Remerciements

Il y a tellement de personnes qui méritent leur place dans mes remerciements que j'espère n'oublier personne. Pour commencer, j'ai eu la chance dans mon parcours d'être entourée d'académiques aussi compétents que bienveillants. Merci à mon promoteur, Christophe Leys, de m'avoir toujours challengée. Nos divergences se sont souvent avérées être une force. Là où ma tendance naturelle m'a fréquemment amenée à proposer des brouillons d'articles truffés de détails noyant le discours, mon promoteur m'a fréquemment rappelé l'importance d'aller à l'essentiel et d'épurer mon discours pour rester aussi didactique que possible. Cela a certainement contribué au succès des articles présentés au sein de cette thèse. Dans un registre plus privé, merci pour avoir toujours su me rassurer dans mes (fréquents) moments de doute et d'avoir toujours été là pour moi, qu'il vente ou qu'il pleuve.

Merci aux membres de mon comité d'accompagnement pour leurs conseils avisés : à Alain Content pour son aide en ce qui concerne la programmation de données (mes scripts sont bien plus optimaux qu'ils ne l'étaient avant son intervention) et à Olivier Klein, qui a toujours pris le temps de m'aiguiller lorque j'étais en difficulté et qui restera pour toujours l'un des chercheurs que j'admire le plus, tant humainement qu'intellectuellement.

Merci également aux autres membres du corps académique que j'ai eu le privilège de cotoyer. Merci à Daniël Lakens dont la rencontre a littéralement bouleversé ma carrière. Sa passion pour la recherche et son énergie positive m'ont continuellement poussée à me dépasser et je ne trouverai jamais les mots pour exprimer à quel point je lui suis reconnaissante pour tout ce qu'il m'a apporté. Merci à Christophe Ley dont la rigueur statistique s'est avérée très précieuse, notamment au cours de notre étude sur les tailles d'effet. Merci à Nicolas Kervyn, un allié et un ami qui m'a toujours soutenu et ce depuis le jour de notre rencontre, ainsi qu'à Vincent Yzerbyt d'avoir accepté de faire partie de mon jury de thèse. Enfin, merci à tous les membres académiques de l'ULB. Vous êtes nombreux à avoir enrichi mon parcours de par votre présence, et même si je n'ose citer de noms, de peur d'en oublier certains, je vous suis très reconnaissante pour tout.

L'ULB a été un lieu riche de rencontres exceptionnelles, et je souhaiterais mettre quelques petites perles en lumière. Pour commencer, je tiens à accorder une place d'honneur à Aude Fenaux. Bien plus que la secrétaire du service d'analyse des données, Aude a le don d'enchaîner

au besoin les casquettes d'amie, de conseillère, de psychologue... Aude est un véritable rayon de soleil et je suis persuadée que tous les étudiants et membres du corps académique qui l'entourent mesurent la chance de travailler avec une personne aussi formidable.

Je tiens également à remercier certains collègues que j'affectionne particulièrement et dont la gentillesse, l'humour et la bonne humeur m'ont apporté beaucoup. Je pense notamment à Kenzo, Youri, Lola, Julia, Jan, mais également à mes collègues du service d'analyse de données tels que Bruno, Caroline et Tania, ainsi que tous ceux qui ont quitté l'équipe mais qui restent chers à mon coeur : Dyna, Doris, Anouk et Damien, pour ne citer que les meilleurs.

Je souhaite également remercier les étudiants de manière générale. Ma première passion et motivation à mener cette thèse à terme a toujours été l'enseignement. Nombreux sont les étudiants m'ayant montré des marques d'affection et de soutien au cours de ces années, et cela a été mon moteur.

Enfin, merci à ma famille d'avoir toujours été derrière moi. La pression peut par moment me rendre "légèrement" insupportable et ils se sont toujours montrés compréhensifs malgré tout. Un merci particulier à mon père, Guy, à mes frères, Laurent et Jean-Pierre ainsi qu'à Marc d'avoir relu toutes les parties francophones de cette thèse. Par dessus tout, merci à Vincent, mon futur mari et père de mes 2 petites filles en construction, pour avoir toujours cru en moi, même lorsque je n'y arrivais plus moi-même.

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 et al., 2012). Cependant, des lacunes semblent persister dans la manière dont les statistiques sont enseignées aux futurs psychologues.

Traditionnellement, depuis plus de 50 ans, les tests de comparaison de moyennes (les traditionnels tests t et les ANOVAs) se trouvent au coeur de la grande majorité des programmes dans les domaines des Sciences Psychologiques et de l'Education (Aiken et al., 2008; Curtis & Harwell, 1998; Golinski & Cribbie, 2009) et des livres d'introduction aux statistiques pour psychologues (Field, 2013). Cela pourrait vraisemblablement expliquer pourquoi ces tests sont si persistants dans la recherche en psychologie (Counsell & Harlow, 2017).

Les tests t et ANOVAs sont les tests 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 et 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 et 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 (Aiken et al., 2008; Hoekstra et al., 2012). Pourtant, certains indices semblent contredire cette conviction.

Au sein de cette thèse, nous nous focaliserons exclusivement sur le cas des plans expérimentaux de type inter-sujets avec un seul facteur (ceux utilisés lorsque les individus sont répartis

aléatoirement au sein des différentes conditions définies par les modalités d'un facteur).

Bien qu'il existe plusieurs types de tests t et d'ANOVAs, les chercheurs en psychologie privilégient souvent par défaut le test t de Student et l'ANOVA de Fisher.[3] Il s'agit de tests paramétriques (soit des tests qui impliquent des conditions relatives aux paramètres des populations étudiées, en vue d'être valides) qui consistent à comparer les scores moyens de deux (ou plusieurs) groupes indépendants de sujets. Ces tests 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 et Christianson (2001), par exemple, avaient trouvé que seulement 8% des auteurs reportaient des informations sur les conditions d'application des tests, soit à peine 1% de plus qu'en 1969. Plus récemment, Hoekstra et al. (2012) ont montré que sur 50 articles publiés en 2011 dans *Psychological Science* utilisant au moins une ANOVA, test t ou régression, seulement trois discutaient des questions de normalité et d'homogénéité des variances. Par ailleurs, les informations reportées sont souvent non exhaustives (Counsell & Harlow, 2017), et la condition d'homogénéité des variances est encore moins fréquemment citée que celle de normalité. Parmi les 61 articles analysés par Keselman et al. (1998), seulement 5% 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 et Cribbie (2009) ont fait un constat similaire : parmi les 140 articles qu'ils ont analysés, 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 et Cribbie (2009), par exemple, ont constaté à travers leur 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 experts des revues (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[4] (Hoekstra et al., 2012). 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 si l'on part du postulat que les conditions ne sont 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 et al., 2017; Erceg-Hurn & Mirosevich, 2008; Grissom, 2000; Micceri, 1989; Yuan et al., 2004).

Bien que l'on ne puisse totalement écarter la possibilité que certains chercheurs prennent des décisions inhérentes aux violations des conditions d'application sans le mentionner dans leur article, l'hypothèse privilégiée par Keselman et al. (1998) est que pour la majorité des chercheurs, le choix d'opter pour un test paramétrique se fait généralement indépendamment du fait que les conditions dont ce type de test dépend soient ou non respectées. Une expérience menée par Hoekstra et al. (2012) semble aller dans le même sens : afin d'étudier les pratiques des chercheurs lorsqu'ils étaient confrontés à un scénario qui suggérait la réalisation d'un test t, d'une ANOVA, d'une régression linéaire ou d'une alternative non paramétrique à ces tests, ils 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. Ils ont constaté que tous les doctorants ont opté pour des tests paramétriques et pourtant, les conditions d'application de ces tests n'ont été testées que dans un faible pourcentage de cas. Après l'expérience, les 30 doctorants ont été soumis à un questionnaire. Celui-ci a révélé que la non-vérification des conditions d'application des tests était due à leur manque de familiarité avec les conditions d'application des tests paramétriques, 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 paraît, 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 et al.,

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'hypothèse, mais il a été démontré qu'un test appliqué conditionnellement aux résultats d'un test statistique préliminaire sera généralement associé à des taux inadéquats d'erreur de type I et II (Zimmerman, 2004).

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 sensibles aux violations de ces conditions, particulièrement à celles de la condition d'homogénéité des variances, et cette sensibilité est accentuée lorsque les échantillons n'ont pas tous la même taille (Keselman et al., 1998).

Compte tenu de tous les éléments précités, il semblerait donc 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 tronquée), mais ces derniers font très souvent face à une forte résistance de la part des chercheurs, qui persistent à vouloir comparer les moyennes (Erceg-Hurn & Mirosevich, 2008; Keselman et al., 1998; Wilcox, 1998).

Dans la mesure où une revue approfondie de la littérature démontre que 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 et al., 2012), 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 et al., 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'homogénéité des variances 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'homogénéité des variances, 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 et al., 2012). L'année 1999 a joué un rôle clé dans la mise en oeuvre de ces recommandations, 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 et al., 2013). Or, il semblerait que ces diverses recommandations aient été associées à des modifications dans les pratiques des chercheurs. Peng et al. (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 de Student, 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 lorsque la condition d'homogénéité des variances n'est pas respectée. Nous tenterons de comparer l'efficacité des différents estimateurs sous des déviations réalistes de la condition de normalité, en nous appuyant sur l'investigation de Cain et al. (2017), qui avaient 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).

Notons que notre choix de nous focaliser exclusivement sur les mesures de la famille d dans ce chapitre s'explique par le fait que les chercheurs utilisent très fréquemment le d de Cohen lorsqu'ils réalisent un test t. La non-prise en compte des mesures non standardisées ne doit nullement être interprétée comme un déni de leur intérêt[6] (pour une discussion intéressante sur l'intérêt des mesures non standardisées, nous recommandons l'article de Pek et Flora, 2018).

Enfin, 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 et al., 2013), même lorsque ces mesures sont utilisées indépendamment d'un test d'hypothèse (Counsell & Harlow, 2017). Le calcul des intervalles de confiance n'est pas toujours chose aisée et dans le cas des mesures standardisées, il s'avère particulièrement complexe puisqu'il requiert l'usage des distributions non centrales (Balluerka et al., 2005). C'est pour cette raison qu'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.

Finalement, les mesures de taille d'effet (qu'elles soient ou non standardisées) et leur intervalle 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 et al., 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 moyennes, que la différence de moyennes entre les groupes ne dépasse pas une certaine valeur (Newman et al., 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 et al., 2018). C'est le principe des tests d'équivalence, qui feront l'objet du cinquième chapitre de cette thèse.

Au sein de ce chapitre, nous commencerons par expliquer l'intérêt des tests d'équivalence avant de nous pencher plus spécifiquement sur le TOST (Two One-Sided Tests) dont nous décrirons le principe. Ensuite, nous présenterons un article coécrit avec Daniël Lakens dans lequel nous comparons le TOST à une technique récemment proposée par Blume et al. (2018), à savoir le SGPV (Second Generation P-Value).

Notes de fin de chapitre

[1]Parmi 68 articles analysés en 2013 par Counsell et al. (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 qu'ils réalisent 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 M. Tomczak et E. 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 parue. Cependant, la mention explicite des conditions d'application ne fait pas partie des mises à jour 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 3, cette traduction commune ne représente que mal cette mesure, qui nous informe plus sur la densité des distributions, au niveau des extrémités, que sur leur aplatissement.

[6]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 standardisée (R^2, coefficient de régression standardisé, différence de moyennes standardisée...; Counsell & Harlow, 2017).

Chapitre 2 : Utiliser le test t de Welch

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

Le matériel supplémentaire est disponible au lien suivant : https://difusion.ulb.ac.be/

Independent samples 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 et al., 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 and Student's t-test, and suggest that Welch's t-test is a better default for the social sciences than Student's t-test. 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 if 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. $n \geq 5$)[7], 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 populations, 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).[8][9] 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 population variances is undesirable. Then, we will discuss whether the assumption of equal population variances is plausible in psychology and point out research areas where this assumption is implausible. We will then review differences between Student's t-test and Welch's t-test and show through simulations that the bias in Type I error rates when Student's t-test is used has a larger impact on statistical inferences than the rather modest impact on the Type II 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 when sample sizes are unequal.

Limitations of Two-Step Procedures

Readers may have learned that the assumptions of normality and of equal population 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 populations. Any departure to this assumption, such as unequal population variances, will therefore lead to the rejection of the assumption of equal distributions (Zimmerman, 2000). Alternatives exist, known as the "modern robust statistics" (Wilcox et al., 2013). For example, Yuen's t-test, also called "20% trimmed means test" (Luh & Guo, 2007; Yuen, 1974), is an extension of the Welch's t-test and according to Wilcox and Keselman (2003), it is allegedly more robust in case of non-normal distributions. 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).

With respect to the assumption of equal variances across populations, if the test of the equality of variance is non-significant, and the assumption of equal population 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 population 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 population variances

Researchers have proposed several tests for the assumption of equal population variances. Levene's test and the F-ratio test are the most likely to be used by researchers because they are available in popular statistical software (Hayes & Cai, 2007). Levene's test is the default option in SPSS. Levene's test is the One-Way ANOVA computed on the terms $|X_{ij} - \hat{\theta}_j|$, where X_{ij} is the i^{th} observation in the j^{th} group, and $\hat{\theta}_j$ is the "center" of the distribution for the j^{th} group ($j = 1, 2$; 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).[10] The F-ratio statistic is obtained by computing $\frac{max(S_1,S_2)}{min(S_1,S_2)}$ where S_j is the sample standard deviation of the j^{th} group ($j = 1, 2$). 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 I and Type II 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 et al., 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 II 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 population standard deviation ratio (SDR = $\frac{\sigma_2}{\sigma_1}$ where σ_j is the population standard deviation of the j^{th} group; $j = 1, 2$) : 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 population is bigger than the standard deviation of the first population, and when SDR < 1 the standard deviation of the second population is smaller than the standard deviation of the first population. 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 $n_1 + n_2 = N$; see Figure 1).[11]

As we can see in the graph in Figure 1, 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 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 80% and about 160 to reach a power of 95%. Since such an SDR is already very problematic in terms of the type I 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

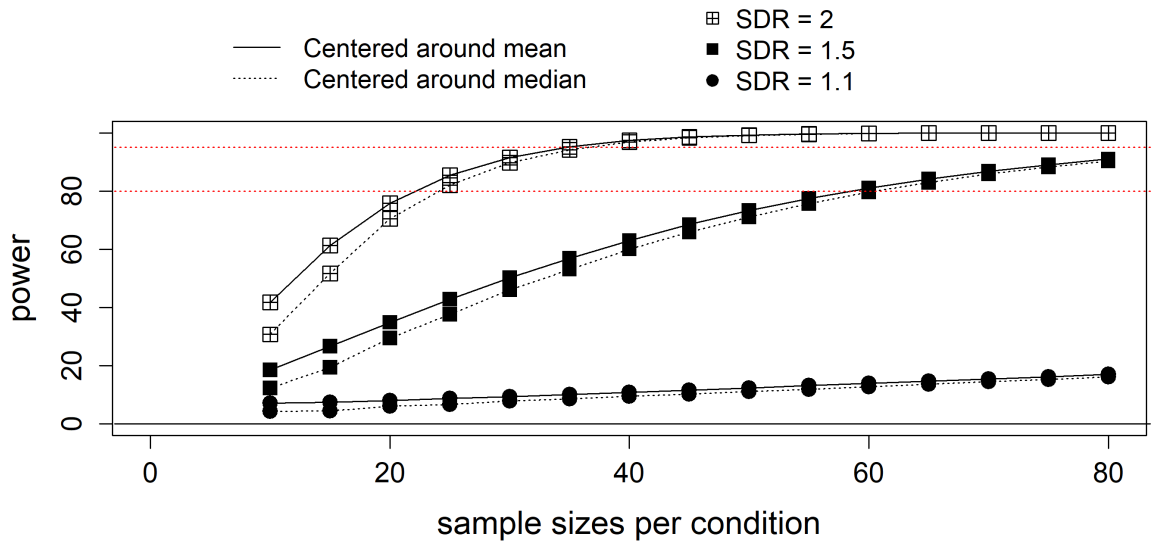

FIGURE 1 – Estimated power of Levene's test as a function of sample size, SDR and centering parameter

becomes even worse for lower SDR : detecting an SDR as small as 1.1 calls for a huge sample size (a sample size of 160 provides a power rate of 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 et al., 2011; Ruxton, 2006; Wilcox et al., 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 to round 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 et al., 2012).[12] 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 population variances is a realistic assumption in psychological research.[13] We will discuss three origins of unequal standard deviations across two populations : 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.

First, 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 et al. (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 populations have different variability by definition. An example from the field of education is the comparison of selective school systems (where students are accepted on basis of selection criterions) versus comprehensive school systems (where all students are accepted regardless of 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.

Second, a quasi-experimental treatment can have a different impact on variances between populations. 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 et al. (1991) noted that intra-individual variability is larger in patients suffering of 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.

Third, while variances of two groups are the same when group assignment is completely randomized, deviation from equality of population variances can occur later, as a consequence of an experimental treatment (Cumming, 2013b; 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 IQ of students from a control group with the IQ 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 estimation $\approx$ 1.32). As proposed by Bryk & 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 et al., 2016), such that we can accurately quantify whether standard deviations differ between populations, and how large the SDR is.

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

So far, we have simply mentioned that Welch's t-test differs from Student's t-test in that is 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) :

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

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

$$df_{Student} = n_1 + n_2 - 2$$

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 population variance is associated with the *larger* sample size, there is a decrease in the nominal Type I error rate (Nimon, 2012; Overall et al., 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 population variance is associated with the *smaller* sample size, the Type I error rate is inflated (Nimon, 2012; Overall et al., 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 population variances as long as the sample sizes of each group are similar (Nimon, 2012; Ruxton, 2006; Wallenstein et al., 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 population variances and sample sizes are unequal (Glass et al., 1972; Overall et al., 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 population variances are unequal becomes visible when we look at the error rates of the test and the influence of both Type I errors and Type II errors. An increase in the Type I error rate leads to an inflation of

the number of false positives in the literature, while an increase in the Type II error rate leads to a loss of statistical power (Banerjee et al., 2009).

To address these limitations of Student's t-test, Welch (1938) 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[14] :

$$t_{Welch} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

The degrees of freedom are computed as follows :

$$df_{Welch} = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1)^2}{n_1 - 1} + \frac{(S_2^2/n_2)^2}{n_2 - 1}}$$

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 II error rate) is very similar for 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. However, differences are very small (in Supplemental Material 5 in the additional file, the larger observed difference is 4.29%). 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 I error rates between Student's t-test and Welch's t-test, we

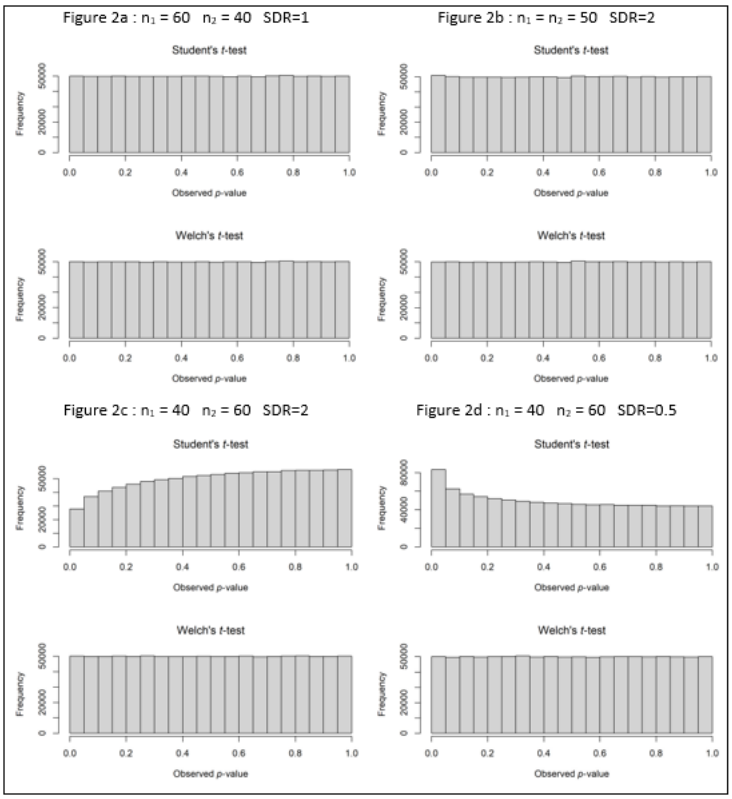

FIGURE 2 – P-value distributions for Student's and Welch's t-test under the null as a function of SDR, and sample size

simulated 1,000,000 studies under the null hypothesis (no difference between the means in each populations) under four scenarios. We chose a small sample ratio ($n_1 = 40$ vs. $n_2 = 60$) to show that when the equal population variances assumption is 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 population (SDR = 1; homoscedasticity assumption met) and sample sizes are unequal (See Figure 2a), with Scenario 2 where the variance differs between populations (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 population variances were unequal between groups and the larger population variance is associated with the larger sample size (SDR = 2; See Figure 2c), and a similar Scenario 4, where the larger population 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 populations or sample sizes are equal, the distribution of Student's p-values is uniform, as expected (see Figures 2a and 2b), 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 population variance is associated with the larger sample size, the frequency of p-values less than 5 percent decreases to .028 (see Figure 2c), and when the larger population variance is associated with the smaller sample size, the frequency of p-values less than 5 percent increases to .083 (see Figure 2d). Welch's t-test has a more stable Type I error rate (see Keselman et al., 1998; Keselman et al., 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 Table 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 I error rates than Student's t-test, even if neither of them give critical values (i.e., values under .025 or above .075, according to the definition of Bradley, 1978). With SDR = 1.1, when the larger population variance is associated with the larger sample size, the frequency of Student's p-value being less than 5 percent decreases to .046 and when the larger population variance is associated with the smaller sample size, the frequency of Student's p-value being less than 5 percent increases to .054. On the other side, the frequency of Welch's p-values being below 0.05 is exactly 5 percent in both cases.

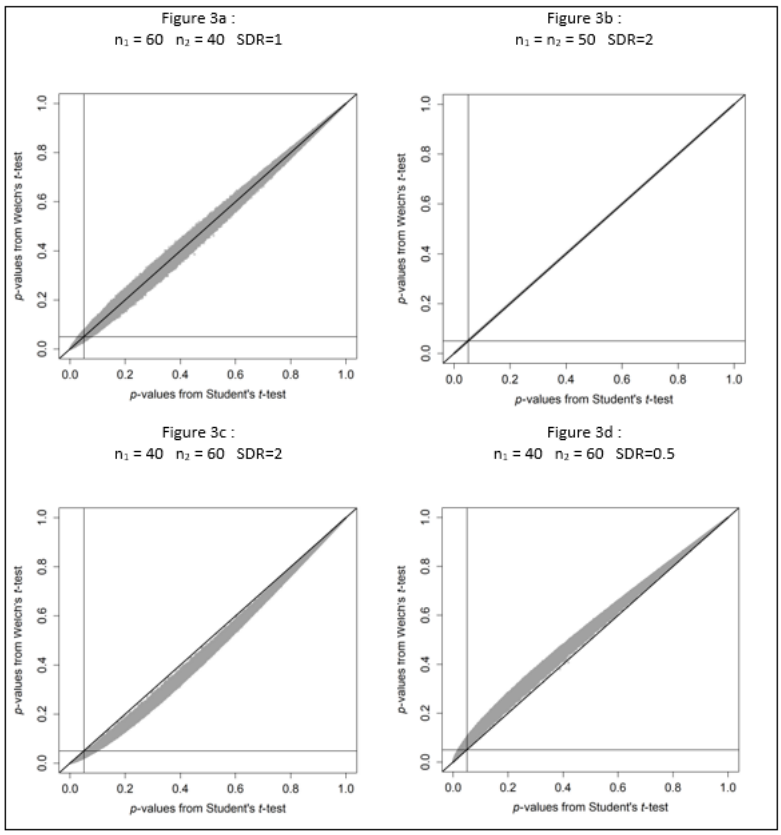

FIGURE 3 – P-value from Student's t-test against p-values from Welch's t-test under the null

In Figure 3, p-values from Welch's and Student's t-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 both population variances are the same (SDR = 1) and sample sizes are unequal. Figure 3b displays Student's p-values plotted against Welch's p-values of Scenario 2, where both population variances differ (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 population variances are unequal and the larger population 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 population 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 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 population variances is not met, Student's t-test yields unreliable results, while Welch's t-test controls Type I 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. 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 I error rate is a lot 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 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.

