zero because we'll always have sampling variability in estimates of δ. I'm finding it hard to think of a good way to get intuitions about the empirical variances of the different ES measures, starting with the relative values reported in Figures 2 and 3. Part of my difficulty is that variance is a squared measure. As usual, I'd much prefer to be able to think in terms of CIs and their lengths.

Thinking in terms of CIs and their lengths require to run new simulations because we compute CIs based on the noncentral t distribution method (the method is explained in CI.pdf, here: https://github.com/mdelacre/Effect-sizes/tree/master/Supplemental%20Material%204/CI.pdf/ ; the R script is available here: https://github.com/mdelacre/Effect-sizes/blob/master/Scripts/Confidence%20intervals/CI.R.

I have made simulations in order to compute CIs around point estimators for all scenarios where $\mu_1 - \mu_2 = 0$ and $\mu_1 - \mu_2 = 1$.

Note that analyses on CIs would probably be very redundant with that of the variance. I agree that the length of CIs is probably more intuitive, but at the same time, everybody knows that the variance is a measure of dispersion, and the greater the variance, the greater the dispersion. We mention that our main goal is to *compare* the relative bias and variance of different estimators, and variance allows this comparison.

The variance formulas in Tables 1 and 2 show that the variance of d and g, and the top two ES measures in Table 2, do indeed depend on δ. To try to get a feel for the extent of this dependence I made a quick spreadsheet of the variance formula for Cohen's d, using the approximate formula in the second line of Table 1. The figure below shows the percentage of variance in d that is given by the second, δ-dependent, term in the formula for variance. I assumed two equal sized groups. The figure suggests that for $\delta < 1$ only a very small % of the variance is dependent on δ, and for $\delta = 1$ (around 10%) and greater, the % increases rapidly with δ—not surprisingly, given the δ^2 term in the formula. The % hardly changes with sample size, for two samples each of at least 20, as you use.



For δ values of 1, 2, 3, and 4 (as are averaged in Figures 2, 3), the %ages are, in round numbers, about 11, 34, 54, and 67, respectively. I expect the %ages for the other ES measures would not be vastly different. Therefore the relative variance values shown in the lower row in Figures 2 and 3 are averages over cases in which dramatically different proportions of the variance are dependent on δ. I think this makes it harder to get good intuitions. Values of the variance of the estimates, rather than relative values, may be more understandable?

This is a very interesting analysis. I was quite confused about this question, and I've been wondering for a long time what the best solution would be. We had many long discussions in order to decide whether we should compute the variance or the relative variance. The issue was indeed the fact that the variance only partly depends on delta. Finally, we decided to compute the relative variance because we had in mind that when $n_1 = n_2$ and $\sigma_1 = \sigma_2$, the bias (variance) of Shieh's d is twice smaller (four times smaller) as the bias for Cohen's d, only because Shieh's d is twice smaller as Cohen's d (see Appendix 1 in our paper). However, I keep in mind that you argue in favor of removing Shieh's d from our analyses.

Taking a different approach, and considering the relative variance values plotted in Figures 2 and 3, would it be reasonable to say that relative SD, the square root of relative variance, is the SE used to calculate the CI on that ES estimate, expressed as a fraction of δ, the population ES? If so, such a 95% CI would be about $\pm 2 \times SE\ value \times \delta$. (Given that the smallest total N being

considered is 40, that approximation should be reasonable? Perhaps less so for much smaller group sizes. The asymmetry of CIs on d is actually very small unless N is very small and δ is large.) If that's at least roughly correct, then your reported values of average efficiency become indicators of the precision of the different ES estimates. An efficiency estimate (lower row in the figures) of, say, 0.04 (a typical low value) would correspond approximately to $SE = 0.2\delta$ and a 95% CI of about $\pm 0.4\delta$. However, that's very rough because the efficiency values plotted are means of four possibly quite different values. (The dependence of relative variance—and thus CI length—on N is clear in Figure 2, and can also be seen in Figure 3.)

This is not very clear to me. If you think this is a key point, perhaps we could discuss it together when I will revise the article?

It looks complicated here but, I think, mainly because I'm trying to get approx CI info from averaged relative variance values. It's probably not worth considering this further, given that somewhere in future plans may be analyses that tell us directly about CI lengths and coverage %ages, and likely errors (biases) in these. I'm more than happy to discuss at revision time.