Endnotes

[7]There is a Type I error rate inflation in a few cases where sample sizes are extremely small and SDR is big (for example, when $n_1 = n_2 = 3$ are sampled from uniform distributions and SDR = 2, the Type I 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 ($n_1 + n_2 \leq 5$), 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.

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

[9]In a simulation that explored Type I 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 .11 or deflated down to .02 (See the Additional file).

[10] Other variants have been proposed such as the 20 percent trimmed mean (Lim & Loh, 1996).

[11]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 % power of Levene's test as $SDR = \frac{1}{X}$. For example, SDR = 2 will return the same answer as SDR = 0.5.

[12] 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) .

[13]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.

[14]Also known as the Satterwaite's test, or the Smith/Welch/Satterwaite test, or the Aspin-Welch test, or the unequal variances t-test.

Chapitre 3 : Utiliser l'ANOVA *W* de Welch

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 *W*-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 populations (or homoscedasticity; see Lix et al., 1996). While a deviation from the normality assumption generally does not strongly affect either the Type I error rates (Glass et al., 1972; Harwell et al., 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 population variances (Grissom, 2000). Unequal population variances can alter both the Type I error rate (David & Johnson, 1951; Harwell et al., 1992) and statistical power (Nimon, 2012; Overall et al., 1995) of the *F*-test.

Although it 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 population variances assumption was taken into account. They reported corrected degrees of freedom for unequal population variances, which could signal the use of the W-test instead of the classical F-test. A similar investigation (Hoekstra et al., 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 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 violations of the homoscedasticity assumption. 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 et al., 2017; Micceri, 1989; Yuan et al., 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 could also change the shape of a distribution, 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 assumption. Subgroups might exist that are 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 lead 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 population variances (or homoscedasticity) is a mathematical requirement that is also ecologically unlikely (Erceg-Hurn & Mirosevich, 2008; Grissom, 2000). In a previous paper (Delacre et al., 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 et al. (2010). For example, Green et al. (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 ($S_{spanish} = .80 > S_{english} = .50$, with S = sample standard deviation). For men, the reverse was true ($S_{spanish} = .97 < S_{english} = 1.33$).[15]

Third, even when the population variances 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, 2013b; 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-β) 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 -.4 and .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, according to Srivastava (1959) and Boneau (1960), 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 populations 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 violation and unequal variances. Indeed, the effect of heteroscedasticity is relatively constant regardless of the shape of the distribution.

Based on mathematical explanations and Monte 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 equation 1, the F statistic is calculated by dividing the inter-group variance by a pooled error term :

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

where S_j^2, $\bar{X}_j$ and n_j are respectively the variance, mean and size of the j^{th} sample ($j = 1, ..., k$), $N = \sum_{j=1}^{k} n_j$, and $\bar{X}_{..}$ is overall mean. The degrees of freedom in the numerator and in the denominator of the F-test are computed as follows :

$$df_n = k - 1$$

$$df_d = N - k$$

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 the larger variance is associated with the larger sample size, the error term, which is the denominator in equation 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 equation 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 :

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

where X_j and S_j^2 are respectively the mean and variance of the j^{th} sample ($j = 1, ...k$), and $\bar{X}_{..}$ is the overall mean. As it can be seen in equation 2, 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, each sample variance is weighted by 1 minus the relative frequency of each sample. This adjustment implies that the variance associated with the smallest sample 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) :

$$df_n = k - 1$$

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

Equation 3 provides the computation of the Welch's statistic (W). In the numerator of the W statistic, the squared deviation between sample 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).

$$W = \frac{\frac{1}{k-1}\sum_{j=1}^{k}\left[w_j(\bar{X}_j - \bar{X}')^2\right]}{1 + \frac{2(k-2)}{k^2-1}\sum_{j=1}^{k}\left[(\frac{1}{n_j-1})(1-\frac{w_j}{w})^2\right]} \tag{3}$$

where $w_j = \frac{n_j}{S_j^2}$, $w = \sum_{j=1}^{k} w_j$ and $\bar{X}' = \frac{\sum_{j=1}^{k}(w_j \bar{X}_j)}{w}$. The degrees of freedom of the W-test are approximated as follows :

$$df_n = k - 1$$

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

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 vs. *W*-test vs. *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 the last group had a

different population mean ($\mu_k = \mu_j + 1$).

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 populations).

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 largest 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 two groups, the F-test becomes increasingly liberal as soon as the variances of the distributions in each population are not similar, even when sample sizes are equal between samples (Harwell et al., 1992; Quensel, 1947).

For didactic reasons, we will report only the results where we compare 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 Di-fusion : "Type I error rate".

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.

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

		***n*-ratio**		
		1	**>1**	**<1**
	1	a	b	c
***SD*-ratio**	**>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 size of the last sample is higher than the size of the first sample, n-ratio > 1, and when the size of the last sample is smaller than the size of the first sample, 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.

In all Figures presented below, averaged results for each subcondition are presented under 7 different configurations of distributions, using the legend provided in Figure 4.

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 all population means are equal, the Type I error rate of all tests should be equal to the nominal alpha level. We assessed the Type

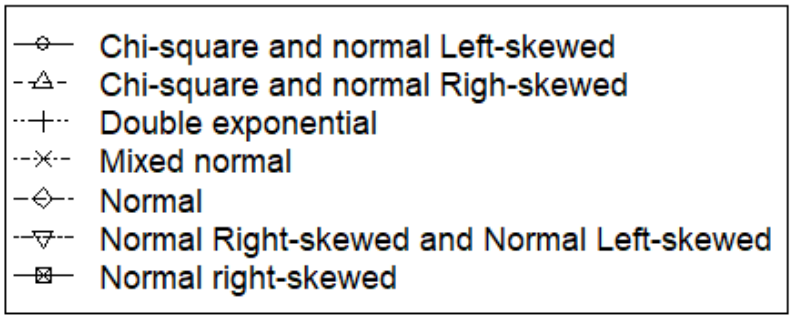

FIGURE 4 – Legend

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 9 cells of Table 1 simplify into 5 subconditions:

— Equal sample sizes and population *SDs* (a);
— Unequal sample sizes but equal population *SDs* (b and c);
— Unequal population *SDs* but equal sample sizes (d and g);
— Unequal sample sizes and population *SDs*, with positive correlation between sample sizes and population *SDs* (e and i);
— Unequal sample sizes and population *SDs*, with negative correlation between sample sizes and population *SDs* (f and h).

In Figures 5 to 9 (see Figure 4 for the legend), we computed the average Type I error rate of the three tests under these 5 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 $[.5 \times \alpha;\ 1.5 \times \alpha]$. 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 $[.9 \times \alpha;\ 1.1 \times \alpha]$.

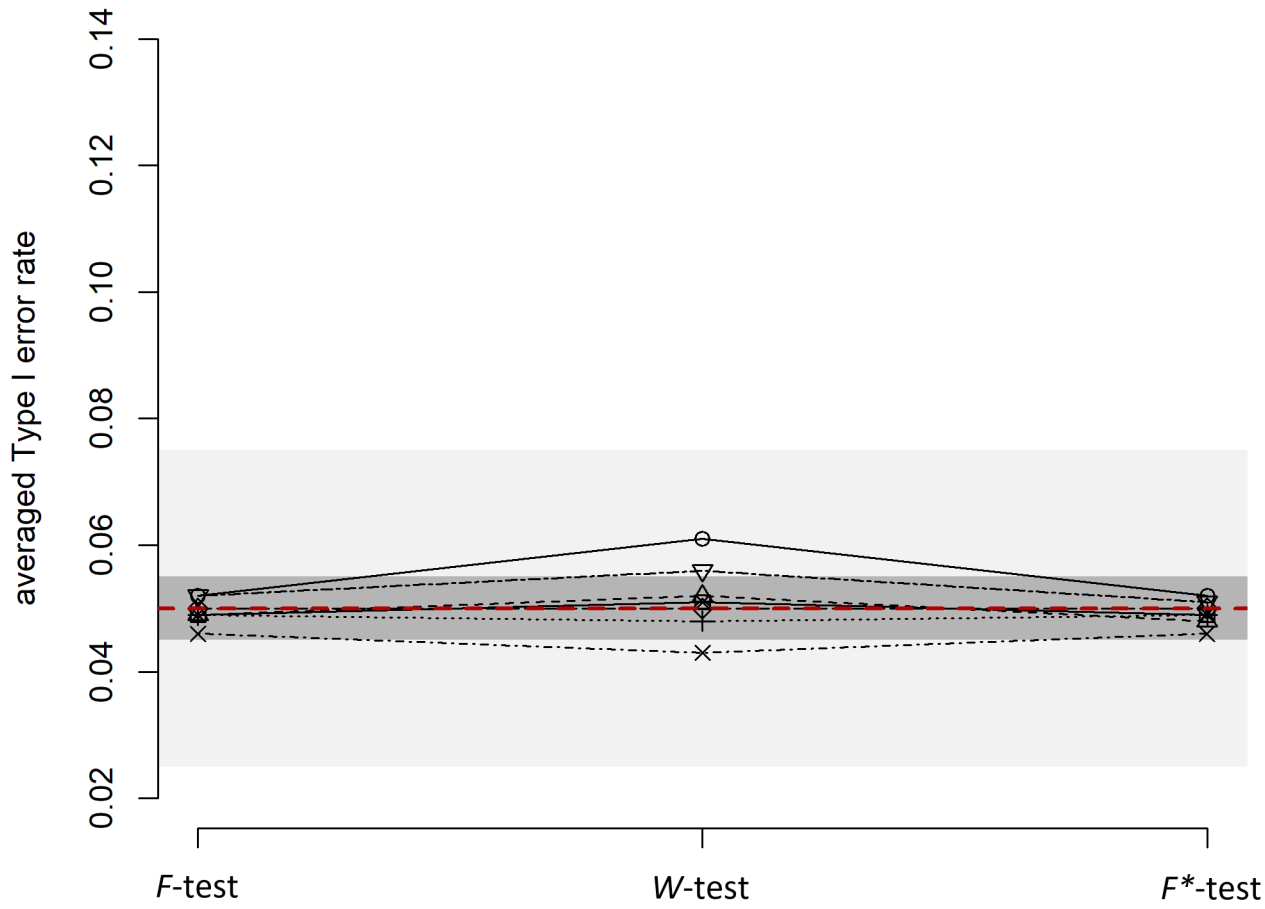

FIGURE 5 – Type I error rate of the F-test, W-test and F^*-test when SDs and sample sizes are equal across groups (cell a in Table 1)

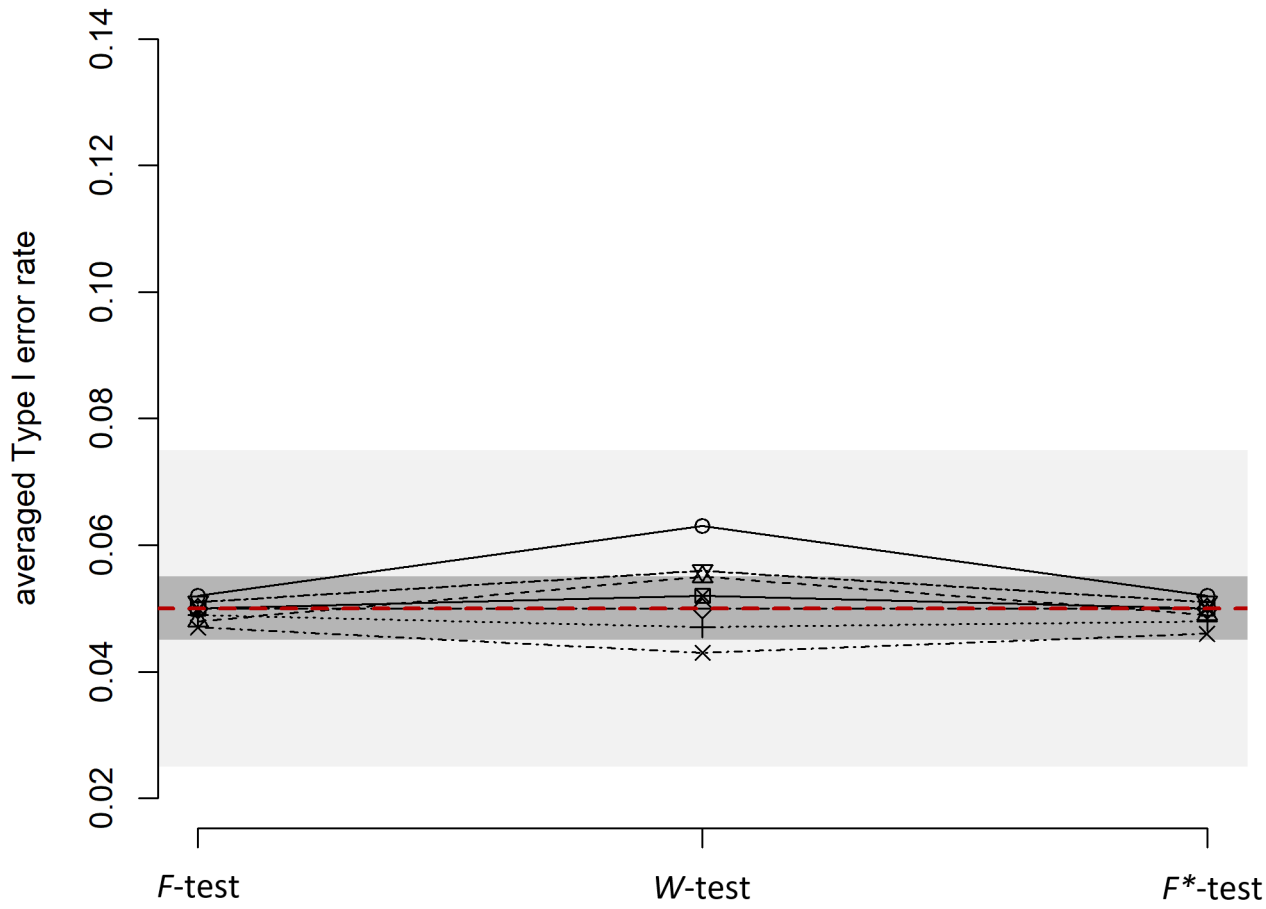

FIGURE 6 – Type I error rate of the F-test, W-test and F^*-test when SDs are equal across groups and sample sizes are unequal (cells b and c in Table 1)

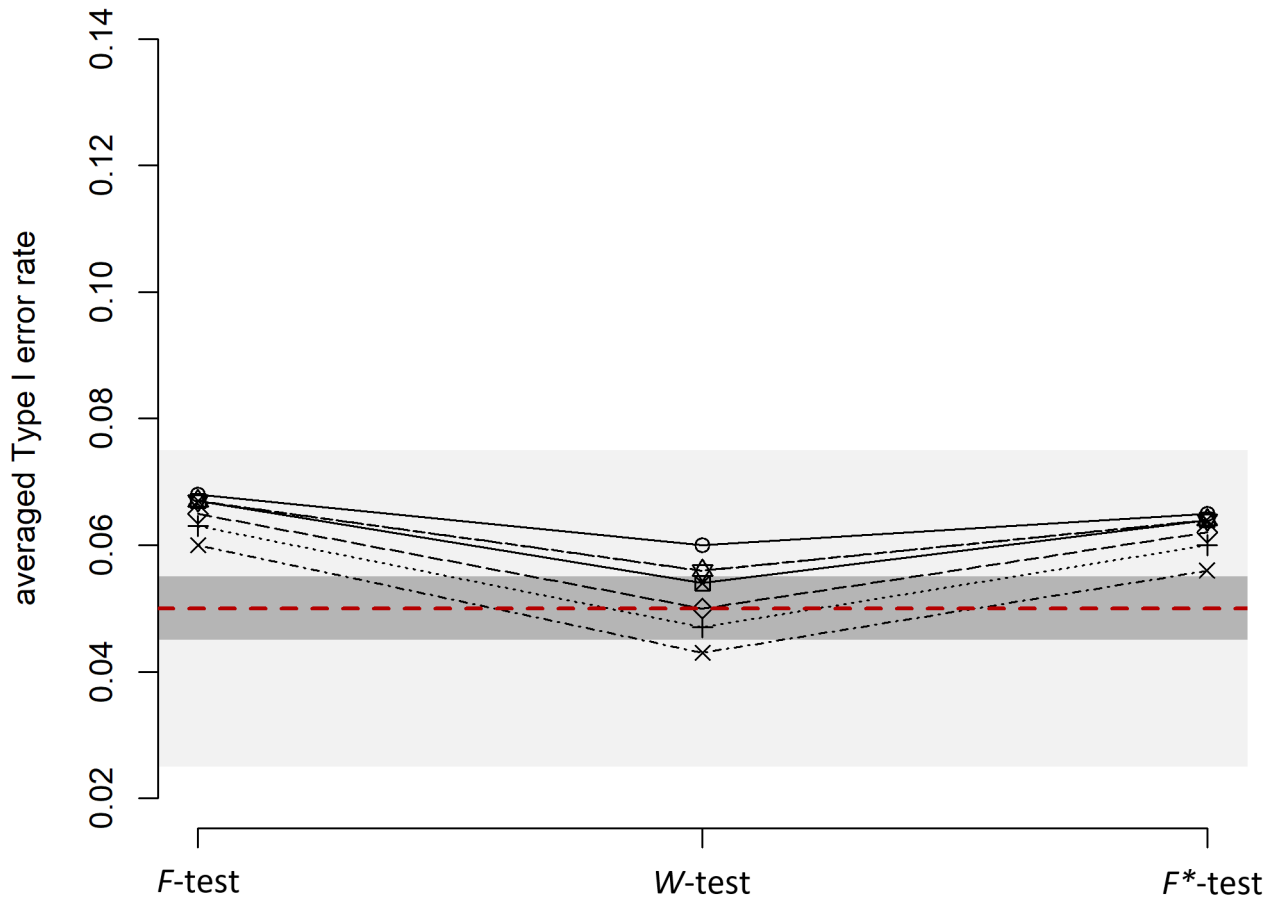

FIGURE 7 – Type I error rate of the F-test, W-test and F^*-test when SDs are unequal across groups and sample sizes are equal (cells d and g in Table 1)

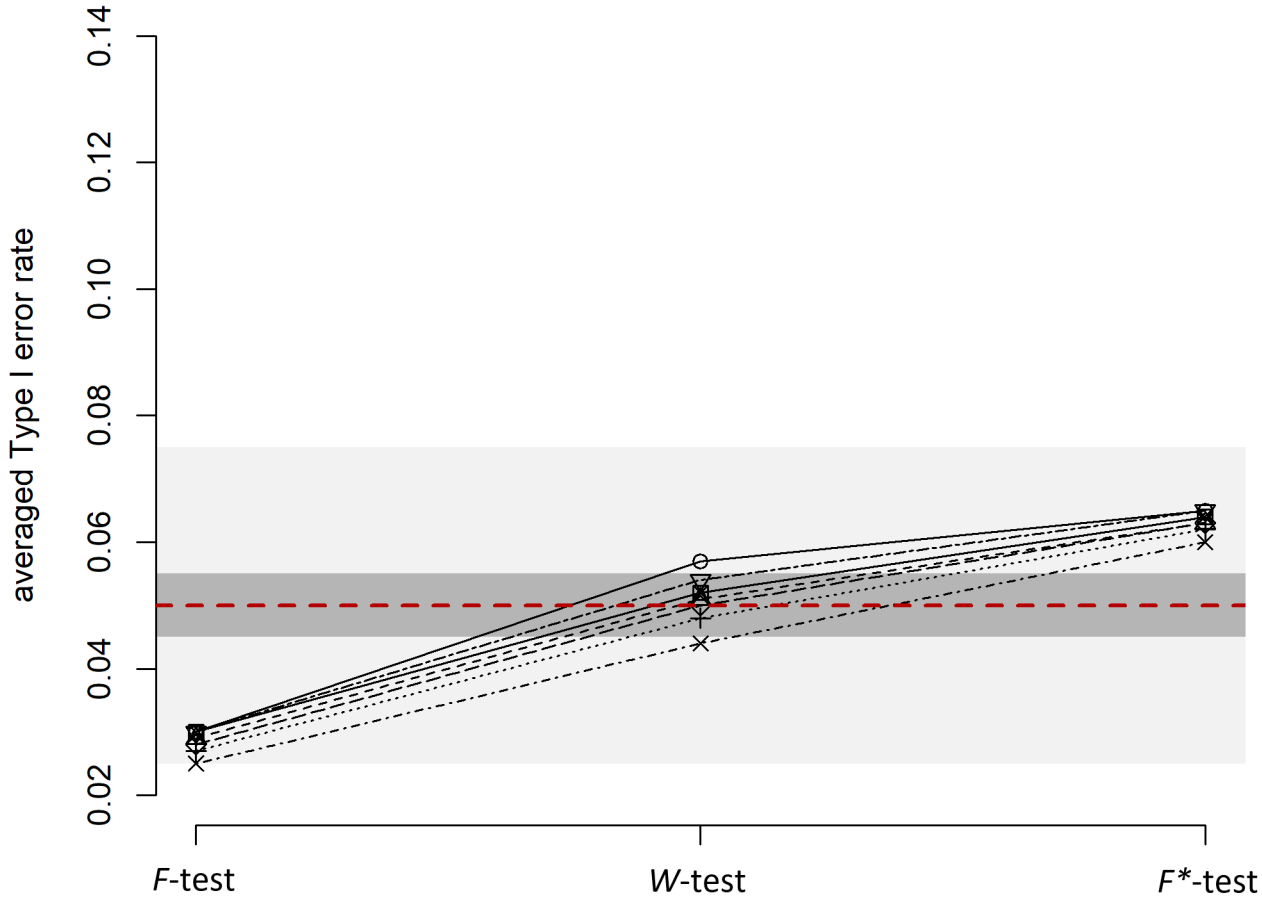

FIGURE 8 – Type I error rate of the F-test, W-test and F^*-test when SDs are unequal across groups, and there is a positive correlation between sample sizes and SDs (cells e and i in Table 1)

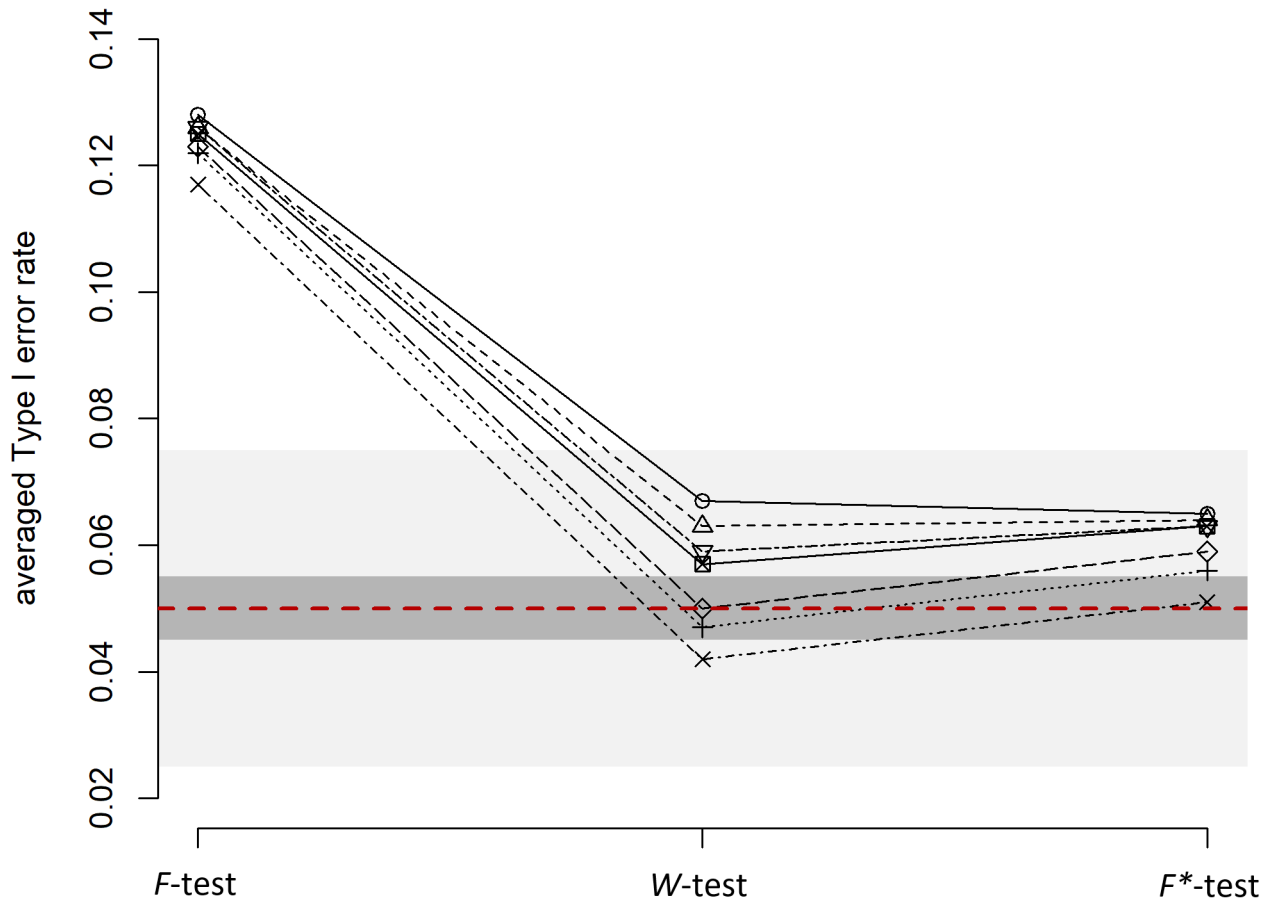

FIGURE 9 – Type I error rate of the F-test, W-test and F^*-test when SDs are unequal across groups, and there is a negative correlation between sample sizes and SDs (cells f and h in Table 1)

In Figures 5 and 6 (cells a, b, and c in Table 1), 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 7, 8, and 9 (cells d to i in Table 1), the homoscedasticity assumption is not met. When sample sizes are equal across groups (Figure 7) and when there is a positive correlation between sample sizes and *SDs* (Figure 8), 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, Rubinstein, Hayes, & Olds (1992), Glass et al. (1972), Nimon (2012) and Overall, Atlas, & Gibson (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 9), 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) have a larger impact on the Type I error rate than when there is homoscedasticity. This is consistent with conclusions by Lix et al. (1996) about 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 robust 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 sample sizes 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-β) 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 has 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 subconditions 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 expected power, divided by the expected power :

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

When consistency equals zero, the observed power is consistent with the expected 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.

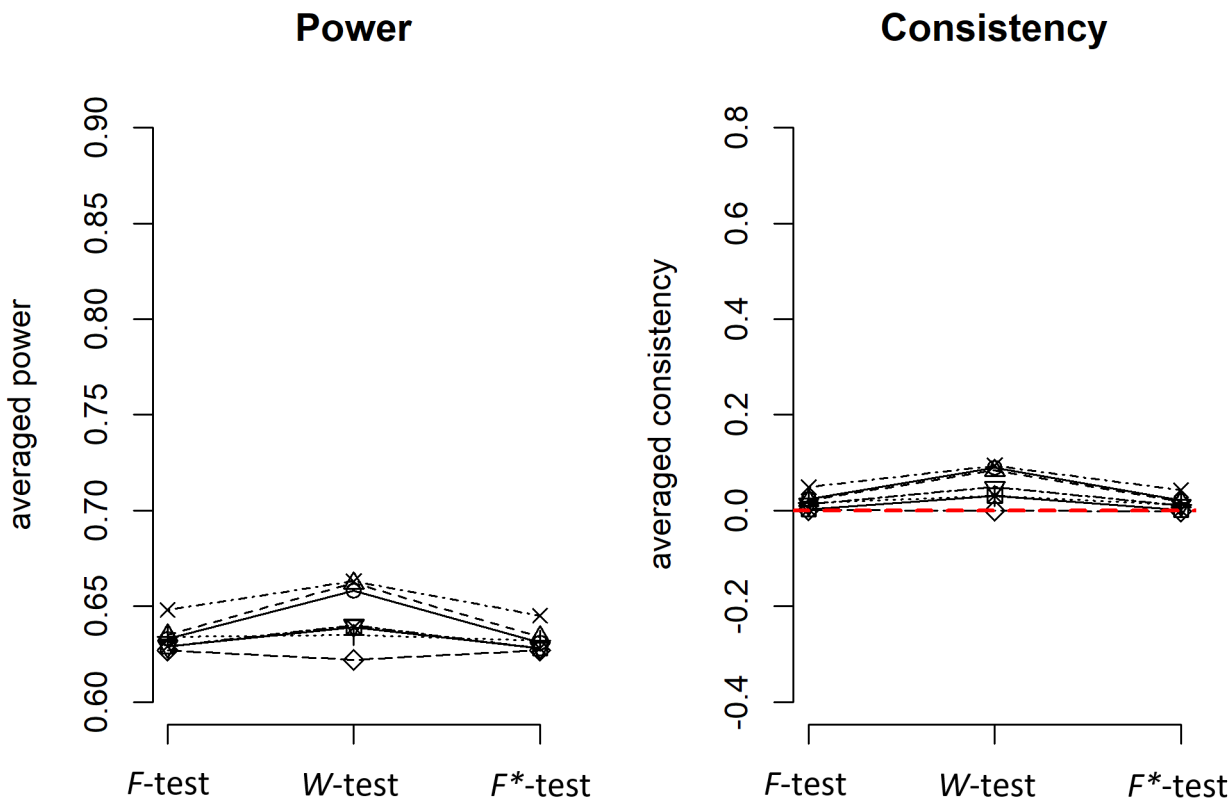

FIGURE 10 – Power and consistency of the F-test, W-test and F^*-test when SDs and sample sizes are equal across groups (cell a in Table 1)

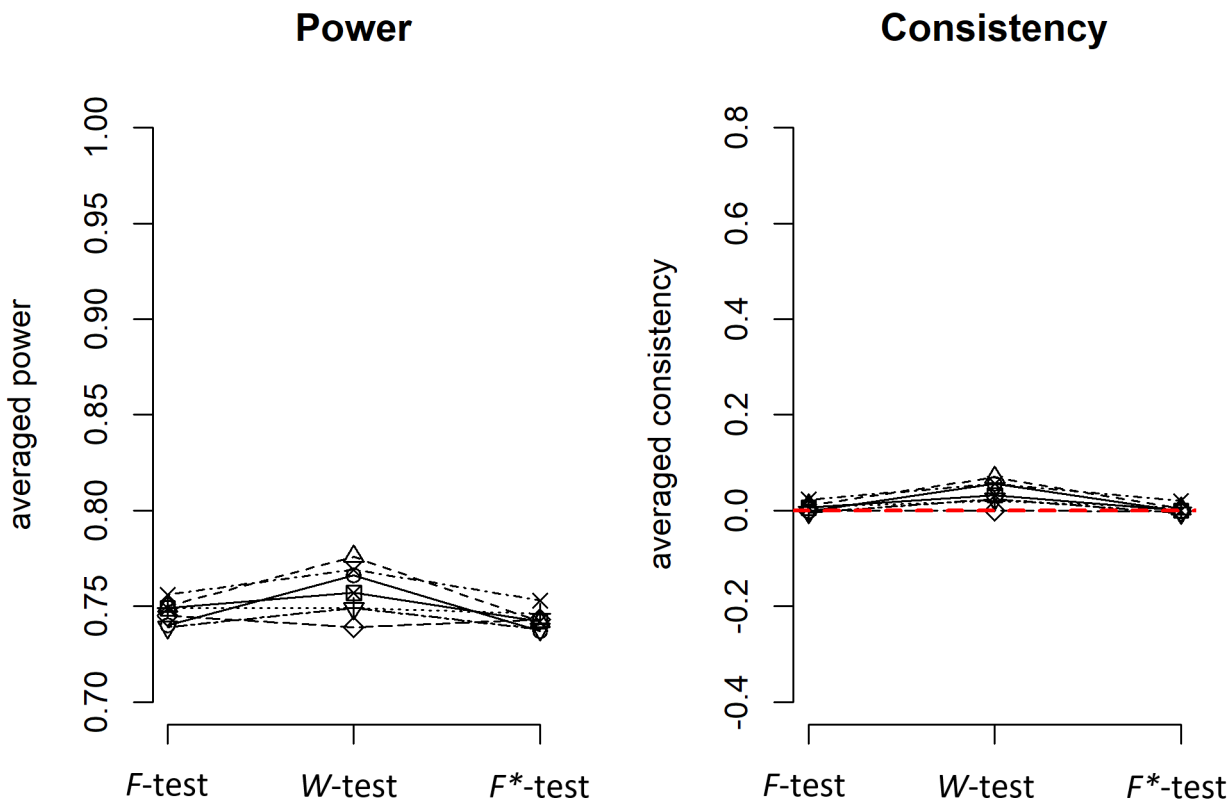

FIGURE 11 – Power and consistency of the F-test, W-test and F^*-test when SDs are equal across groups, and there is a positive correlation between sample sizes and means (cell b in Table 1)

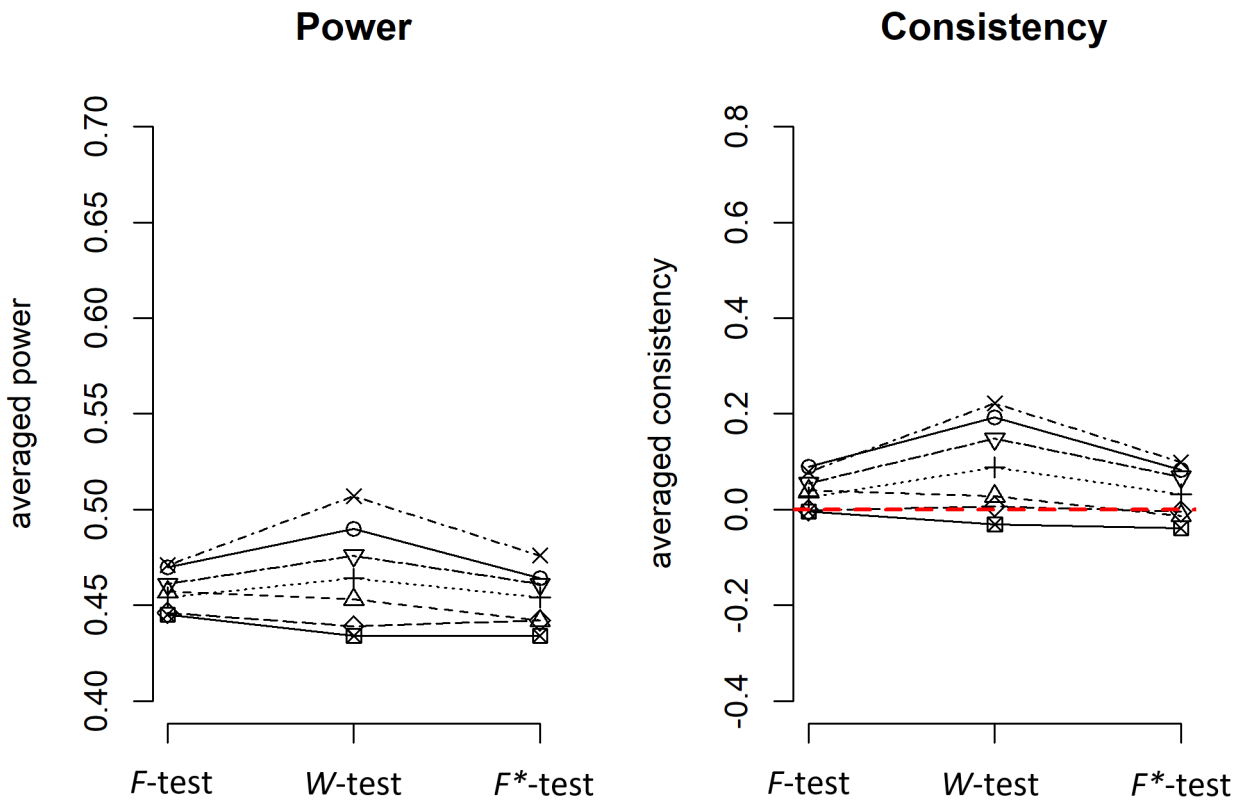

FIGURE 12 – Power and consistency of the F-test, W-test and F^*-test when SDs are equal across groups, and there is a negative correlation between sample sizes and means (cell c in Table 1)

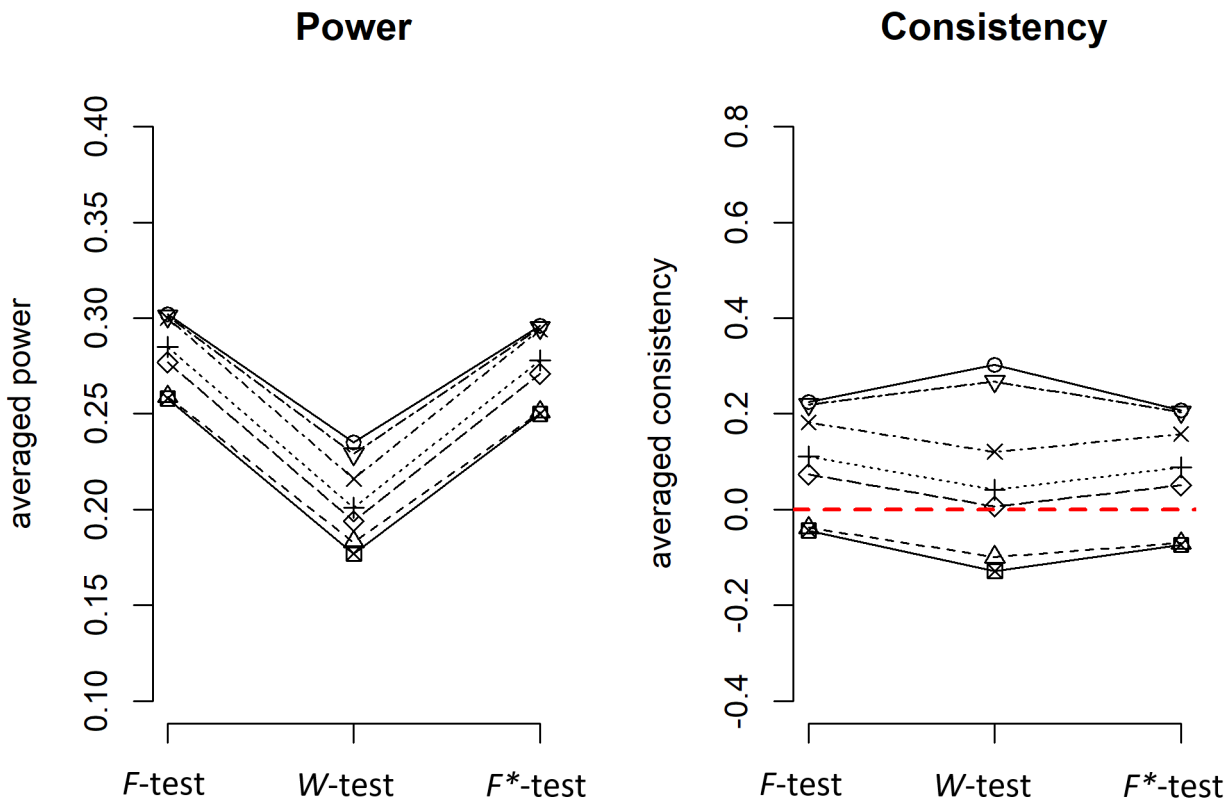

FIGURE 13 – Power and consistency of the F-test, W-test and F^*-test when SDs are unequal across groups, there is a positive correlation between SDs and means, and sample sizes are equal (cell d in Table 1)

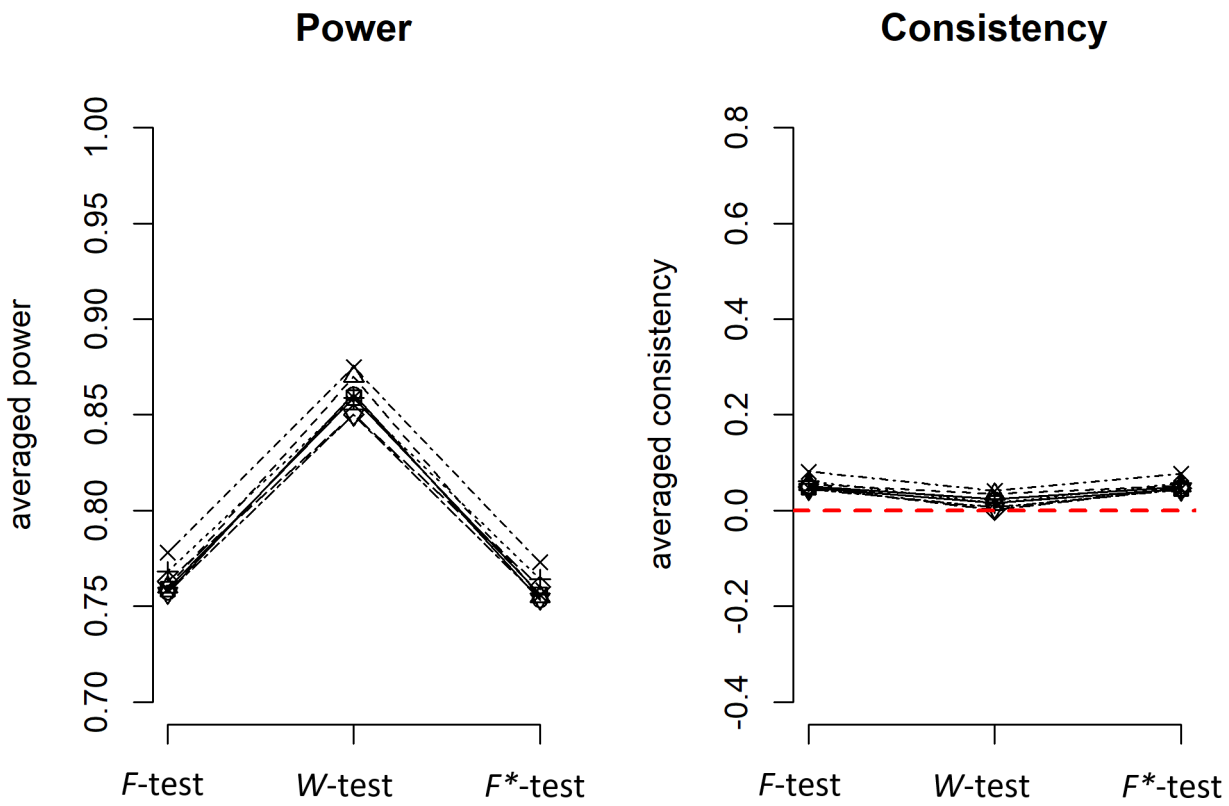

FIGURE 14 – Power and consistency of the F-test, W-test and F^*-test when there SDs are unequal across groups, there is a negative correlation between SDs and means, and sample sizes are equal (cell g in Table 1)

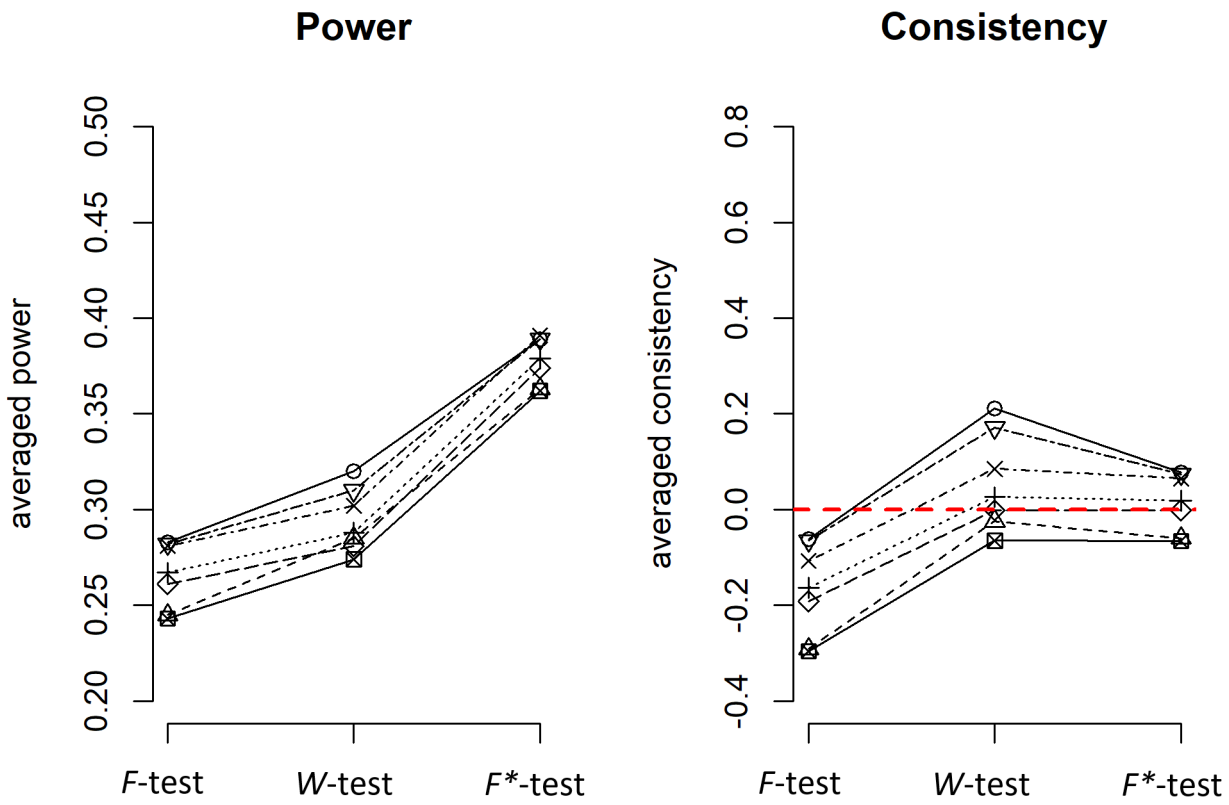

FIGURE 15 – Power and consistency of the F-test, W-test and F^*-test when SDs are unequal across groups, there is a negative correlation betwen sample sizes and SDs, and a positive correlation between SDs and means (cell f in Table 1)