Shieh's *d* and *g*

I can see that including these does give a comprehensive exploration of ES possibilities. Seeing them here led me to go back to Shieh's original article and my comment. You cite that comment (thank you) as reason to doubt the value of the Shieh approach. You also point out (p. 15, bottom) that in the base case the Shieh value is half d. This seems to me so crazy that it alone may provide sufficient reason to dump the Shieh approach. For example, a difference of 7.5 between two IQ means is $d = 0.5$, half an SD. That makes sense. But Shieh would calculate 0.25! Units of double the SD! How could we make sense of that? How could I even attempt to justify that to students as a meaningful standardised ES measure?

But as you say there's more. Calculating a pooled SD weights by sample size, for good reasons. The Shieh calculation, however counter-weights, so the SD of the *smaller* group is weighted more heavily. That's what we need to calculate the variance, but, surely, not a value that could justifiably be used as standardiser—the units in which to express our standardised ES.

Even more, again as you say: Simply change the n-ratio and the SD estimate changes, so we can't see the ES value we get as an estimate of any readily understood population quantity.

All that seems to me more than enough reason simply to dump the Shieh approach. Imho, it's worth no more than a para or two to explain why it's not worth pursuing.

I get your point and I agree with all what you say. At the same time, Shieh's d is based on the same statistical quantity as Welch's t-test. By "same quantity", I mean that the percentage of p-values associated with Welch's t-test below the alpha risk exactly equals the proportion of confidence intervals around Shieh's d that does not include 0. Our first motivation to write this paper was to advice researchers with a measure they could use when performing Welch's t-test, in the continuity of our previously written articles on the subject ("Why Psychologists Should by Default Use Welch's t-test Instead of Student's t-test" and "Taking Parametric Assumptions Seriously: Arguments for the Use of Welch's F-test instead of the Classical F-test in One-Way ANOVA"). In that context, I think it's a bit weird to not include this effect size estimator in our comparisons. Moreover, before running all my simulations, I thought that this test was the best from an inferential point of view (again, due to this mathematical relation). The study of bias and variance reveals that even for inference, it has big flaws. For that alone, I find it interesting to include it in the comparisons. At the same time, we can (and we will) insist a little more on the fact that the mathematical relation between Welch's t-test and Shieh's d explains why we took this ES estimator into account, and specify at the outset that the simulations will reveal that this measure is not very defensible, even purely statistically speaking.

Results section p. 24, ll. 356-8: moving left to right, columns 2-4: in both figures the bias increases slightly then decreases considerably—not "moving left to right… the larger the bias"

Have you noticed that in column 4, we don't use the same scale as in column 2 and 3 (this is true for both Figures 2 and 3)? So indeed the bias is larger in column 3 than in column 2, and larger in column 4 than in column 3.

In some places I did note the differing vertical scales in your figures, esp. Column 4. But I slipped up in my comments on interpretation of those figures. Sorry.

p. 27, l. 403: Unequal variances. In Figure 4, what are the variances? A single pair of values in all cases, or are the pictured values averaged over a range of variance pairs?

Scenarios in Figure 4 are those where σ_1 always equals 1, and σ_2 = .1, .25, .5, 2, 4 or 10, so

pictured values are averaged over a range of variance pairs. I'll mention it explicitly in the next version to avoid confusion.

p. 27, ll. 407-8: "when moving from left to right... the larger the bias". But for all Figures 4-9, except 7, bias increases slightly (or hardly) from column 2 to 3, then drops to be smallest in column 4.

Again, we don't use the same scale in column 4 vs. columns 2 and 3. It was mentioned p.24, l. 339.

p. 27, ll. 408-9: "the bias of Hedges' g_s remains very small..." But it is often large, and in every case no smaller than that of Hedges' g_s^*.

I meant that the bias of Hedges' g remains very small when the normality assumption is met (but this needs to be clarified better in the sentence!).

I'm wondering what value you use for δ when assessing relative bias and variance for Hedges' g. Perhaps what you define for δ_{Cohen} back on p. 8, l. 152? If so, that value depends not only on the two variances, but on the two sample sizes. So we're estimating a population ES that doesn't reflect any relevant population in the world and, moreover, assuming a population ES that changes merely because we happen to use different sample sizes.

What would you recommend to use? Indeed I'm using δ_{Cohen}. Mathematically speaking, the expectation of Hedges' g always equals δ_{Cohen} when both assumptions of normality and equal population variances are met, so we compare mean(Hedges' g) with δ_{Cohen} in order to compute the raw bias. This explains why I also used δ_{Cohen} in the denominator when I computed the relative bias.

I had in mind the introduction of Cohen's d on pp. 7-8. We are assuming homogeneity of variance. I made a comment:
p. 8, l. 152: "Use σ rather than σ_{pooled} (twice) because here we're assuming homogeneity of variance, with $\sigma = \sigma_1 = \sigma_2$ as the common population SD? (Also Table 1, line 1 of Note.)" Ok!