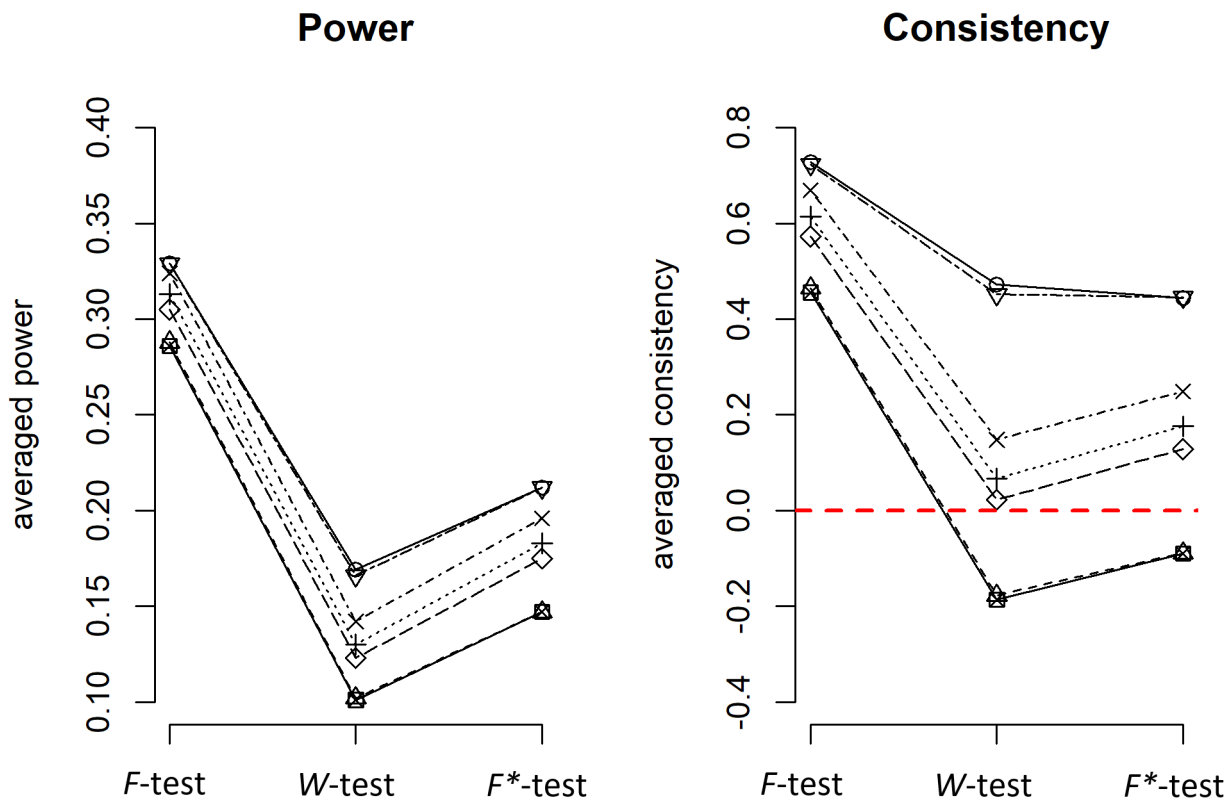

FIGURE 16 – Power and consistency of the F-test, W-test and F^*-test when SDs are unequal across groups, there is a negative correlation betwen sample sizes and SDs, and a negative correlation between SDs and means (cell h in Table 1)

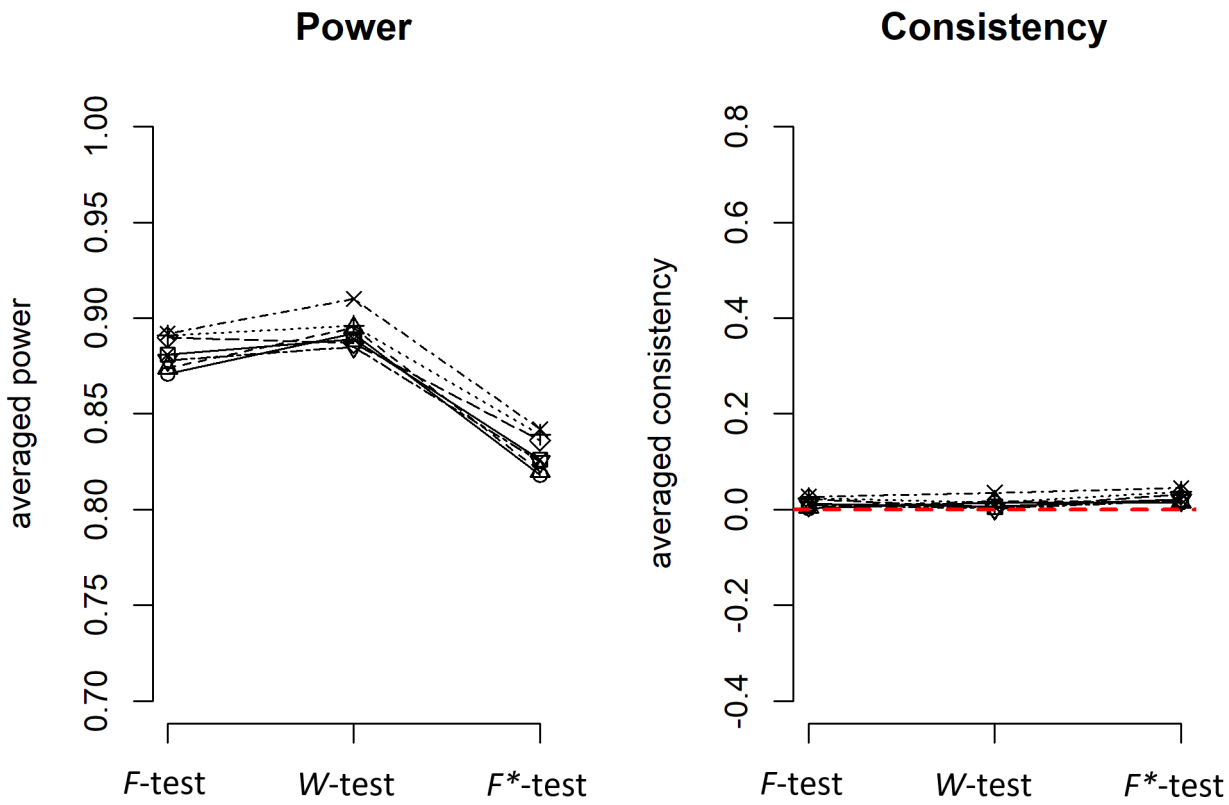

FIGURE 17 – Power and consistency of the F-test, W-test and F^*-test when SDs are unequal across groups, there is a positive correlation betwen sample sizes and SDs, and a positive correlation between SDs and means (cell e in Table 1)

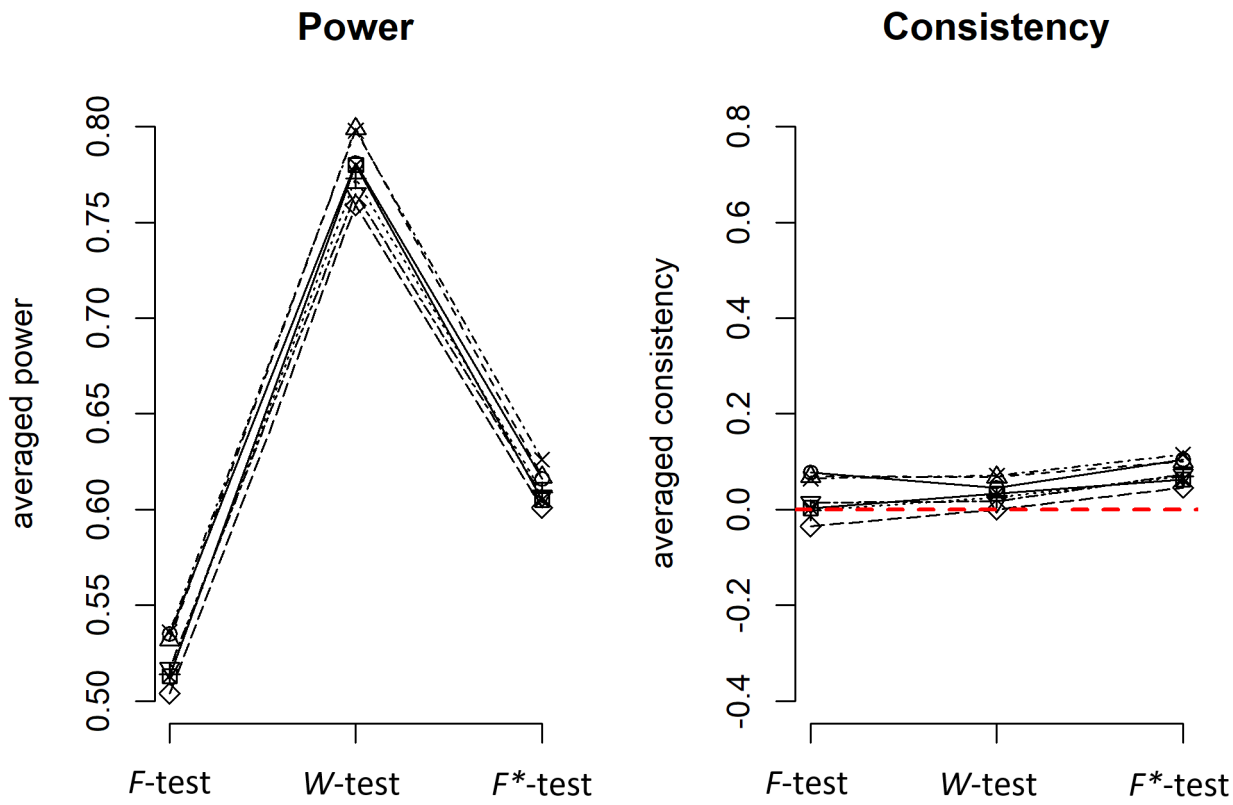

FIGURE 18 – Power and consistency of the F-test, W-test and F^*-test when SDs are unequal across groups, there is a positive correlation betwen sample sizes and SDs, and a negative correlation between SDs and means (cell i in Table 1)

In Figures 10, 11 and 12 (cells a, b, and c in Table 1), the homoscedasticity assumption is met. When distributions are normal, the W-test is slightly less powerful than the F-test and F*-test, even through 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 13 to 18 (cells d to i in Table 1, see Figure 4 for the legend), the homoscedasticity assumption is not met. When sample sizes are equal across groups (Figures 13 and 14), the F-test and the F*-tests are equally powerful, and have the same consistency, the correlation between the *SD* and the mean does not matter. 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 largest population mean is associated with the largest population variance (Figure 13), 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 largest population mean is associated with the smallest population variance (Figure 14), 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 15 and 16), 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 17 and 18), 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 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, 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 positive correlation between the *SD* and the sample size (meaning that the largest group has the largest 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 populations and distributions are not normal, the W-test is slightly more powerful (but a bit less consistent) than both the F-test and the F^*-test, even with heavy-tailed distributions.
3) When the assumption of equal population 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 population 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. We therefore 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)[16] or nonparametric tests. For more information about robust alternatives that are based on other parameters than the mean, see Erceg-Hurn and Mirosevich (2008).

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

Why Hedges' g^* 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 dans ce chapitre n'a pas encore été accepté pour publication.

Le matériel supplémentaire est disponible au lien suivant : https://difusion.ulb.ac.be/

Effect sizes are an important outcome of empirical research. Moving beyond decisions about statistical significance, there is a strong call for researchers to report and interpret effect sizes and associated confidence intervals. This practice is highly endorsed by the American Psychological Association (APA) and the American Educational Research Association (American Psychological 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, where the sample mean difference is divided by the pooled sample standard deviation (Peng et al., 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 et al., 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, Glass's d, Shieh's d and Cohen's d^*) and with correction for bias (Hedges' g, Glass's g, Shieh's g and Hedges' g^*). 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 et al., 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).

Inferential 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 et al., 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 19). 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

$$\delta_{bias} = E(\hat{\delta}) - \delta \tag{4}$$

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 2 and 3 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 :

$$\delta_{relative\ bias} = \frac{E(\hat{\delta}) - \delta}{\delta} \tag{5}$$

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, since any single estimate might be very far from the true population value. Therefore it is not

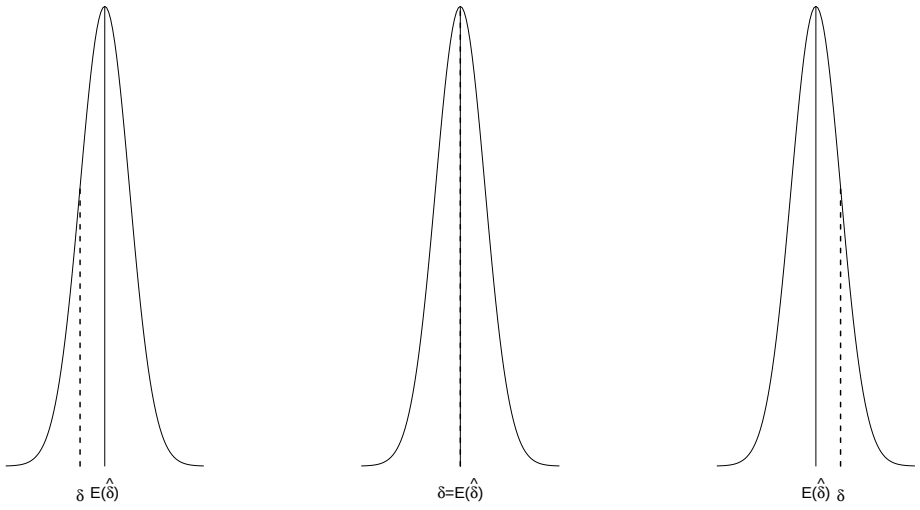

FIGURE 19 – Sampling distribution for a positively biased (left), an unbiased (center) and a negatively biased estimator (right)

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

$$Var(\hat{\delta}_1) \leq Var(\hat{\delta}_2) \tag{6}$$

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.[17] 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 :

$$relative\ var(\hat{\delta}_1) = \frac{Var(\hat{\delta})}{\delta^2} \tag{7}$$

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

$$\delta = \frac{\mu_1 - \mu_2}{\sigma} \tag{8}$$

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 and Hedges' g), and effect sizes that do not assume equal variances (Glass's d, Shieh's d, Cohen's d^*, Glass's g, Shieh's g, and Hedges' g^*). 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, where the sample mean difference is divided by a pooled error term (Cohen, 1965) :

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

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 (Keselman et al., 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 is directly related to Student's t-statistic :

$$t_{Student} = \frac{Cohen's\ d}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \leftrightarrow Cohen's\ d = t_{Student} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{9}$$

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

$$df_{Student} = n_1 + n_2 - 2 \tag{10}$$

and noncentrality parameter[18]

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

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 9 and the theoretical distribution of Student's t-statistic allow us to determine the sampling distribution of Cohen's d, and therefore, its expectation and variance when the assumptions of normality and equal variances are met. All these equations are provided in Table 2. For interested readers, Supplemental Material 1 provides a detailed examination of the theoretical bias and variance of Cohen's d based on Table 2, as well as the bias and variance of all other estimators described later, based on Tables 3 and 4, 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 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 bias with small sample sizes, Hedges and Olkin (1985) defined a bias-corrected version :

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

where $df_{Student}$ has been defined in equation 10 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 :

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

where N is the total sample size. Hedges' g is theoretically unbiased when the assumptions of normality and equal variances are met (see Table 2). Moreover, while the variance of both Cohen's d and Hedges' g depend on the same parameters (i.e. the total sample size (N) and the sample sizes ratio $\left(\frac{n_2}{n_1}\right)$), Hedges' g is less variable, especially with small sample sizes.[19]

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; Grissom & Kim, 2005; Kelley, 2005; Shieh, 2013). When variances are unequal between groups, the expression in equation 8 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 et al., 2017; Delacre et al., 2017; Delacre et al., 2019; Erceg-Hurn & Mirosevich, 2008; Glass et al., 1972; Grissom, 2000; Micceri, 1989; Yuan et al., 2004), both Cohen's d and Hedges' g should be abandoned in favor of an alternative robust to unequal population variances.

TABLE 2: Expectation, bias and variance of Cohen's d and Hedges' g under the assumptions that independent residuals are normally distributed with equal variances across groups.

	df	Expectation	Variance
Cohen's d	$N-2$	$\delta_{Cohen} \times c_f$	$\frac{N \times df}{n_1 n_2 \times (df-2)} + \delta^2_{Cohen}\left[\frac{df}{df-2} - c_f^2\right]$
		$\approx \frac{\delta_{Cohen}}{\left(1-\frac{3}{4N-9}\right)}$	$\approx \frac{N \times df}{n_1 n_2 \times (df-2)} + \delta^2_{Cohen}\left[\frac{df}{df-2} - \left(\frac{1}{1-\frac{3}{4N-9}}\right)^2\right]$
Hedges' g	$N-2$	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 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 is an unbiased estimator, because its expectation equals δ_{Cohen}; equations in this table 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, (B) the Shieh's d and (C) the Cohen's d^*.

Glass's *d* When comparing one control group with one experimental group, Glass et al. (1981) recommend using the standard deviation of the control group as standardizer. This yields

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

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, 2013a). 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 is defined as in Algina et al. (2006) :

$$Glass's\ d \sim \sqrt{\frac{1}{n_c} + \frac{\sigma_e^2}{n_e \times \sigma_c^2}} \times t_{df,ncp} \tag{11}$$

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

$$df = n_c - 1 \tag{12}$$

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

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 11, we can compute its theoretical expectation and variance when the assumption of normality is met (see Table 3), and therefore determine which factors influence bias and variance, and how they do so (see Supplemental Material 1).

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

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

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 in this table require $df \geq 3$ and at least 2 subjects per group.

Shieh's *d* Kulinskaya and 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 and Staudte, proposed a modification of the exact *SD* of the sample mean difference :

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

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

$$Shieh's\ d = \frac{t_{Welch}}{\sqrt{N}} \leftrightarrow t_{welch} = Shieh's\ d \times \sqrt{N} \tag{13}$$

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 (Welch, 1938) :

$$df_{Welch} = \frac{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^2}{\frac{(\sigma_1^2/n_1)^2}{n_1 - 1} + \frac{(\sigma_2^2/n_2)^2}{n_2 - 1}} \tag{14}$$

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

where $\delta_{Shieh} = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n1/N} + \frac{\sigma_2^2}{n2/N}}}$. The relationship described in equation 13 and the theoretical distribution of Welch's t-statistic allow us to approximate the sampling distribution of Shieh's d. Based on the sampling distribution of Shieh's d, we can estimate its theoretical expectation and variance under the assumption of normality (see Table 3), and thereby determine which factors influence bias and variance, and how they do so (see Supplemental Material 1).

As demonstrated in Appendices 1 and 2, when variances and sample sizes are equal across groups, the biases and variances of Cohen's d and Shieh's d are identical except for multiplication by a constant. The same is true for the estimators δ_{Cohen} and δ_{Shieh} :

$$\delta_{Cohen} = 2 \times \delta_{Shieh} \quad (\text{considering } \sigma_1 = \sigma_2 \text{ and } n_1 = n_2) \tag{15}$$

$$Bias_{Cohen's\ d} = 2 \times Bias_{Shieh's\ d} \quad (\text{considering } \sigma_1 = \sigma_2 \text{ and } n_1 = n_2) \tag{16}$$

$$Var_{Cohen's\ d} = 4 \times Var_{Shieh's\ d} \quad \text{(considering } \sigma_1 = \sigma_2 \text{ and } n_1 = n_2) \tag{17}$$

We can deduce from equations 15, 16 and 17 that relative to their respective population effect size, Cohen's d and Shieh's d 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 5 and 7) over the most commonly used raw bias and variance (previously defined in equations 4 and 6).

When sample sizes are not equal, according to the statistical properties of Welch's statistic under heteroscedasticity, Shieh's d 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 (2013a) 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 will vary as a function of the sample sizes allocation ratio (unlike Cohen's d^* 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 *d** 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 (based on the pooled standard deviations) and Cohen's d^* with an asterisk. This yields :

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

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^* (Huynh, 1989) :

$$Cohen's\ d^* \sim \sqrt{\frac{2(n_2 \times \sigma_1^2 + n_1 \times \sigma_2^2)}{n_1 n_2 (\sigma_1^2 + \sigma_2^2)}} \times t_{df^*, ncp^*} \tag{18}$$

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

$$df^* = \frac{(n_1 - 1)(n_2 - 1)(\sigma_1^2 + \sigma_2^2)^2}{(n_2 - 1)\sigma_1^4 + (n_1 - 1)\sigma_2^4} \tag{19}$$

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

where $\delta^*_{Cohen} = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2 + \sigma_2^2}{2}}}$. Using equation 18 we can compute its theoretical expectation and variance when the assumption of normality is met (see Table 3), 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; 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, Shieh's g and Hedges' g^* As for Cohen's d, a Hedges' correction can be applied in order to compensate for the bias of Glass's d, Shieh's d and Cohen's d^* with small sample sizes (see Table 3). This correction has the following general form :

$$g = d \times \frac{\Gamma(\frac{\nu}{2})}{\sqrt{\frac{\nu}{2}} \times \Gamma(\frac{\nu-1}{2})}$$

where the distinct values of ν are provided in equation 12 for Glass's g, in equation 14 for Shieh's g and in equation and 19 for Hedges' g^*. 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 4. In summary :

— The variances of Hedges' g^* and Shieh's g 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 $\left(\frac{n_2}{n_1} \times \frac{\sigma_2}{\sigma_1}\right)$.

— The variance of Glass's g 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 4: Expectation, bias and variance of Glass's d and Cohen's d^* and Shieh's d under the assumption that independent residuals are normally distributed.

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

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 in this table 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 Hedges g, Glass's g (using respectively the sample SD of the first or second group as a standardizer), Hedges' g^* and Shieh's g.

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 et al. (2017). Cain et al. (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 G_2 = -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. G_2=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 G_1 = -2.08 and G_1 = 6.32. We also simulated samples extracted from population where $G_1 = 0$, 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 5.

TABLE 5: 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 5. 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 5), 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 = $\frac{n_2}{n_1}$), the population SD-ratio (i.e. $\frac{\sigma_2}{\sigma_1}$), and the sample size and population variance pairing $\left(\frac{n_2}{n_1} \times \frac{\sigma_2}{\sigma_1}\right)$. In our scenarios, μ_2 was always 0 and μ_1 varied from 1 to 4, in steps of 1 (so does $\mu_1 - \mu_2$).[20] Moreover, σ_1 always equals 1, and σ_2 equals .1, .25, .5, 1, 2, 4 or 10, and therefore, the SD-ratio were 10, 4, 2, 1, .5, .25 or .1. 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 6. 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 6: 4 conditions based on the n-ratio and the *SD*-ratio.

		***n*-ratio**		
		1	**>1**	**<1**
***SD*-ratio**	**1**	a	b	b
	>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 at https://github.com/mdelacre/Effect-sizes/.

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 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 20 and 21 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).[21] 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 should always be avoided when the equality of variance assumption is met. Hedges' g, Hedges' g^* and Shieh's g perform equally well as long as the sample size ratio is close to 1 (condition a; see Figure 20). However, when designs are highly unbalanced (condition b; see Figure 21), Shieh's g is not consistent anymore, while Hedges' g^* remains consistent, Hedges's g is a better estimator. For interested readers, these findings are detailed in the three paragraphs below.

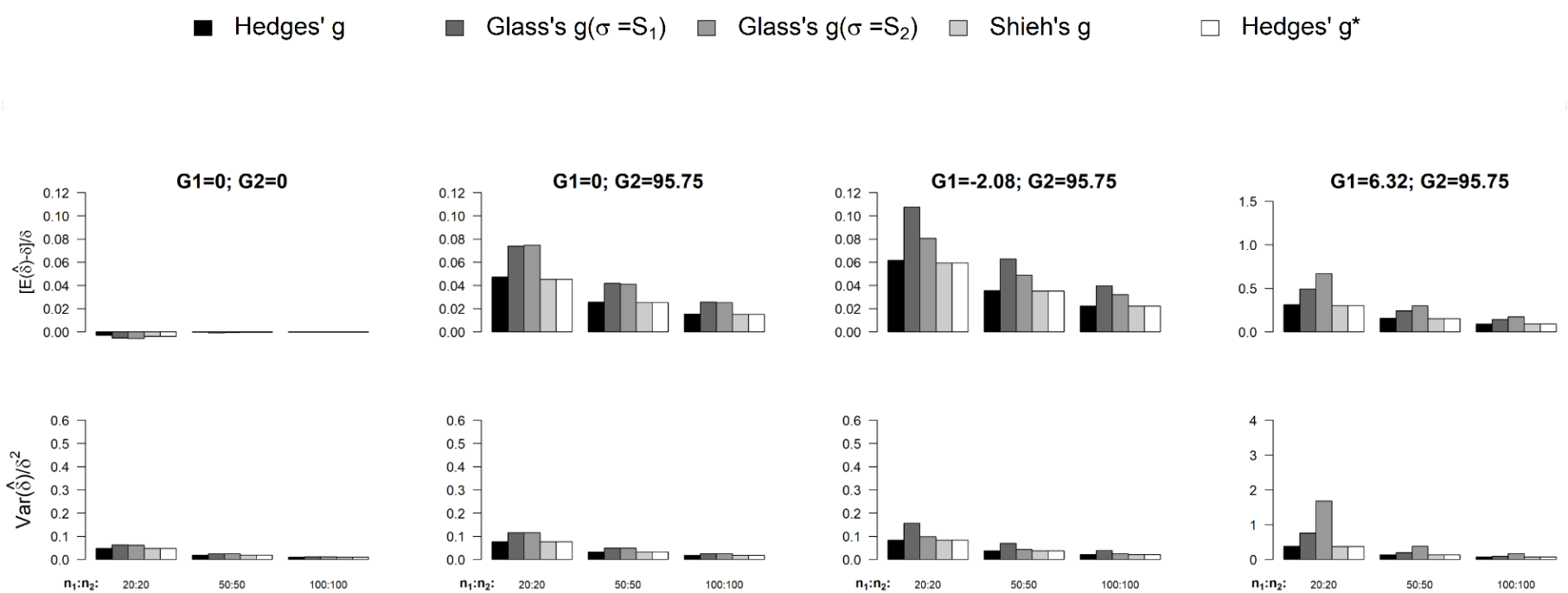

FIGURE 20 – Bias and efficiency of estimators of standardized mean difference, when variances and sample sizes are equal across groups (condition a)

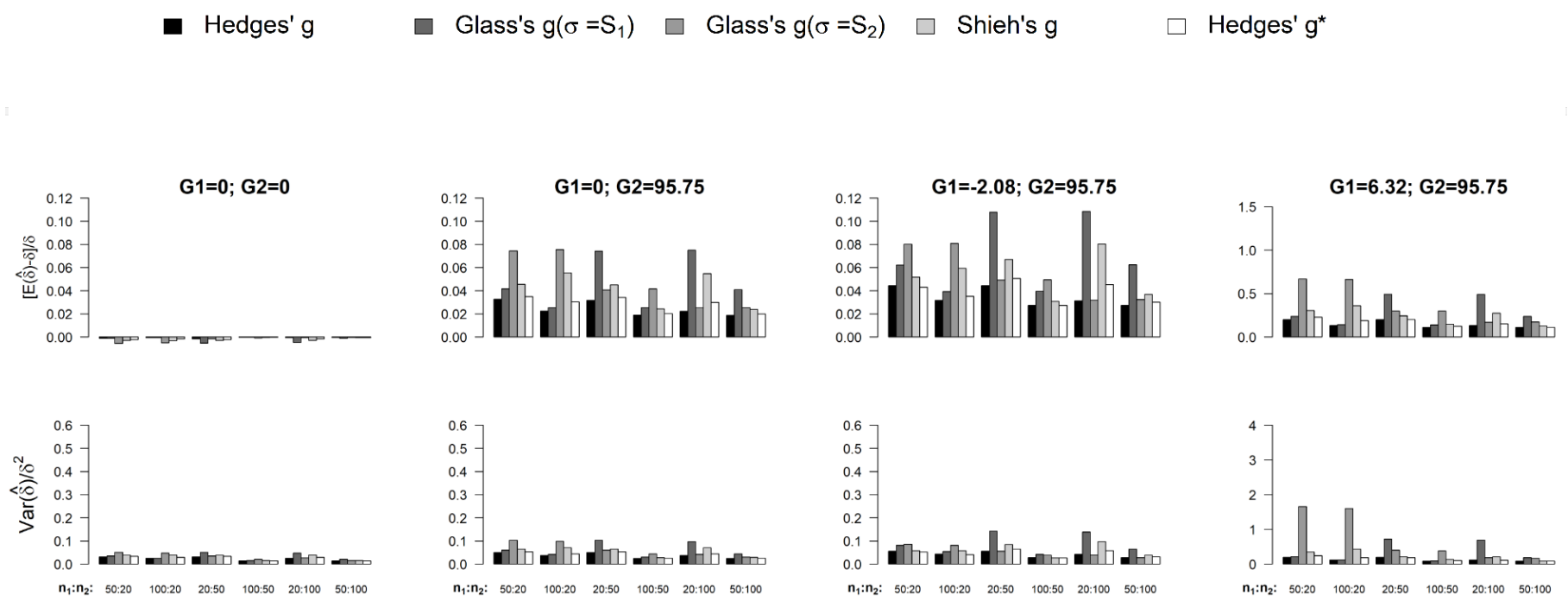

FIGURE 21 – Bias and efficiency of estimators of standardized mean difference, when variances are equal across groups and sample sizes are unequal (condition b)

Figure 20 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, Shieh's g and Hedges' g^* are similar[22] and smaller than the bias and variance of Glass's g estimates using either S_1 or S_2 as a standardizer. Moreover, when samples are extracted from skewed distributions, Glass's g 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 21 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 shows the smallest bias and variance. Hedges' g and Hedges' g^* are consistent estimators (i.e. the larger the sample sizes, the lower the bias and the variance), unlike Shieh's g and Glass's g. The bias of Glass's g 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 is therefore to add subjects in the control group. On the other hand, the variance of Glass's g 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 relative size of the experimental group decreases so does the variance, even when the total sample size is decreased. Regarding Shieh's g, 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.[23] 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 21), for a constant total sample size, Glass's g, Shieh's g and Hedges' g^* 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 are sometimes a bit smaller and sometimes much larger than the bias and variance of Shieh's g and Cohen's d^*.[24]

When variances are unequal across groups

Figures 22 to 27 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.[25] It might be considered surprising that the bias of Hedges' g remains very small throughout these conditions. As discussed in the section "Different measures of effect size", Hedges' g 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 will not be taken into account in the following comparisons.

We will observe that when variances are unequal across populations, Glass's g sometimes performs better, but also sometimes performs much worst than Shieh's g and Hedges' g^*, both in terms of bias and variance. The performance of Glass's g 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 and Hedges' g^* are both appropriate but the further the sample sizes ratio is from 1, the larger the bias of Shieh's g in order that in the end, the measure that we believe performs best across scenarios is Hedges' g^*.

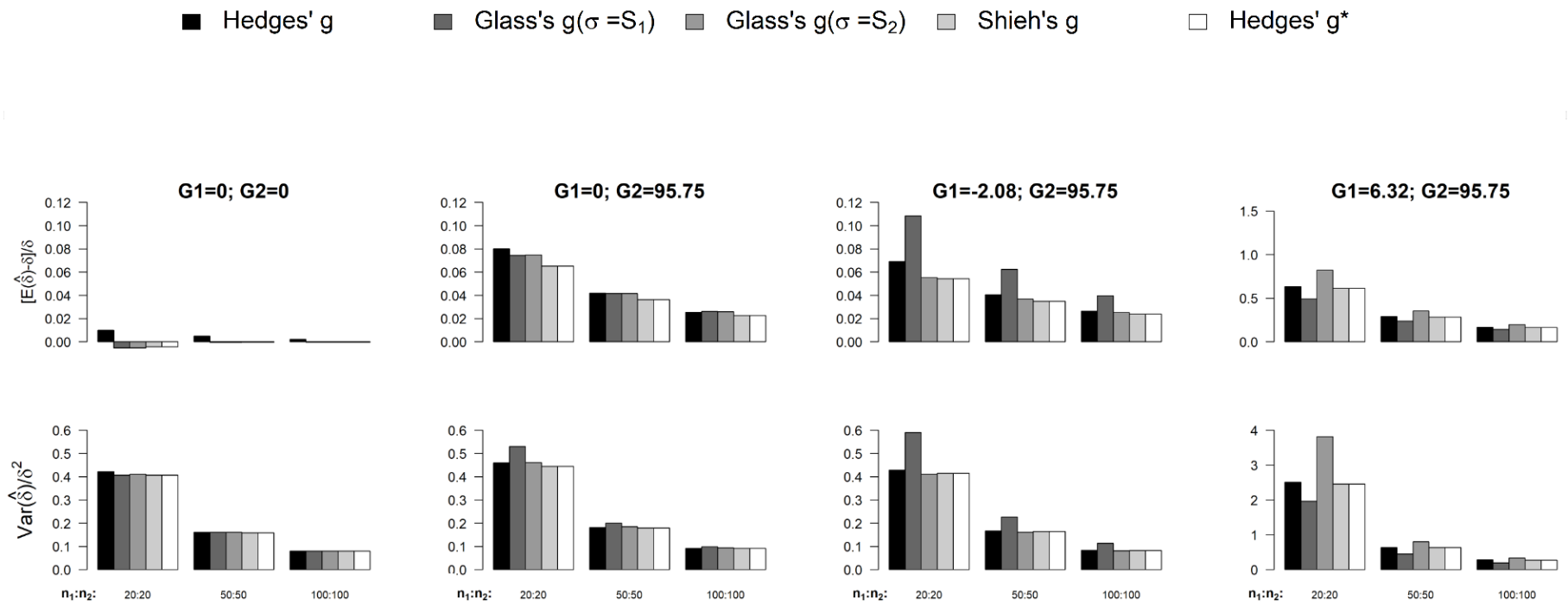

FIGURE 22 – Bias and efficiency of estimators of standardized mean difference, when variances are unequal across groups and sample sizes are equal (condition c), as a function of sample sizes

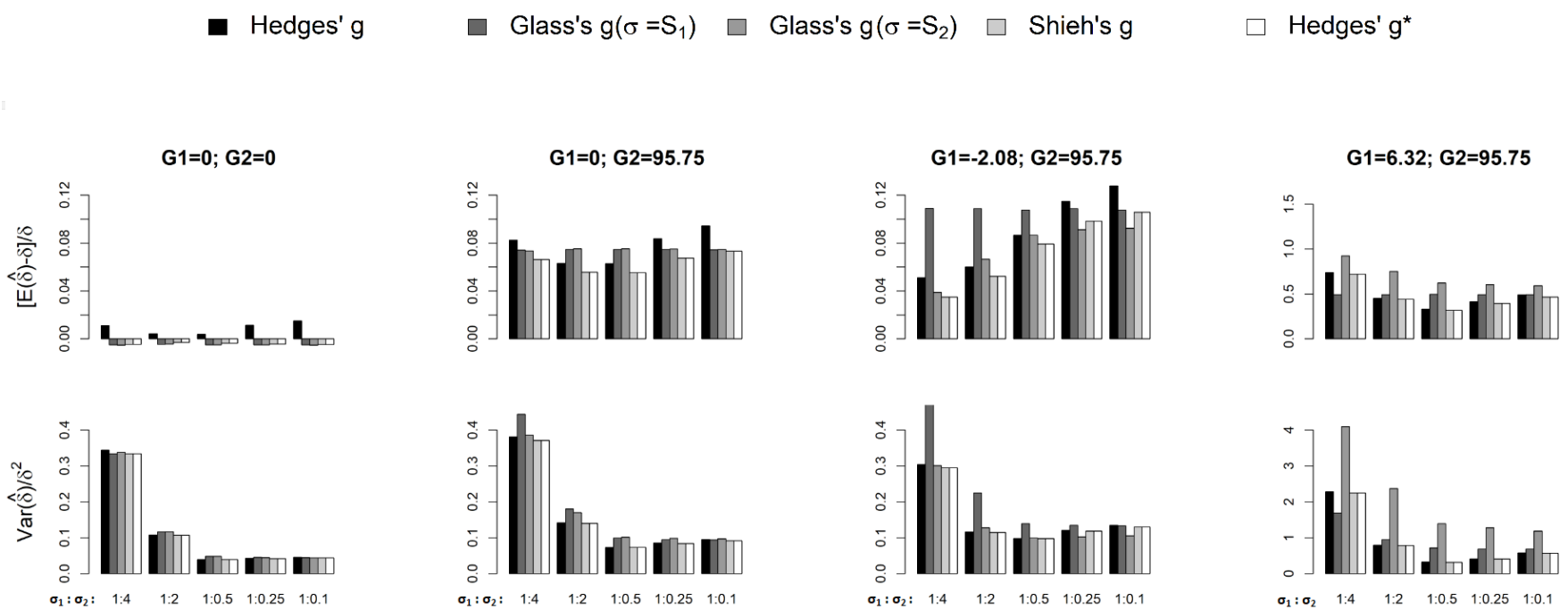

FIGURE 23 – Bias and efficiency of estimators of standardized mean difference, when variances are unequal across groups and sample sizes are equal (condition c) as a function of the SD-ratio (when $n_1 = n_2 = 20$)

Figures 22 and 23 are dedicated to scenarios where population variances are unequal between groups and sample sizes are equal (condition c). In Figure 22, 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 23, 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 23, the relative variance seems to be much larger when $S_2 > S_1$.[26] 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.[27]

When samples are extracted from skewed distributions, the bias and variance of Glass's g are sometimes smaller and sometimes larger than the bias of Shieh's g and Hedges' g^*. 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 23), 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.[28]

Figures 24 to 27 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 4 revealed an interaction effect between the sample sizes ratio and the SD-ratio on the bias and variance of Hedges' g^* and Shieh's g (see Supplemental Material 1), we chose to present all configurations where the larger SD is 10 times larger than the smaller SD (Figures 24 and 25), and configurations where the larger SD is twice larger than the smaller SD (Figures 26 and 27), 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 ($\frac{\sigma_2}{\sigma_1} = 10$ or .1) or medium ($\frac{\sigma_2}{\sigma_1} = 2$ or .5).

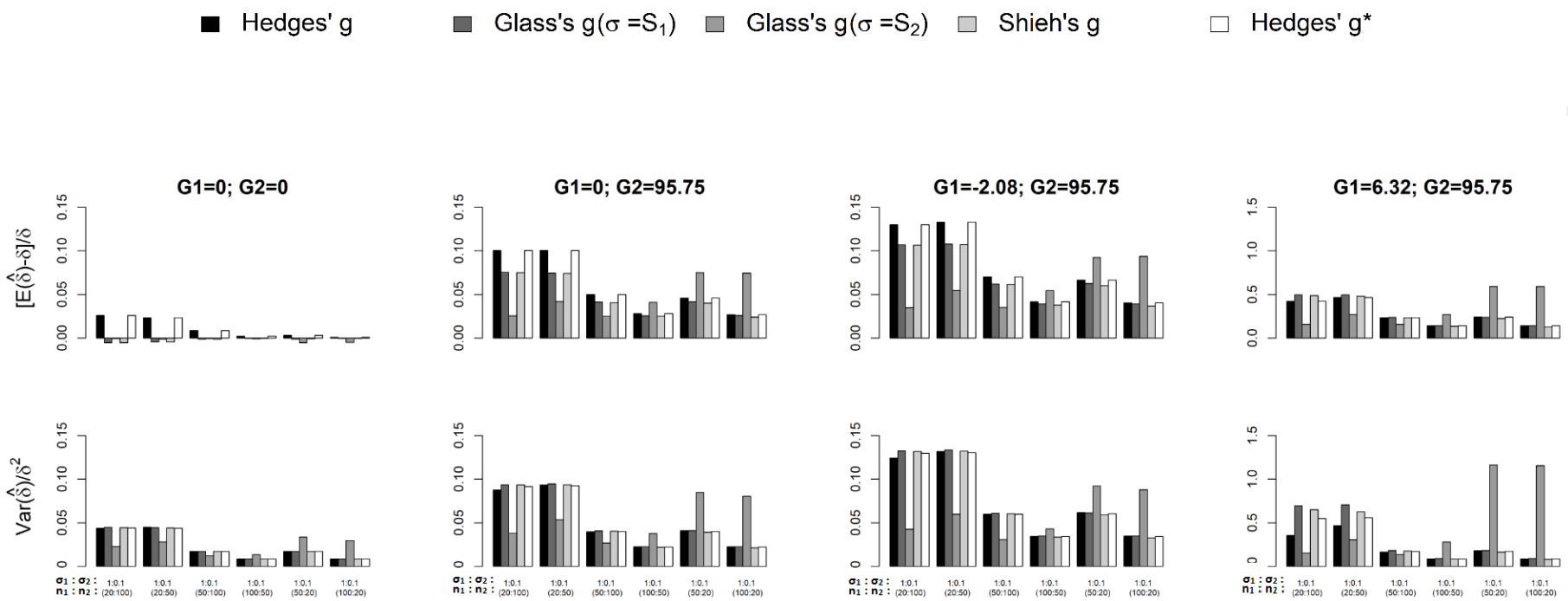

FIGURE 24 – Bias and efficiency of estimators of standardized mean difference, when variances and sample sizes are unequal across groups (condition d), and σ_1 is 10 times larger than σ_2

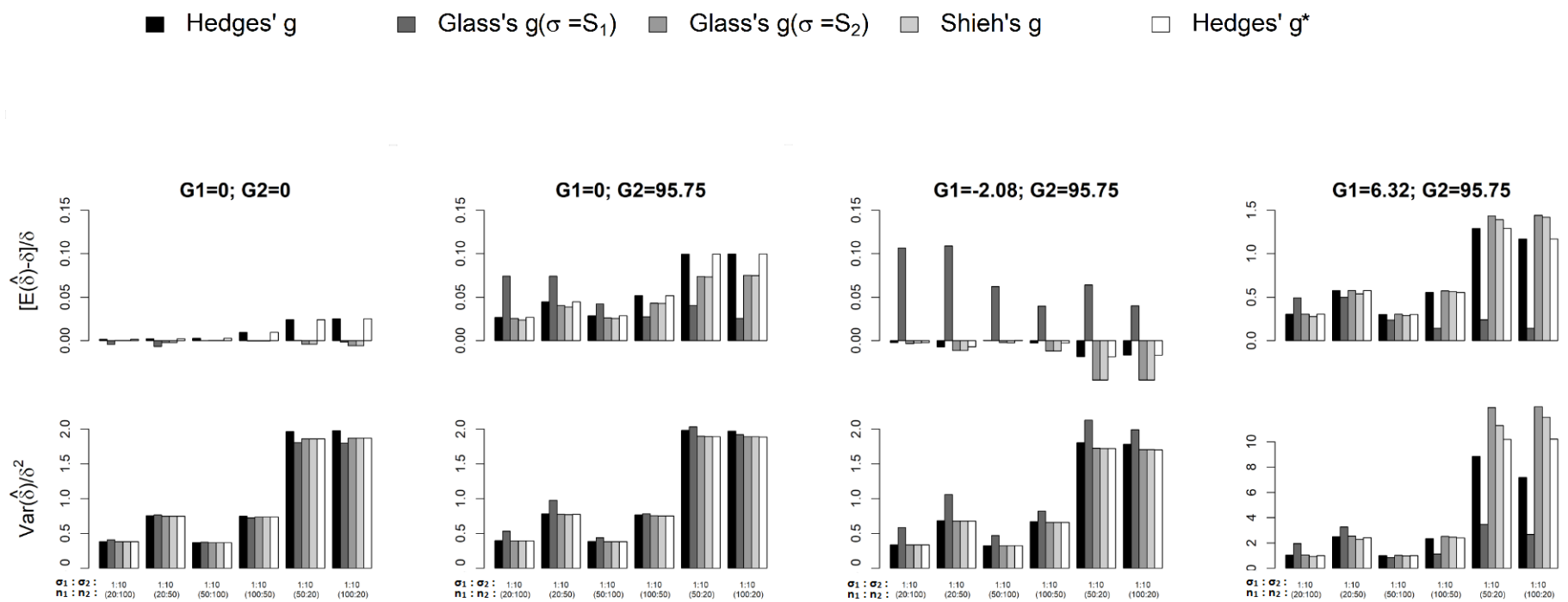

FIGURE 25 – Bias and efficiency of estimators of standardized mean difference, when variances and sample sizes are unequal across groups (condition d), and σ_2 is 10 times larger than σ_1

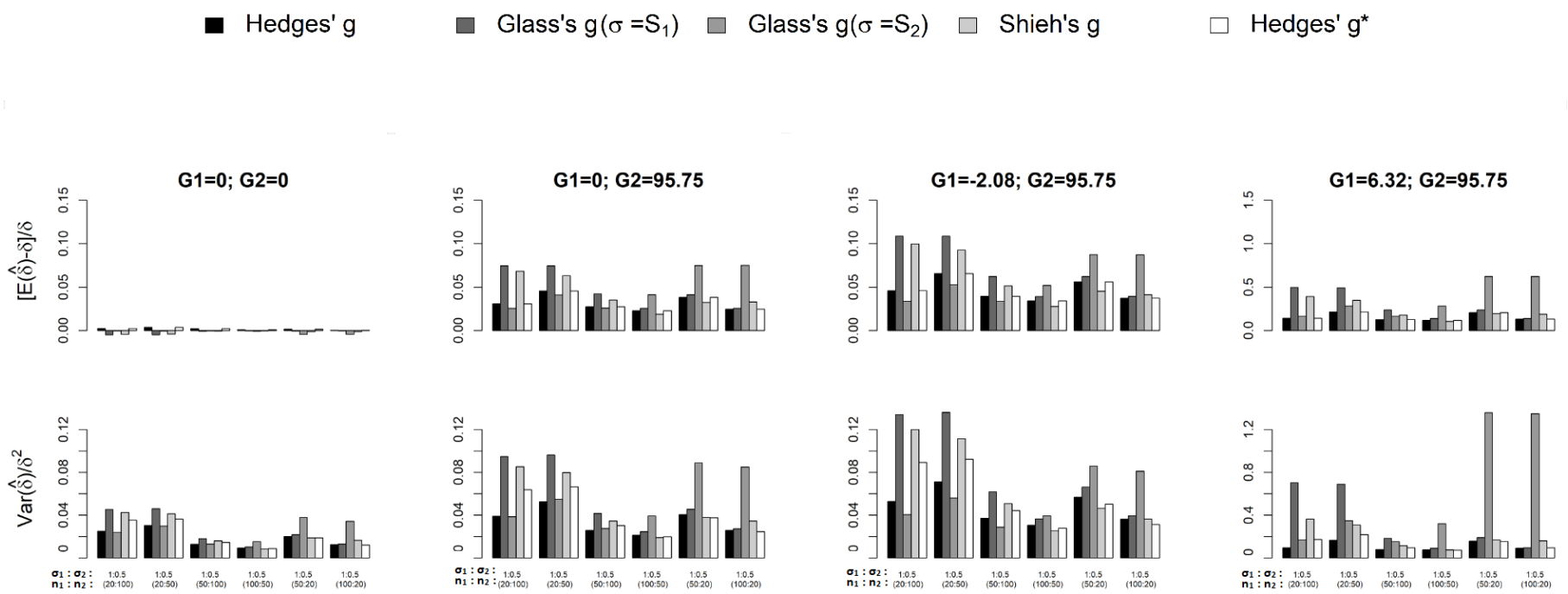

FIGURE 26 – Bias and efficiency of estimators of standardized mean difference, when variances and sample sizes are unequal across groups (condition d), and σ_1 is twice larger than σ_2