→ you agreed with that. So the formula, line 152, with the square root (I'll call that 'the weird formula') reduces to $\sigma = \sigma$. I guess I'm assuming that Cohen's d, and also Hedges' g, are

defined assuming homogeneity of variance and their formulas use s_p, the pooled value. They both estimate δ, defined as in your (5).

Thinking further about this, I take your point. In a simulation we know σ_1^2 and σ_2^2, as well as the two sample sizes, but we know of no 'true' single σ to use in defining the δ we would like to use as the benchmark to assess the estimates calculated in the simulation. Perhaps using the weird formula is the best we can do; if anything using this means we're bending over backwards to minimise likely estimation error.

Perhaps you could consider shifting the weird formula from the basic intro to Cohen's d on pp. 7-8, where to me it doesn't seem to fit, and introducing it, as a sort of necessary evil, when explaining the simulation evaluation of d and g?

However, on p. 27, from l. 411, you say that pooling variance estimates in this situation results in an [effect size] estimator that "will be smaller… or larger… than it should be." What should it be? Doesn't such an estimator match the definition of δ I refer to above? Then you make the same statement about population values, although that would not be true if you are using the definition of δ I refer to above—which is based on just such a pooled (and weighted by sample size) population SD? Even so, your conclusion at the top of p. 28, dismissing Hedges' g, may be justified—when variances differ.

I'm not sure that I understand your comment. Could you please clarify? P.27, we meant that when variances are unequal across populations, the pooled error term is not valid (because it assumes equal population variances). When we compute the bias of Cohen's d in scenarios with unequal population variances, we compare an invalid estimate with an invalid population measure of effect size. Because both estimate and population value are invalid, the problem of Cohen's d in case of heteroscedasticity is not revealed by the calculation of the bias. This is actually the same explanation as p.10:

172 [redacted] If we pool the estimates of two unequal population variances, the
173 estimator of effect size will be smaller as it should be in case of positive pairing (i.e. the
174 group with the larger sample size is extracted from the population with the larger variance)
175 and larger as it should be in case of negative pairing (i.e. the group with the larger sample
176 size is extracted from the population with the smaller variance). [redacted]

Given your explanation about using the weird formula, I think I now understand what you were

getting at. Your statement that the estimate and the population measure (calculated using the weird formula) are both invalid expresses the problem that there's probably no good way to calculate the population measure when variances are unequal. No sensible way to specify what the estimate should be estimating. Indeed!

Of course, if we try to avoid the double invalidity you point to by simply averaging the two different σ values, we get the d^* and g^* estimators. These may solve the unequal n problem of the weird formula, but still have problems (in my view, severe problems) in terms of interpretability.

p. 28, l. 422: "parameters that we cannot control…". You include the n-ratio, which is true if, for example, we are meta-analysing past research. But in many cases a researcher chooses sample sizes, and your investigations are likely to lead to useful advice about choice of sample sizes.

I think it's partly true. About the bias, the larger the control group size, the lower the bias. About the variance, it's a bit more complex: in Supplemental Material 1 we write that when the normality assumption is met "The variance always decreases when the control and/or the experimental group increases. The benefit of adding subjects in the control, in the experimental, or in both groups, in order to reduce the variance, varies as a function of the SD-ratio and the population effect size. The only situation where it is optimal to maximize the experimental group is when $\sigma_e > \sigma_c$ and $\delta_{Glass} \approx 0$. Most of the time, it is more efficient to maximize the control groups (e.g. anytime $\sigma_e < \sigma_c$, and when δ_{Glass} is very large) or to uniformly add subjects in both groups (e.g. when $\sigma_e > \sigma_c$ and δ_{Glass} is neither null nor huge)" (This is a summary but you can see more details and figures here: https://github.com/mdelacre/Effect-sizes/blob/master/Supplemental$/%$20Material$/%$201/Theoretical$/%$20Variance$/%$20of$/%$20all$/%$20estimators$/%$20as$/%$20a$/%$20function$/%$20of$/%$20population$/%$20parameters.pdf).

As we can see in this summary, recommendations about sample sizes would be different as a function of SD-ratio and most of the time, we don't know the SD-ratio in advance. But thanks for pointing this out, it probably deserves at least a footnote!

Discussion

I think your three purposes of ES estimators (pp. 3-4) are spot on. Your three desirable properties

are indeed the desirable statistical properties, but I'd list interpretability in the research context as the first essential property. If we don't have interpretability for a particular standardised ES estimator then we shouldn't be using it.