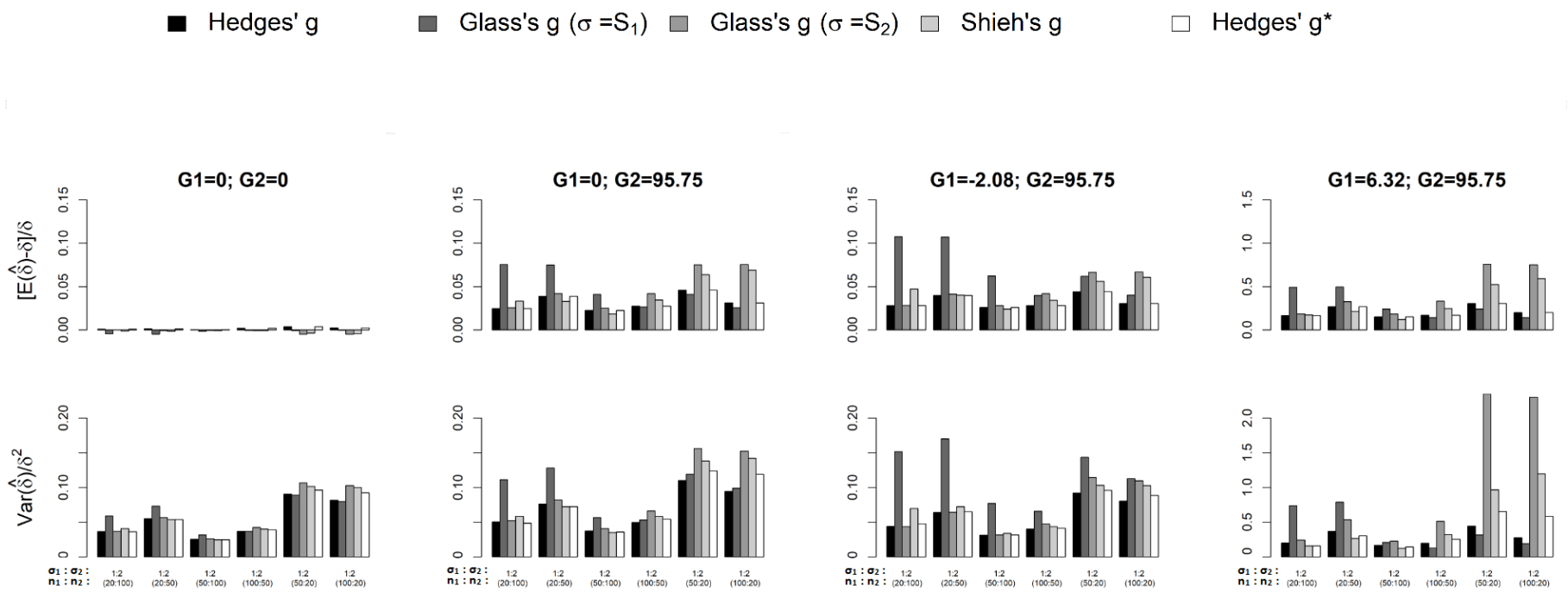

FIGURE 27 – Bias and efficiency of estimators of standardized mean difference, when variances and sample sizes are unequal across groups (condition d), and σ_2 is twice larger than σ_1

When distributions are symmetric, the bias of Glass's g 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 24 to 27, one can also notice that the bias of Glass's g does not depend on the SD-ratio either. Unlike the bias of Glass's g, 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, to add subjects in the control group. Regarding Hedges' g^* and Shieh's g, their respective biases and variances depend on an interaction effect between the sample sizes ratio and the SD-ratio $\left(\frac{n_2}{n_1} \times \frac{\sigma_2}{\sigma_1}\right)$: 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 24 and 25) than when it is only twice more variable (Figures 26 and 27). However, the respective biases and variances of Hedges' g^* and Shieh's g 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 are sometimes smaller and sometimes larger than the bias of Shieh's g and Hedges' g^*, 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^* 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), b) it is always consistent (unlike Shieh's g), c) it is easy to interpret (Hedges' g^* can be interpreted in the same way as Hedges' g) 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^* should always be reported with a confidence interval. To help researchers compute Hedges' g^* and its confidence interval we created the R package *deffectsize* (see https://github.com/mdelacre/deffectsize). 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

Hedges' g^* is computed but it is also possible to compute biased estimators (e.g. Cohen's d^*) and/or to use a pooled error term as standardizer by assuming that the equality of population variances is met (e.g. Hedges' g or Cohen's d, 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 (or Shieh's d) and Glass's g (or Glass's d) 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/.

Endnotes

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

[16]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.

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

[18]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.

[19]In Table 2, one can see that the variance of Hedges' g equals the variance of Cohen's d, 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 and Hedges' g.

[20]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.

[21]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.

[22]While the bias and variance of Cohen's d, Cohen's d^* and Shieh's d are identical, the bias and variance of Hedges' g are marginally different from the bias and variance of Hedges' g^* and Shieh's g (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 does not contain the sample variances (unlike the correction applied on both other estimators), the bias and variance of Hedges' g are slighly different from the bias and variance of Hedges' g^* and Shieh's g.

[23]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 21 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.

[24]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 and other estimators is larger when Glass's g is the least efficient estimator.

[25]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.

[26]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.

[27]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.

[28]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 et al. (2001), par exemple, ont rapporté que parmi 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'hypothèse nulle. Plus récemment, Lakens (2017) a noté que les mots "no effect" ont été utiliséé dans 108 articles publiés 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 et 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 (2011), comme le soulignent 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 et al. (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 7 résume les résultats de simulations Monte Carlo pour un ensemble de 42 scénarios de comparaison des moyennes de 2 échantillons indépendants. En fonction du scénario, nous faisons varier la taille des échantillons (n_j) ainsi que la différence entre les moyennes de population dont sont extraits les échantillons ($\mu_1 - \mu_2$).

Bien que comme nous l'avons longuement expliqué dans les chapitres antérieurs, la condition d'homogénéité des variances est peu fréquemment rencontrée en situation réelle, nous avons choisi de générer systématiquement deux échantillons extraits de populations aux variances identiques, afin de pouvoir simplifier la formule de l'erreur standard de la différence de moyennes qui sera ultérieurement mentionnée. Nous nous sommes permis ce choix puisque le message clé de l'illustration ne dépend nullement du ratio entre les variances de population.

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[29] et extrait la p-valeur du test. Ensuite, nous avons calculé la proportion d'itérations associées à une p-valeur supérieure 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 7, à l'exception de la première), cette proportion correspond au taux d'erreur de type II (communément appelé β).

Table 7.

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 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 que l'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[30] et si cette estimation atteint 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 8, dans laquelle nous envisageons les mêmes scénarios que dans la Table 7 et ajoutons une contrainte de puissance : nous décidons qu'on ne peut conclure à l'équivalence que si l'on atteint une puissance de 80% pour détecter une différence de moyennes 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'atteint jamais la puissance minimale de 80% pour détecter une différence de moyennes de .3.[31] Par contre, une fois les échantillons assez grands pour s'assurer cette puissance, lorsque la différence entre les moyennes de population 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.[32]

Table 8.

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 grande (Meyners, 2012; Schuirmann, 1987), plus la probabilité de conclure à l'équivalence augmentera. Ce dernier point est illustré au sein de la Figure 28, dans le contexte de la comparaison de deux moyennes. Sur l'axe des abscisses, on représente différentes estimations de la différence de moyennes ($\bar{X}_1 - \bar{X}_2$) 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).[33] 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 28, cette propriété peu désirable n'est pas partagée par le TOST, un test d'équivalence que nous allons décrire ci-après (Schuirmann, 1987).

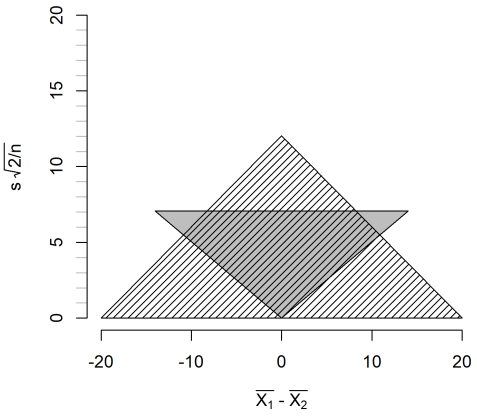

FIGURE 28 – Région d'équivalence pour l'approche de la puissance (zone grisée) et pour le TOST (zone hachurée), pour l'exemple où $|\theta|$=20, n = 50 et $\alpha = .05$

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 et al., 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 et al., 2018). Il existe ensuite plusieurs approches pour démontrer que l'effet observé se situe dans la zone d'équivalence (voir Meyners, 2012). Parmis celles-ci, une approche très simple est celle du "Two one-sided tests" (Lakens, 2017; Schuirmann, 1987), que l'on appelle plus communément le TOST.[34] 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 seconde 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 nulles peuvent être simultanément rejetées, on peut conclure à l'équivalence (Seaman & Serlin, 1998). Cela équivaut, statistiquement parlant, à montrer que l'intervalle de confiance autour de θ 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 résultats 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 et al., 2018; Rogers et al.,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 et al., 1993).

Notons également qu'il n'est pas nécessaire de procéder à une correction du risque alpha due 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.[35] 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 29 dans le contexte de la comparaison de deux moyennes indépendantes :

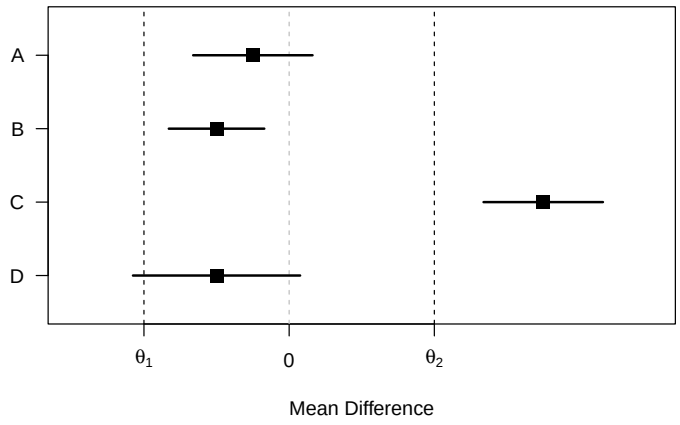

FIGURE 29 – Différence de moyennes $(\bar{X}_1 - \bar{X}_2)$ et IC à $(1 - 2\alpha)\%$ autour de la différence de moyennes $(\bar{X}_1 - \bar{X}_2)$ pour 4 scénarios distincts.

(1) La différence de moyennes observée est simultanément (et significativement) supérieure à θ_1 et inférieure à θ_2. On ne peut par contre montrer d'écart significatif par rapport à 0 (scénario A, Figure 29) : dans ce cas, on conclura à l'absence d'un effet au moins aussi grand que les bornes d'équivalence.

(2) La différence de moyennes observée est simultanément (et significativement) supérieure à θ_1, inférieure à θ_2 et différente de 0 (scénario B, Figure 29) : 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 et al., 1993).

(3) La différence de moyennes observée diffère significativement de 0, par contre, on ne peut démontrer qu'elle est significativement supérieure à θ_1, significativement inférieure à θ_2 ou ces deux affirmations simultanément (autrement dit, on ne peut rejeter au moins une des deux hypothèses unilatérales contre les bornes de la zone d'équivalence; scénario C, Figure 29) : on conclura alors à la présence d'un effet non nul (Rogers et al., 1993).

(4) Non seulement, on ne peut démontrer que la différence de moyennes observée diffère significativement de 0 mais en plus, on ne peut conclure au rejet d'au moins une des deux hypothèses unilatérales contre les bornes de la zone d'équivalence (scénario D, Figure 29): 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 et al., 2018). Dans ce cas, établir l'équivalence revient à rejeter la présence d'un effet ayant un quelconque intérêt pratique (Rogers et al., 1993). Par exemple, Burriss et al. (2015) avaient émis l'hypothèse qu'une augmentation de la rougeur de la peau chez les femmes, en période d'ovulation, les rendraient plus attractives pour les hommes. 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 et al., 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 de 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 leurs réponses à la différence de scores obtenus à l'aide du BDI[36] dans ce même laps de temps (Lakens et al., 2018). Cela leur a permis de déterminer ce qu'ils considéraient 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 et al., 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'occurrence, 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 quantitatives de la littérature. Gignac et 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 et Ozer (2019).[37] 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 et al. (2018) consiste à définir comme borne d'équivalence le plus petit effet que l'étude d'origine aurait pu détecter comme étant significative.[38] Voici l'idée sous-jacente : idéalement, les chercheurs devraient toujours spécifier ce qu'ils considèrent comme étant le plus petit effet d'intérêt mais 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 qu'il 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 chercheur 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, il 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'il 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 et al. (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.[39]

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 et al., 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, le résultat 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 et al., 1993; Serlin & Lapsley, 1985; Spiegelhalter et al., 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, D'Agostino McGowan, Dupont, & Greevy, 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 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 et al., 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 30 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 30 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.

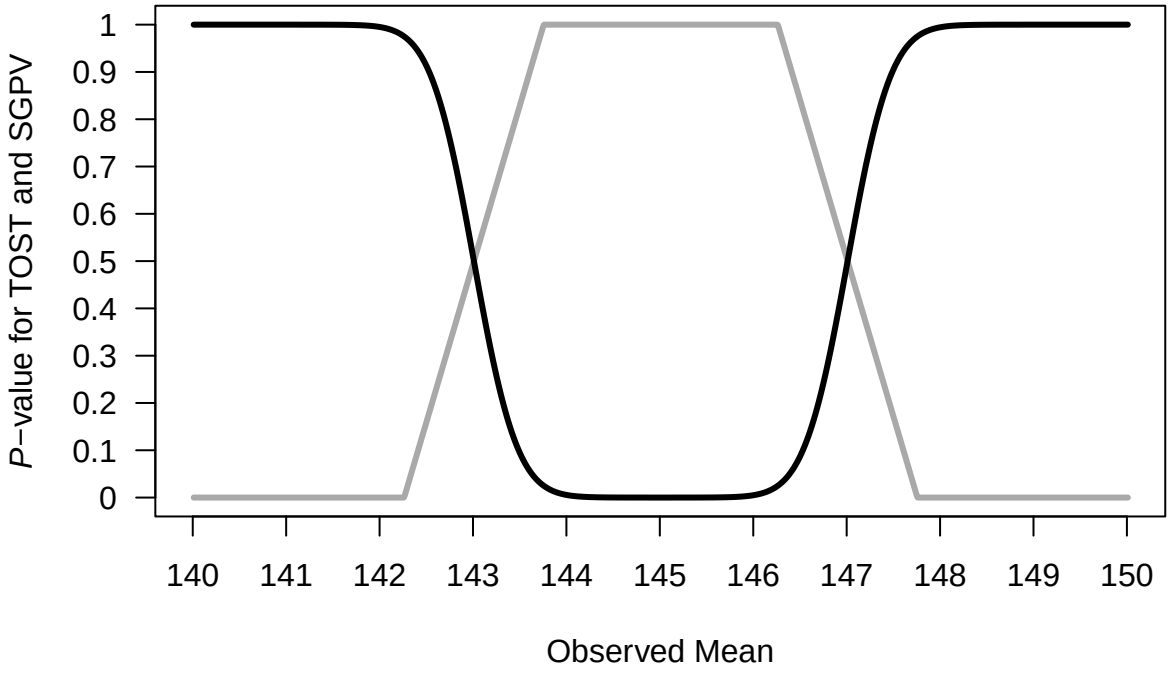

FIGURE 30 – Comparison of p-values from TOST (black line) and SGPV (grey line) across a range of observed sample means (x-axis) tested against a mean of 145 in a one-sample t-test with a sample size of 30 and a standard deviation of 2, illustrating that when the TOST p-value = 0.5, the SGPV = 0.5, when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

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 31) 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 = 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 31 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 30). 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 30).

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 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

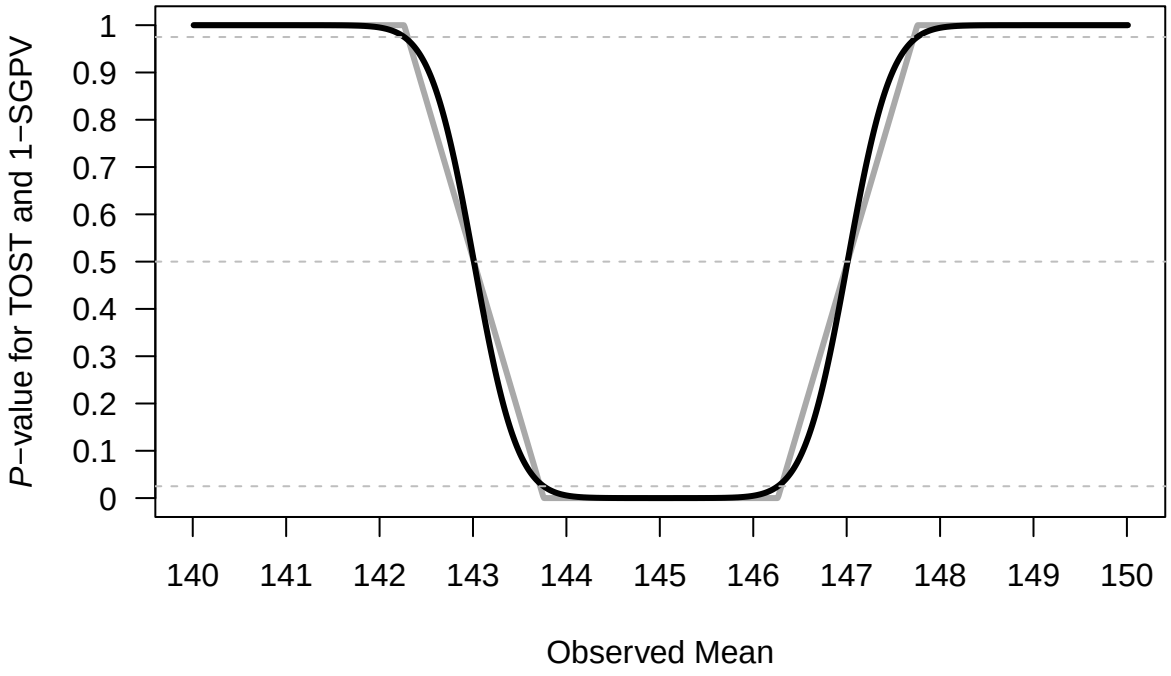

FIGURE 31 – Comparison of p-values from TOST (black line) and 1-SGPV (grey line) across a range of observed sample means (x-axis) tested against a mean of 145 in a one-sample t-test with a sample size of 30 and a standard deviation of 2.

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 32).

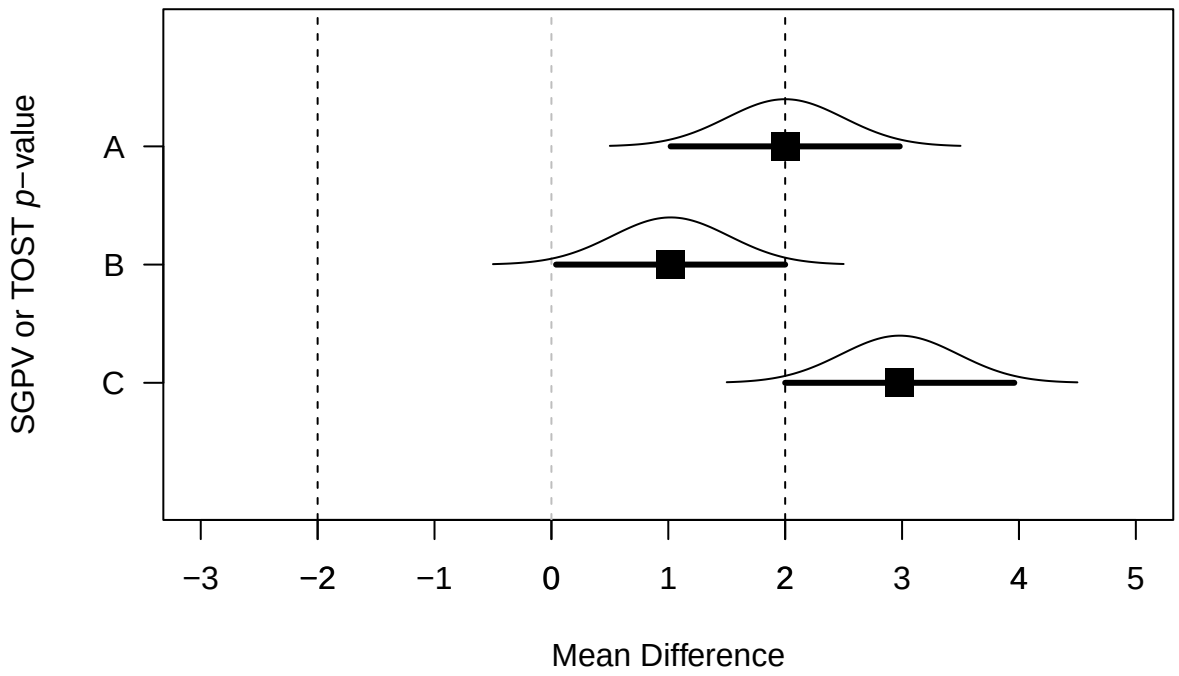

FIGURE 32 – Means, normal distribution, and 95% CI for three example datasets that illustrate the relationship between p-values from TOST and SGPV.

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 32) 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 32), 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 et al.,

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, Scheel, & Isager, 2018).

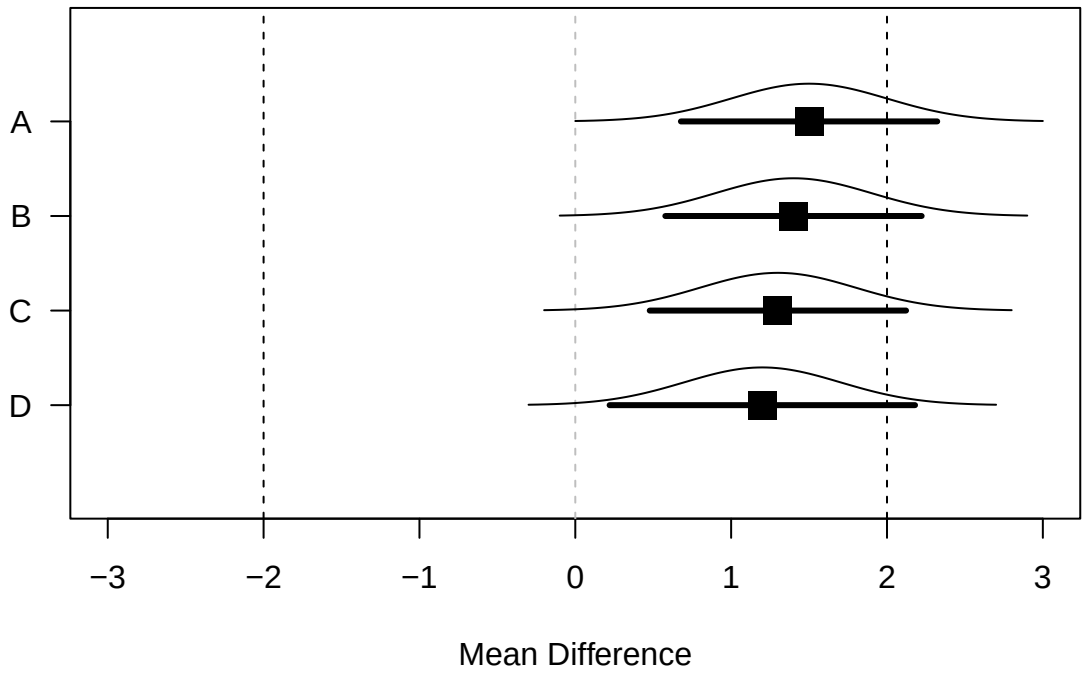

FIGURE 33 – Means, normal distribution, and 95% CI for samples where the observed population mean is 1.5, 1.4, 1.3, and 1.2.

For example, in Figure 33 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 (0.0436) is not the same as the difference between C and D (0.026). 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 30 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 30. When 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

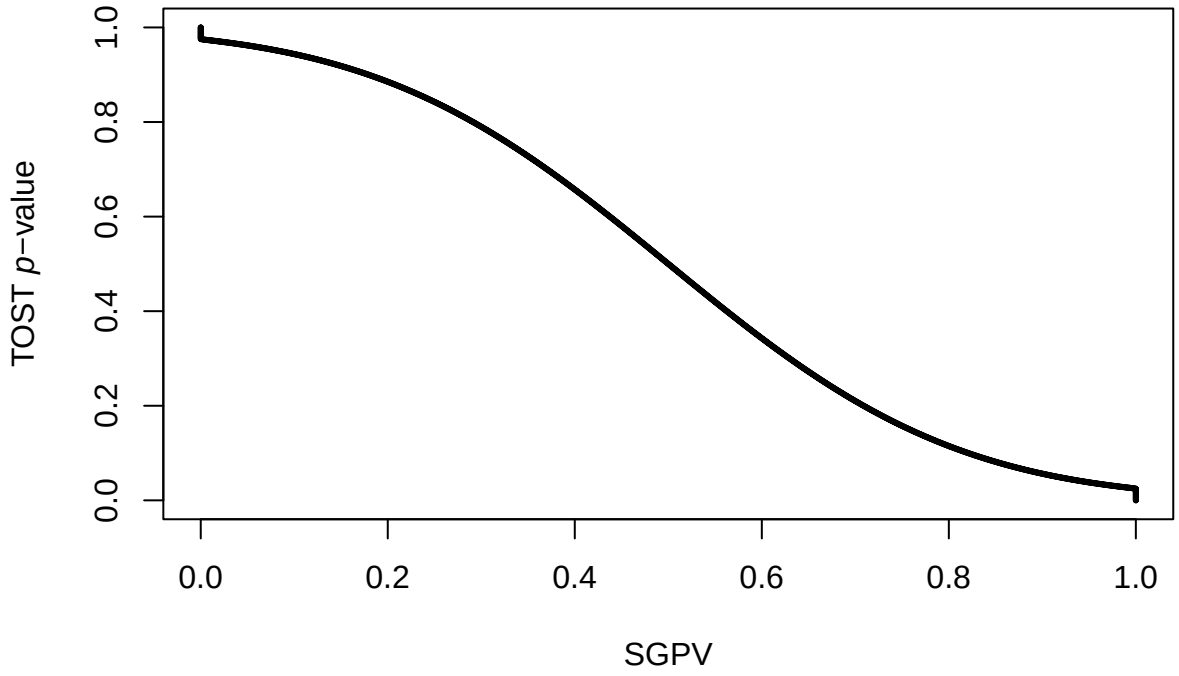

FIGURE 34 – The relationship between p-values from the TOST procedure and the SGPV for the same scenario as in Figure 30.

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.

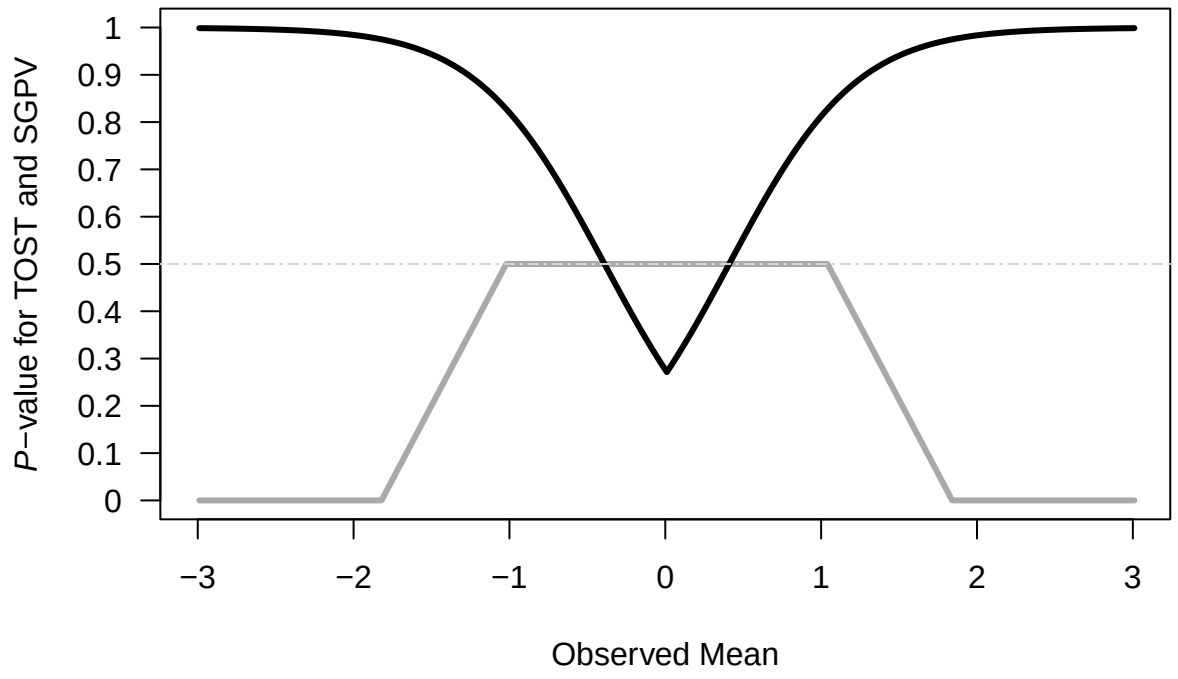

FIGURE 35 – Comparison of p-values from TOST (black line) and SGPV (grey line) across a range of observed sample means (x-axis). Because the sample size is small ($n = 10$) and with a standard deviation of 2 the CI is more than twice as wide as the equivalence range (set to -0.4 to 0.4), the SGPV is set to 0.5 (horizontal lightgrey line) across a range of observed means.

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 35). 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 36). 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.

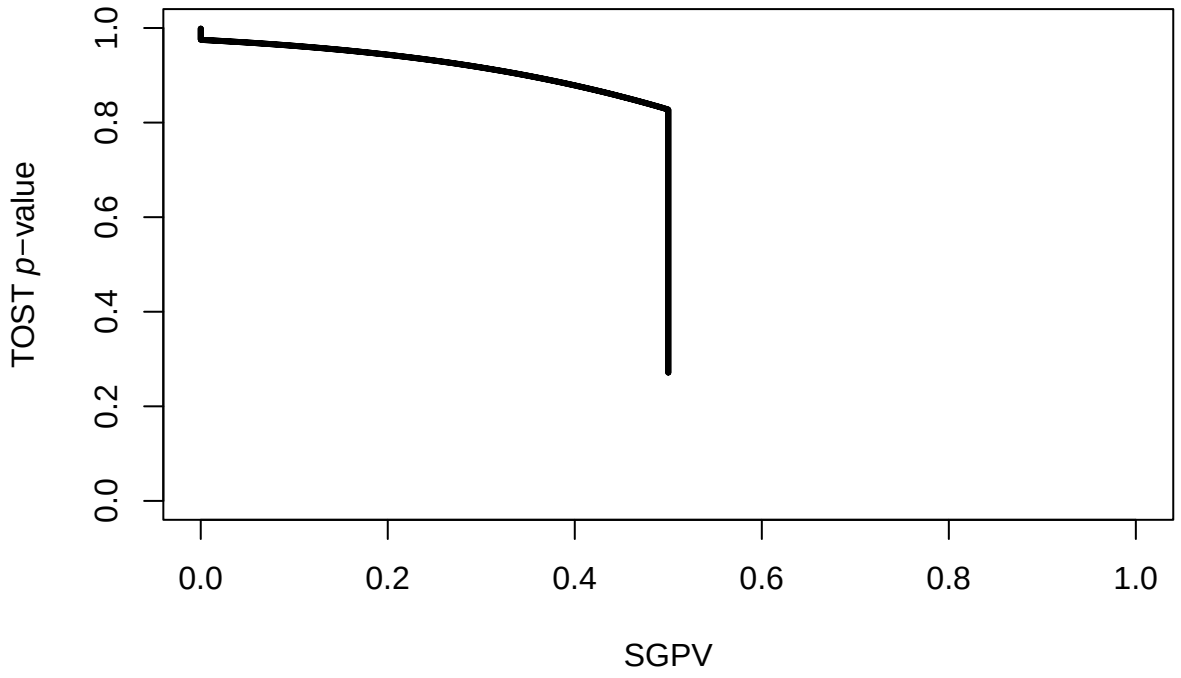

FIGURE 36 – The relationship between p-values from the TOST procedure and the SGPV for the same scenario as in Figure 35.

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 37). 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.

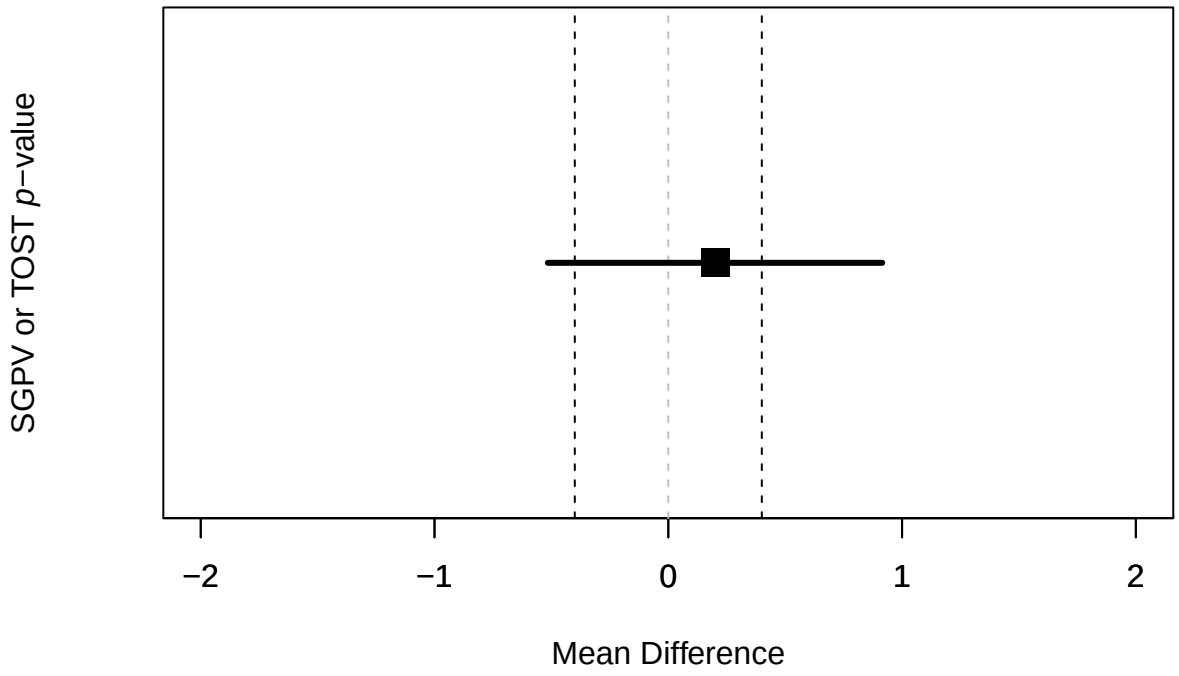

FIGURE 37 – Example of a 95% CI that overlaps with the lower and upper equivalence bound (indicated by the vertical dotted lines).

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 38 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.

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 39).

To conclude this section, there are situations where the p-value from the TOST procedure

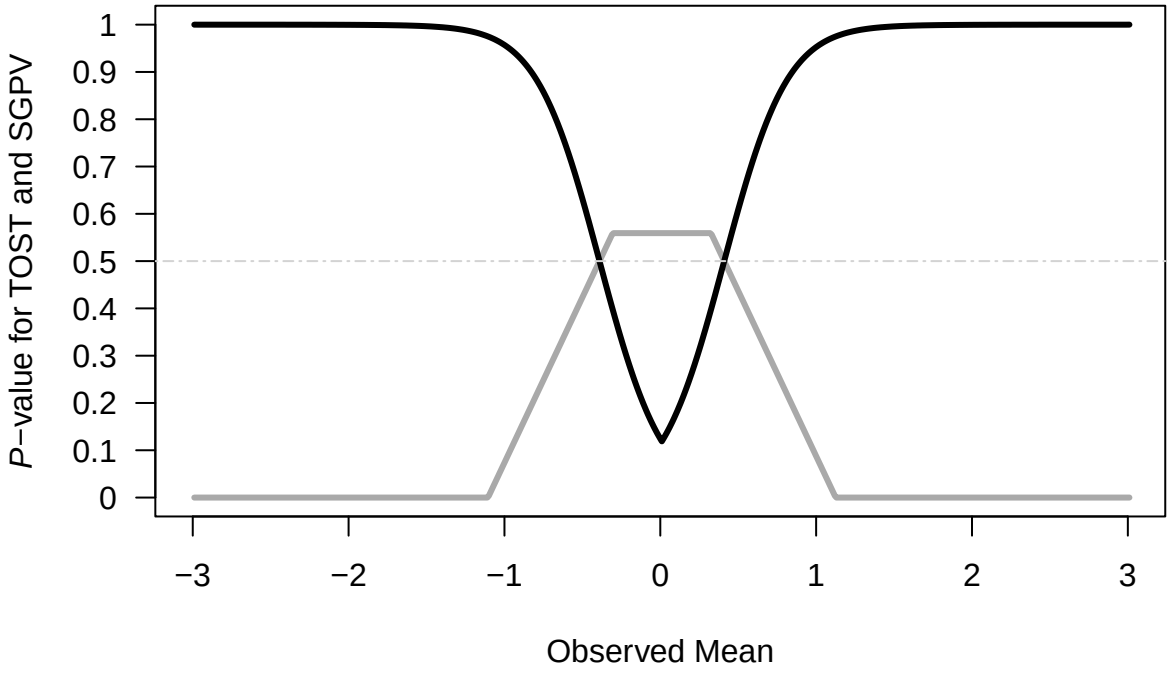

FIGURE 38 – Comparison of p-values from TOST (black line) and SGPV (grey line) across a range of observed sample means (x-axis). The sample size is small ($n = 10$), but because the sd is half as big as in Figure 36 (1 instead of 2) the CI is less than twice as wide as the equivalence range (set to -0.4 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.

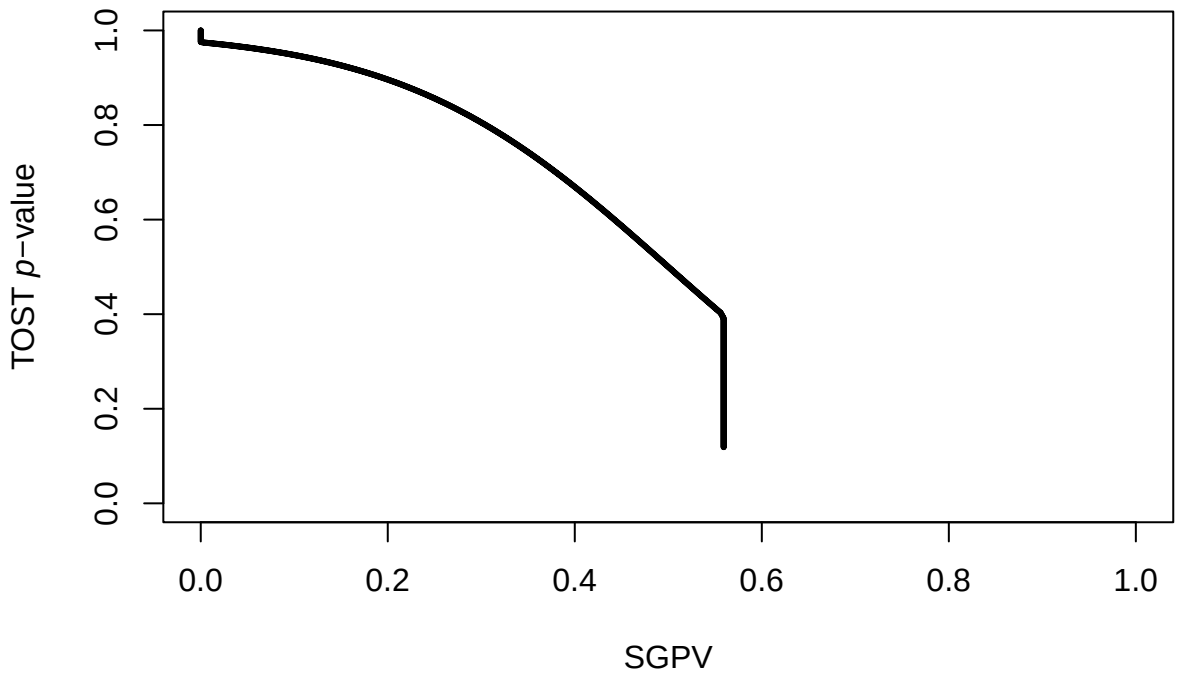

FIGURE 39 – The relationship between p-values from the TOST procedure and the SGPV for the same scenario as in Figure 38.

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 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 40 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 31. 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,

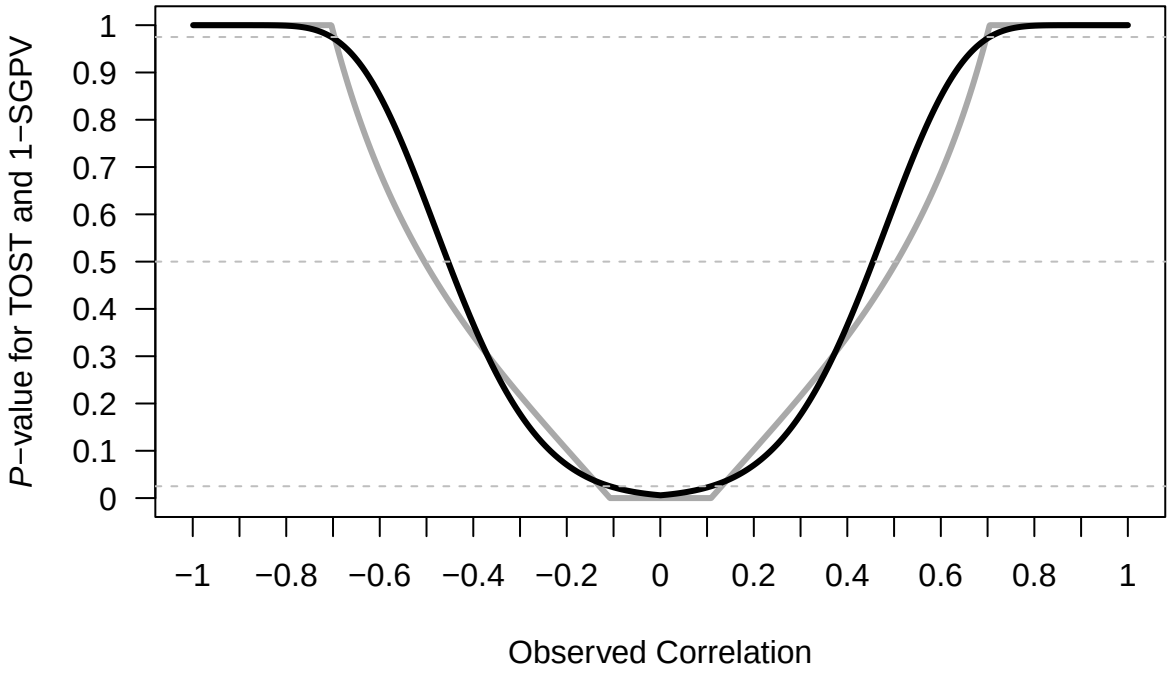

FIGURE 40 – Comparison of p-values from TOST (black line) and 1-SGPV (grey curve) across a range of observed sample correlations (x-axis) tested against equivalence bounds of r = -0.45 and r = 0.45 with n = 30 and an alpha of 0.05.

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 than the true correlation. As can be seen in Figure 41, 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 0.7) according to Blume et al. (2018) 58.106% 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) 97.6% of the confidence interval overlaps with the equivalence range, even though in the long run only 35% 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

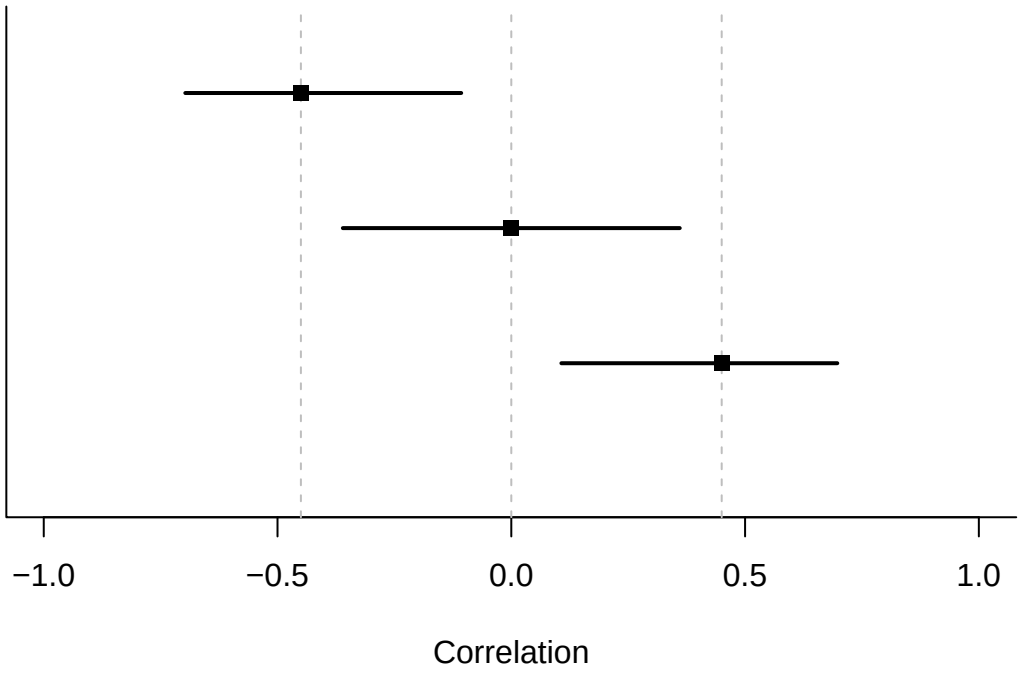

FIGURE 41 – Three 95% confidence intervals for observed effect sizes of r = -0.45, r = 0, and r = 0.45 for n = 30. Only the confidence interval for r = 0 is symmetric.

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 35). An important benefit of equivalence tests is that it does not need such a correction to prevent misleading results.

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 42. 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

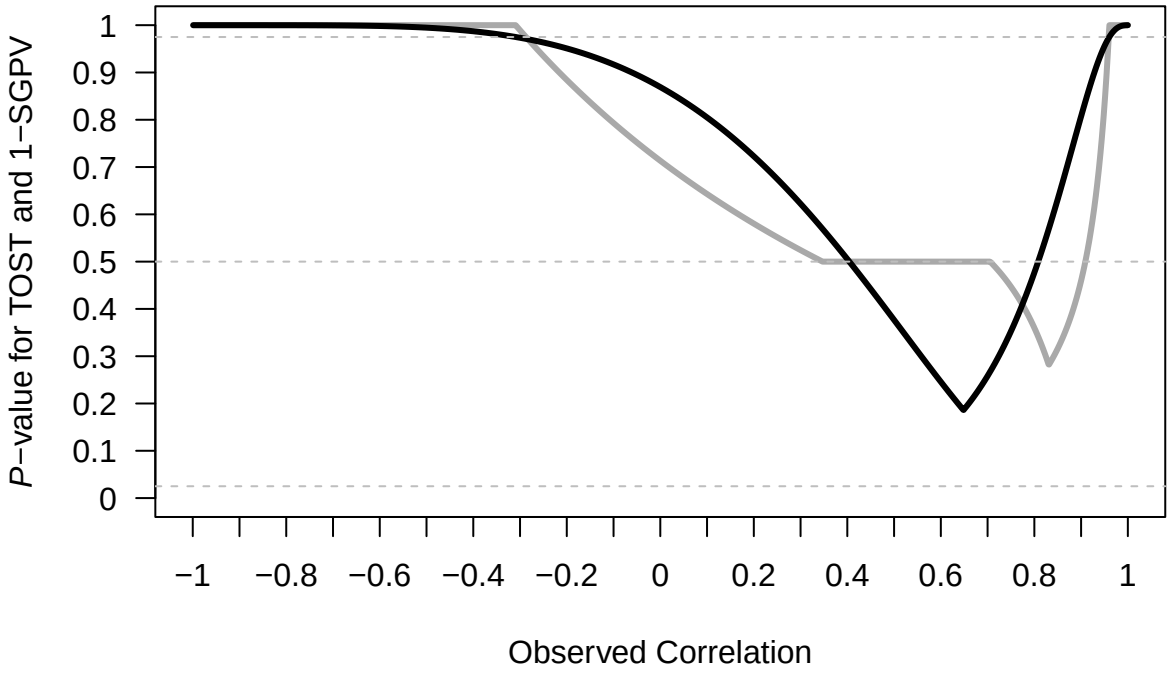

FIGURE 42 – Comparison of p-values from TOST (black line) and 1-SGPV (grey curve) across a range of observed sample correlations (x-axis) tested against equivalence bounds of $r = 0.4$ and $r = 0.8$ with $n = 10$ and an alpha of 0.05.

from a continuous 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.