Theoretically speaking, I agree with you. But practically, as I've mentioned before, I'm not sure if it's a good idea to rule out an estimator that has such a direct connection to Welch's t-test. It is probably possible to introduce things differently, to explain in advance that the Shieh's d is not desirable (and to add that even from an inferential point of view, it has big flaws).

Apart for that, you're right, interpretability should be our first priority. For example, if we don't have an interpretable measure, we cannot use it in order to make informative null hypothesis (e.g. when performing equivalence test). However, I'm a bit concerned with the fact that interpretability and good inferential properties seem so difficult to reconcile. How could we interpret an estimator correctly if it is very biased? This makes me very reserved about Glass's d.

I note these recommendations:

1. p. 10: "Because the assumption of equal variances... is rarely realistic... both Cohen's d and Hedges' g should be abandoned...". I think that's arguable. I'm not convinced that assumption is rarely realistic. The assumption is very often made, for example, I think, in medicine when calculating SMD. I suspect a large proportion of Cochrane systematic reviews include meta-analyses that make this assumption. Of course that doesn't make it justified in every situation, but I suggest the emphasis should be on informed judgment in context rather than simply abandoning these ES estimates.

Perhaps do we need to be more nuanced in asserting this. However, there are many authors who believe that the assumption of homogeneity of variances often does not hold (see for example Erceg-Hurn & Mirosevich, 2008; Zumbo & Coulombe, 1997). In a previous paper (Delacre et al. 2017) we develop many reasons why we think equal population variances are very rare in practice.

Moreover, it's very hard to check for the homogeneity of variances assumption, because:
* the assumption is about population parameters that we don't know (σ_1 and σ_2);

* inferential statements about the homogeneity of variances assumptions based on assumptions tests often lack power to detect assumption violations.

Finally, when we look at figures, we notice that when variances are equal across groups, Cohen's g and Cohen's g^* are either identical (Figure 2) or very close (Figure 3). The only exception is when both skewness and kurtosis are very large (as reminder, we used different scale in the fourth column in comparison with all other columns). Most of the time, there is therefore little cost in choosing Cohen's d^* by default. On the contrary, Cohen's d cannot be used in case of heterogeneity of variances. – Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use 521 Welch's t-test instead of Student's t-test. International Review of Social Psychology, 522 30 (1), 92–101. https://doi.org/10.5334/irsp.82

– 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. DOI: https://doi. org/10.1037/0003-066X.63.7.591

– 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

2. pp. 17-18: You note the wide criticism of Cohen's d^* (and by implication Hedges' g^*) because the standardiser is not the SD of an existing relevant population. You state that, even so, these estimators have very good inferential properties. Grissom and Kim (2012), referring to this averaging of variances, state that we "are estimating the σ of a hypothetical population... which is a concern for some ... but not others" (p. 72). I tend towards the concerned end of that range, but I know there are arguments in favour of using the square root of the average of two different variances as standardiser. Maybe it can sometimes be possible to give a reasonable interpretation using such a strange standardiser. Think of two overlapping normal curves with different SDs: What SD unit is best for expressing the difference between the two means? Does it make sense to use a compromise between those two SD values? Again, this should be recognised as a judgment call, and lack of interpretability in context should never be over-ruled by good statistical

properties.

I find this convincing (especially the fact that interpretability should never be over-ruled by good statistical properties) and will have to think about a better way to introduce this estimator.

1. p. 28, l. 424: "We do not recommend using [Glass's g]." I suggest that Glass vs something else is the choice that most clearly should be based on the context. Does it make sense to use the SD of one group as the standardiser? If so, we should do so, unless there are very strong reasons against. We should use choice of sample sizes and perhaps other strategies to minimise any disadvantage of the Glass ES estimate. You make several observations that can help, mainly that increasing the control group sample size is desirable.

As previously mentioned, we need to increase the control group in order to reduce the bias, but sometimes, we need to increase the experimental group in order to reduce the variance. This makes recommendations about sample sizes very complicated.

2. p. 30, l. 473: "We recommend using Hedges' g^*..." You state some strong advantages of this ES estimator, but I remain hesitant because of the possible interpretive difficulty, given its standardiser.

Overall, you provide a wonderful range of detailed findings. I'd love to see you develop the discussion, drawing further on those findings, in a way that guides a researcher's likely full decision sequence. For example, first, do we have some relevant population SD value? If so, use this as standardiser. If not, Glass or something else, considering the context? If not Glass, homogeneity of variance or not? Then, can we improve the precision of estimation of our chosen standardiser by, for example, pooling over studies, or meta-analysing our data with past research? Meaningfulness and interpretability in the context should always be the primary consideration.

The great value of your findings is then to help us be specific about the trade-offs. If I choose Glass, how large is the likely bias? If only a few % I might be prepared to tolerate that, or I might consider a data transformation to reduce the departure from normality. (Perhaps a log transform of RT data?) Of course, all such choices should where possible be made in advance of seeing the data, when formulating the data analysis plan to be preregistered.

This is a set of considerations that I will have to take into account when I rework my article in September. If you're ok with that, I'd love to show you our next draft before submitting it.

I'd be more than happy to comment on future drafts, in due course. Or to discuss any of these fascinating issues!

Earlier I made a very sketchy attempt to estimate roughly what CI length might be, given your relative variance results. You refer (e.g., p. 17, l. 253) to "very good inferential properties". That's highly relevant, but I suggest CI length and coverage should be central issues in assessing the inferential effectiveness of any estimator. Then, when considering some judgment call, I could weigh up any price there might be to pay in terms of a longer CI for one of the options under consideration, or a slight departure from 95% coverage. No doubt investigation of CI length and coverage would be a further project, although perhaps you could derive from your variance results some relevant guidance based on CIs?

Ideally, we'd like to see the performance of CIs based on non-central t distributions, and also bootstrapped CIs, for the various estimators and conditions you investigate. Ideally, again, we could then weigh up the various choices, with some idea as to the cost in terms of bias or CI length or coverage error of a choice that offers superior interpretability—which should always be paramount.

As mentioned above, I already computed CIs for all the conditions, based on non-central t distributions. However, I haven't analyzed them yet but I could use them in order to compute the coverage probability and CI length. Therefore these comments could be part of the "further studies" section, and be the subject of a second paper.

Overall, your investigations give us highly valuable findings to guide such discussions.

Title

Current title: "Why Hedges' g_s^* based on the non-pooled standard deviation should be reported with Welch's t-test".

Why mention that t-test? Yes, it fits with your message that the assumption of homogeneity of variance should be avoided, but it's hardly mentioned in the paper, and if an ES estimate and

CI are reported there's no need to report or even mention any t-test. I would argue that it's much better to avoid doing so.

More positively, your investigations warrant broader recognition and description. Perhaps something like: "d-family standardised effect size estimators: Interpretability, bias, efficiency, and robustness" "d-family standardised effect sizes: Bias, efficiency, and robustness" "Cohen's d and related effect size estimators: Interpretability, bias, precision, and robustness" [my current favourite; assumes some discussion of CIs] "Cohen's d, Hedges' g^*, and related effect size estimators: Interpretability, bias, precision, and robustness" [includes the largest set of terms likely to be used in an online search]

We wanted to mention t-test because this answering the question many people asked us: Which effect size should I report for Welch's t-test. Yes, the paper discusses more, adding Interpretability, bias, precision, and robustness is a good addition. What do you think about: "Arguments to use the Hedges'g_s^* with the Welch's t-test: an investigation of interpretability, bias, precision, and robustness of d-family effect sizes"?

As I said in the blog post, [Voir lien vers le blog post] I'm not convinced that Welch's t-test deserves mention, even though you have made a strong case for using that form of the t-test in your earlier paper. If people ask 'which ES?' I'd answer that they need to make a judgment in context, partly guided by your robustness results—but I've said all that in the post.

I know that you make a general recommendation that, when variances are or may be unequal, we should use Hedges' g^*, and, separately, Welch's t-test, but I don't see any other link between the two.

Your Equation (10), p. 15, expresses the very strong link between Welch's t and Shieh's d. Somewhere I've read a statement that the CI on Shieh's d corresponds with Welch's t test in the sense that the two give the same p values, tho' I can't find that statement in Shieh (2013) or in your preprint or supp material—sorry if I've missed it. However, I'd expect that the CI on your recommended Hedges' g^* and Welch's t would not give the same p values. If that's correct, we have one more reason not to link the two together, for example in the title.

Grissom, R. J., & Kim, J. J. (2012). *Effect sizes for research*, 2*nd* ed. New York: Routledge.

It's nice to see a bunch of likes and retweets to the tweet about the blog post. [See https://twitter.com/TheNewStats/status/1405386935074496512]

There's a useful comment (link below) from Brenton Wiernik about relevant work by Doug Bonnett. Very likely you know this, tho' it's new to me. I'd better have a squiz.

Cheers, Geoff

Annexe D: A VOIR SI ERRATUM DU CHP 5?