Notes de fin de chapitre

[29]Lorsque la condition d'homogénéité des variances est respectée, le test de Student est le test de comparaison de 2 moyennes indépendantes le plus robuste, tant en termes d'erreur de type I qu'en termes d'erreur de type II. A nouveau, si nous avions choisi d'utiliser le test t de Welch, cela n'aurait en rien altéré le message-clé de l'illustration

[30]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 connaît pas et devra donc estimer sur base de S, l'écart-type de l'échantillon (Schuirmann,1987).

[31]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.

[32]En comparant les Tables 7 et 8, 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 8, sont inférieures aux proportions d'itérations de chaque scénario qui amènent à ne pas rejeter l'hypothèse nulle, dans la Table 7. 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.

[33]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.

[34]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).

[35]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.

[36]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).

[37]Funder et Ozer (2019) ont relevé plusieurs enquêtes ayant calculé un r de Pearson moyen de .21 sur base d'effets publiés dans la littérature 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.

[38]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 borne d'équivalence au singulier.

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

Chapitre 6 : Discussion et conclusion générale

Objectifs de départ, résumé et apports de la thèse

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 et (3) de proposer des recommandations pour les améliorer.

Dans un premier temps, nous nous sommes focalisés sur l'usage des tests t de Student et F de Fisher, soit des tests communément utilisés par les chercheurs en psychologie, en vue de comparer les moyennes de deux ou plusieurs groupes de sujets indépendants, et qui reposent sur les conditions que des résidus, indépendants et identiquement distribués, soient extraits d'une distribution normale et que les variances des populations dont sont extraits chaque groupe soient identiques (soit la condition d'homogénéité des variances).

Bien que les enjeux des conditions statistiques de ces tests aient déjà été largement explorés par le passé, ils semblaient toujours largement ignorés par de nombreux chercheurs appliqués. Notre principale motivation à aborder cette thématique était dès lors d'ordre pédagogique : il nous semblait nécessaire de combler le fossé entre les méthodologistes et la majorité des chercheurs appliqués.

Il nous est d'abord apparu que la littérature manquait d'articles expliquant de manière compréhensible les raisons pour lesquelles les conditions statistiques des tests t de Student et F de Fisher étaient peu réalistes. Nous avons dès lors mis en évidence toute une série d'arguments qui permettent de remettre en cause la crédibilité, dans de nombreux domaines de la psychologie, des conditions statistiques de normalité (comme la présence de sous-populations définies par des facteurs non identifiés dans le plan expérimental, l'étude de mesures bornées, comme le temps qui ne peut prendre des valeurs négatives, ou encore le fait qu'un traitement est susceptible de modifier la forme des distributions étudiées) et d'homogénéité des variances (comme l'étude de groupes préexistants à l'expérience, définis par des variables telles que le genre ou l'origine ethnique[40], ou encore le fait qu'un traitement, qu'il soit expérimental ou quasi expérimental, est susceptible d'agir sur tous les paramètres d'une distribution, incluant

sa variance).

Ensuite, grâce aux avancées informatiques récentes, nous avons pu étendre les travaux déjà engagés par de nombreux auteurs avant nous (voir par exemple Harwell, 1992), en vue de montrer les conséquences réelles de la violation des conditions de normalité et d'homogénéité des variances pour respectivement les tests statistiques t de Student (chapitre 2) et F de Fisher (chapitre 3), dans des conditions qui se veulent les plus réalistes possibles dans le contexte de la recherche en psychologie (en termes d'hétéroscédasticité et d'écarts à la condition de normalité). Nous avons, à cette fin, réalisé des simulations intensives, avec 1,000,000 d'itérations pour un nombre très vaste de scénarios, variant en fonction d'un ensemble de paramètres connus pour jouer un rôle clé sur les taux d'erreur de type I et II des test t de Student et F de Fisher.

Il est ressorti de nos simulations que de manière cohérente avec nos attentes théoriques, lorsque les deux échantillons comparés sont de même taille, le test t de Student est robuste aux violations de la condition d'homogénéité des variances. Par contre, il en est différemment pour des échantillons de tailles différentes : sur le long terme, la probabilité de rejeter l'hypothèse nulle avec ce test est supérieure aux attentes théoriques lorsque le plus petit échantillon est extrait de la population ayant la plus grande variance, et est inférieure aux attentes théoriques lorsque le plus petit échantillon est extrait de la population ayant la plus petite variance.

Dans la mesure où l'ANOVA F de Fisher est une généralisation du test t de Student[41], il n'est pas surprenant que nos simulations relatives à l'ANOVA F de Fisher aient amené à des constats semblables à ceux obtenus sur base de nos simulations relativement au test t de Student. De plus, ces simulations nous ont permis de faire deux constats supplémentaires : d'abord, lorsqu'on compare plus de deux groupes, l'ANOVA F de Fisher est affectée par les écarts à la condition d'homogénéité des variances, même lorsque tous les échantillons sont de tailles identiques. Dans ce cas, le test devient plus libéral, ce qui signifie qu'il amène à rejeter l'hypothèse nulle plus souvent qu'attendu théoriquement, sur le long terme. Ensuite, plus le nombre d'échantillons comparés est important, plus le test est affecté par les violations de la condition d'homogénéité des variances.

Si la prise de conscience des limites d'une méthode est un premier pas très important, il est tout aussi important de savoir comment pallier ces limites. C'est pour cette raison que nos

simulations incluaient également les résultats de tests théoriquement jugés comme constituant de bonnes alternatives, plus robustes en cas de violation de la condition d'homogénéité des variances, à savoir les tests t de Welch, F de Fisher et F^* de Brown-Forsythe.

De plus, il est souvent recommandé aux chercheurs de tester préalablement la condition d'homogénéité des variances et ensuite d'utiliser soit le test t de Student (ou F de Fisher) soit une alternative plus robuste aux écarts à la condition d'homogénéité des variances, suivant que cette condition soit ou non respectée. Nous avons dès lors expliqué et illustré une faille importante du test de Levene, le test d'égalité des variances le plus susceptible d'être utilisé par les chercheurs en psychologie, de par son accessibilité dans les logiciels conviviaux tels qu'SPSS et Jamovi : la puissance du test de Levene à détecter les écarts à la condition d'homogénéité des variance est souvent très faible, si bien qu'il conduira le plus souvent à privilégier le test t de Student (ou l'ANOVA F de Fisher) aux alternatives plus robustes.

En ce qui concerne la comparaison des taux d'erreur de type I et II des tests t de Student et t de Welch, il est apparu que le test t de Welch est pratiquement aussi puissant que le test t de Student lorsque la condition d'homogénéité des variances est respectée, et contrôle bien mieux les taux d'erreur de type I et II lorsqu'elle ne l'est pas. De même, le test W de Welch est très légèrement inférieur aux tests F^* de Brown-Forsythe et F de Fisher en cas d'homogénéité des variances, tant en termes de contrôle des erreurs de type I et II qu'en termes de consistance entre les puissances théoriques et observées. Par contre, il leur est bien supérieur dans les cas les plus fréquents en psychologie, à savoir les cas de violation de la condition d'homogénéité des variances.

Après avoir décrit tous ces résultats, il nous semblait indispensable de résumer le message clé de ces deux premiers articles par des recommandations claires et précises. Cela nous a semblé d'autant plus important que bien souvent, 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 (Mills et al.,2010). La formulation de directives précises nous semblait être un moyen opportun de limiter ce temps.

Compte tenu du fait que la condition d'homogénéité des variances est plus souvent l'exception que la norme, qu'il est parfois très difficile (voire impossible) de détecter les écarts à cette

condition à travers des tests, et que la très légère perte de puissance des tests t et F de Welch lorsque la condition d'homogénéité des variances est respectée est largement compensée par le gain que constitue leur usage (en termes de contrôle des erreurs de type I et II) lorsque la condition d'homogénéité des variances n'est pas respectée, nous recommandons d'utiliser ces tests par défaut. Cette recommandation s'applique au moins au cas où les échantillons sont de tailles différentes, lorsqu'on ne compare que deux groupes, et s'applique dans tous les cas lorsqu'on compare plus de deux groupes.

Les choix de comparer les tests t de Student et F de Fisher respectivement aux tests de Welch et de Brown-Forsythe et finalement de recommander l'usage des tests de Welch par défaut étaient fortement guidés par le désir de proposer des stratégies qui pourraient être facilement comprises et appliquées par la grande majorité des chercheurs. Comme nous l'avions déjà mentionné en introduction, il existe des tests qui sont plus robustes simultanément aux violations des conditions de normalité et d'homogénéité des variances, tels que les tests où l'on compare des moyennes tronquées (Wilcox, 1994; Wilcox, 1998) ou encore les tests non paramétriques. Cependant, ces tests étaient à nos yeux moins susceptibles de provoquer l'adhésion de la majorité des chercheurs, pour deux raisons essentielles. Premièrement, ces tests ne reposent pas sur la même hypothèse nulle que les tests t de Student et F de Fisher, puisqu'on n'y compare plus les moyennes de chaque groupe. L'usage des tests de Welch, au contraire, constitue un moyen simple d'améliorer les pratiques sans pour autant obliger à repenser la manière de définir l'hypothèse nulle. Deuxièmement, les tests de Welch sont déjà implémentés dans la plupart des logiciels courants tels qu'SPSS, Jamovi et R. C'est même la stratégie proposée par défaut dans Jamovi et R, ce qui est important compte tenu de la propension des chercheurs à privilégier les méthodes proposées par défaut dans les logiciels (Counsell & Harlow, 2017).

Nous ne sous-entendons pas que les tests reposant sur les moyennes tronquées ou les tests non paramétriques sont à bannir et encore moins qu'un outil statistique n'est pas digne d'intérêt s'il n'est pas déjà implémenté dans les logiciels courants (il est de plus en plus abordable de proposer de nouveaux outils, par exemple via R). Nous pensons simplement qu'il était plus réaliste, dans un premier temps, de s'assurer que les hypothèses généralement définies par les chercheurs soient testées correctement, avant d'amener une réflexion sur la manière dont on peut améliorer leur définition.

Finalement, afin d'assurer l'accessibilité de nos travaux et de permettre à tous les chercheurs de disposer d'un maximum d'éléments pour les critiquer de manière éclairée, nous avons accordé une grande importance au fait de rendre gratuitement disponibles, en ligne, tant nos articles que l'ensemble des outils qui nous ont permis de les écrire. Nous avons effectué plusieurs démarches en ce sens : les articles présentés au sein des chapitres 2 et 3 ont été publiés dans *l'International Review of Social Psychology*, une revue Open Access. De plus, avant qu'ils n'aient été acceptés pour publication, nous avions diffusé des preprints de ces articles sur les réseaux sociaux (Facebook, Twitter...). Enfin, nous avons rendu disponibles en ligne tous les scripts de nos simulations et analyses, en utilisant la plateforme de l'*Open Science Framework* dans un premier temps, et Github ensuite.

Dans un deuxième temps, nous nous sommes intéressés à la significativité pratique des effets étudiés, au delà de leur significativité statistique. Cela implique d'étudier la taille des effets étudiés, au delà de la p-valeur. Plus spécifiquement, nous nous sommes focalisés sur le contexte de la comparaison de deux moyennes, dans la continuité de l'article présenté au sein deuxième chapitre de cette thèse.

En entamant ce chapitre, nous avions deux missions principales à l'esprit. Premièrement, nous voulions rappeler aux chercheurs qu'à l'instar des tests t de Student et F de Fisher, la mesure de taille d'effet la plus connue et la plus utilisée en vue de comparer la moyenne de deux groupes, à savoir le d de Cohen, n'est souvent pas appropriée. A travers l'article présenté au sein du chapitre 4, nous avons rappelé deux limites importantes de cette mesure. La première limite est que le d de Cohen est biaisé, même lorsque toutes les conditions dont il dépend sont respectées. Heureusement, il peut être transformé de sorte à annuler son biais lorsque la condition de normalité des résidus est respectée : la mesure transformée se nomme le g de Hedges, en référence à l'auteur ayant proposé cette transformation. La deuxième limite a été illustrée par de nouvelles simulations intensives pour un nombre très vaste de scénarios : une violation de la condition d'homogénéité des variances amène à une forte augmentation de la variance des estimateurs d de Cohen et g de Hedges, et ce même lorsque les deux échantillons sont de tailles identiques.

Deuxièmement, nous souhaitions apporter notre contribution scientifique, suite au constat d'un désaccord, de la part des méthodologistes, quant à la mesure de taille d'effet la plus appropriée à

utiliser lorsqu'on compare deux groupes sur base de leur moyenne. A cette fin, nous avons inclus dans nos simulations l'étude de différents estimateurs qui ont été proposés dans la littérature en vue de remplacer le traditionnel d de Cohen (et le g de Hedges) en cas de violation de la condition d'homogénéité des variances.

Par rapport aux simulations présentées au sein des chapitres 2 et 3, nous avons accordé plus d'importance au réalisme des scénarios envisagés, en nous appuyant sur l'investigation de Cain et al. (2017), de sorte à définir des déviations de la condition de normalité qui semblent crédibles dans les domaines de la recherche en psychologie. Il en est ressorti quelques constats très intéressants : parmi les estimateurs fréquemment proposés en vue de remplacer le traditionnel d de Cohen, on retrouve fréquemment le d de Glass. Celui-ci peut être transformé de sorte à obtenir le g de Glass, théoriquement non biaisé lorsque les résidus se distribuent normalement. Nos simulations ont révélé que la variance du g de Glass varie fortement en fonction de paramètres que l'on ne peut contrôler. Il en est de même pour son biais, lorsque les résidus sont extraits de populations qui ne se distribuent pas normalement. Il s'agit là d'un argument fort pour décourager l'usage de cette mesure. Or, ceci nous semble être un important apport théorique de notre article, dans la mesure où, à notre connaissance, personne avant nous n'avait révélé aux psychologues les failles du g de Glass de manière aussi détaillée. Dans la littérature, on retrouve également la mesure d de Shieh, qui entretient une relation mathématique directe avec le t de Welch, ainsi que la mesure d^* de Cohen qui, contrairement au d de Cohen classique, implique le calcul de la moyenne non poolée des variances de chaque groupe. De même que pour les estimateurs précédemment cités, il est possible de transformer ces mesures en vue de supprimer le biais lorsque la condition de normalité des résidus est respectée. Cela donne respectivement lieu aux mesures g^* de Hedges et g de Shieh. Grâce à nos simulations, nous avons révélé que le g^* de Hedges est supérieur au g de Shieh, non seulement d'un point de vue inférentiel (contrairement au g de Shieh, le g^* de Hedges est consistant, ce qui signifie que sa variance diminue toujours lorsque les tailles d'échantillon augmentent, de même que son biais lorsque les résidus sont extraits d'une population anormale) que d'un point de vue interprétatif (sa valeur est constante, peu importe que les deux échantillons soient de tailles identiques ou non). Finalement, lorsqu'on compare les mesures g de Hedges et g^* de Hedges, on constate que le g^* de Hedges n'est très légèrement inférieur au g de Hedges, en termes de biais et de variance, que lorsque des échantillons de tailles différentes sont extraits de populations aux variances

identiques. Il est tout aussi efficace que le g de Hedges lorsque tant les tailles d'échantillons que les variances de population sont identiques. De plus, il reste valide lorsque la condition d'homogénéité des variances n'est pas respectée, contrairement au g de Hedges.

Pour des raisons similaires à celles avancées précédemment, il nous semblait indispensable de conclure cet article par des recommandations très pratiques. C'est ce que nous avons fait en recommandant de privilégier le g^* de Hedges par défaut. Dans la mesure où cette solution n'est pas encore proposée dans la plupart des logiciels conviviaux tels que Jamovi et SPSS[42], notre article a été accompagné d'outils gratuits pour aider les chercheurs à mettre cette recommandation en oeuvre. Nous avons créé un nouveau package qui se nomme *deffectsize*[43] et pour ceux qui ne sont pas familiers avec R, nous avons également créé une application shiny.[44]

De même que pour les articles présentés au sein des chapitres 2 et 3, nous avons voulu assurer l'accessibilité et la transparence de notre recherche, et pour ce faire, nous avons été un cran plus loin que précédemment : alors que les articles antérieurs n'avaient été rendus accessibles que lorsqu'ils avaient atteint une forme relativement aboutie, via les preprints, la découverte de l'outil Github nous a permis de rendre cette recherche disponible dès le début de sa création.[45]

Le fait de le rendre rapidement disponible en ligne a permis la génération de diverses ressources théoriques et pratiques, grâce à la contribution d'autres chercheurs. D'un point de vue théorique, d'abord, le preprint a donné lieu à des échanges très enrichissants avec Geoff Cumming. Un des points abordés dans cet échange sera décrit dans la section "limites et perspectives futures" de cette thèse. D'un point de vue pratique, certains chercheurs se sont inspirés de notre article et des références que nous y citons pour améliorer des outils disponibles dans Jamovi et dans R. C'est le cas d'Aaron Caldwell[46] qui s'est appuyé sur nos travaux sur le g^* de Cohen pour améliorer la fonction du package "TOSTER" (disponible dans Jamovi) qui sert à réaliser un test d'équivalence dans le contexte de la comparaison des moyennes de deux échantillons indépendants.[47] C'est également le cas de Mattan S. Ben-Shachar[48] qui a pu constater des divergences entre la manière dont les bornes de l'intervalle de confiance autour de certains estimateurs de tailles d'effet étaient calculées dans le package *effectsize* dont il est le créateur (disponible sur le CRAN) et dans notre package *deffectsize*, et qui a pu corriger son package en conséquence.[49]

Dans un troisième temps, nous nous sommes concentrés sur la tendance des chercheurs à définir par défaut, comme hypothèse nulle, une hypothèse d'absence d'effet. Nous avons souligné que cette tendance persiste même lorsque l'objectif est de prouver une absence d'effet : c'est alors sur base d'un non-rejet de l'hypothèse nulle que les chercheurs affirment pouvoir valider leur hypothèse. Pourtant, nous avons vu que ce n'est pas une stratégie adéquate puisque non seulement le test utilisé de cette manière présente de faibles propriétés asymptotiques, mais en plus, la probabilité que le test amène à conclure à l'absence d'effet augmente à mesure que l'erreur de mesure augmente.

Nous avons également souligné qu'en réalité, il n'existe aucun test d'hypothèse qui permette de démontrer l'absence totale d'effet. Par contre, il est possible de démontrer qu'un effet observé ne s'éloigne pas de l'absence d'effet d'une quantité supérieure à une valeur définie (dit autrement, qu'il est *équivalent*), à condition de comprendre qu'il est théoriquement possible de définir n'importe quelle différence (ou intervalle de différences) entre les groupes comme hypothèse nulle. C'est le principe sur lequel repose le TOST (Two One-Sided Tests), à travers lequel on conclut à l'équivalence à condition que l'intervalle de confiance à $(1 - 2\alpha)\%$ autour de l'effet étudié soit entièrement inclus à l'intérieur de la zone d'équivalence.

Nous avons conclu cette thèse par une comparaison du TOST et du SGPV (Second Generation *P*-Value), récemment proposé par Blume et al. (2018) et défini par ses auteurs comme un nouvel outil permettant de calculer la proportion des valeurs de l'intervalle de confiance à $(1-\alpha)\%$ qui sont également compatibles avec l'hypothèse nulle (ou autrement dit, qui se situent à l'intérieur de la zone d'équivalence). Cette comparaison nous semblait pertinente, dans la mesure où les deux stratégies reposent sur un principe similaire, à savoir la comparaison de l'intervalle de confiance de l'effet observé avec la zone d'équivalence.

A travers notre investigation, nous avons révélé de nombreuses failles à l'usage du SGPV (par exemple, dans la mise en place d'une correction sous-optimale) et ne sommes pas parvenus à mettre en évidence une réelle plus-value de cet outil, par rapport à l'usage du TOST. Cette investigation démontre bien à quel point il est important, lorsqu'on propose un nouvel outil, de le comparer à des outils déjà existants afin d'en établir les forces et les faiblesses.

Limites et perspectives futures

Nous n'avons pas considéré le bootstrap dans les simulations que nous avons présentées au sein du chapitre 2 de cette thèse. Or, un ouvrage récemment écrit par Wilcox (2017) suggère de l'utiliser pour remplacer le test de Welch. Le bootstrap peut être utilisé pour tester l'hypothèse nulle d'égalité des moyennes (Efron et Tibshirani, 1993, expliquent dans leur ouvrage comment utiliser la technique du bootstrap à cet escient) tout en présentant l'avantage de ne pas reposer sur l'hypothèse de normalité des résidus, contrairement au test de Welch. Cette stratégie est toutefois connue pour ne pas être appropriée en toute circonstance (elle peut notamment conduire à des résultats peu fiables lorsqu'on utilise des échantillons contenant très peu de sujets; Efron & Tibshirani, 1993). Reproduire nos simulations en incluant le bootstrap pourrait dès lors constituer un intérêt, en vue de déterminer précisément dans quel cas le bootstrap constitue une réelle plus-value.

Après avoir soumis un preprint de l'article présenté au sein du chapitre 4 de cette thèse, nous avons eu le plaisir de recevoir un feedback très détaillé de Geoff Cumming.[50] Ce feedback nous a fait prendre conscience d'une limite importante de l'article : au cours de l'étude que nous y présentons, nous avons accordé plus d'importance aux propriétés inférentielles des estimateurs étudiés qu'à leurs propriétés interprétatives dans le choix des estimateurs comparés via nos simulations. Au delà de la significativité *statistique*, les mesures de taille d'effet donnent une information relative à la significativité *pratique* des effets étudiés.

Comme nous le rappelons en introduction du chapitre 4, l'un des objectifs des mesures de taille d'effet est de fournir une information qui aidera le chercheur à statuer sur la *pertinence* d'un effet en situation réelle. Il est important de pouvoir déterminer, dans un contexte donné, à partir de quelle valeur une mesure de taille d'effet présentera un intérêt aux yeux du chercheur (ou autrement dit, d'être capable d'*interpréter* la mesure obtenue). Dans le contexte des analyses de puissance a priori, par exemple, cela permettra de déterminer les tailles des échantillons qui nous assureront une puissance suffisante en vue de détecter des effets jugés dignes d'intérêt. Cela permettra également de définir des hypothèses de test plus informatives que la traditionnelle hypothèse de présence d'un effet non nul, comme nous l'avons abordé au sein du chapitre 5 de cette thèse.

Bien sûr, cela ne retire en rien l'importance d'avoir de bonnes propriétés inférentielles. Il est difficile, par exemple, de concevoir qu'un estimateur puisse fournir une interprétation adéquate s'il est extrêmement biaisé et si ses propriétés dépendent fortement de paramètres que l'on ne peut contrôler. C'est d'ailleurs la raison qui nous empêche de partager l'enthousiasme du Dr. Cumming à l'égard du d de Glass. Cependant, si avoir de bonnes propriétés inférentielles est requis pour un estimateur, cela ne suffit pas.

Peut-être avons-nous un peu trop mis l'accent sur les propriétés inférentielles afin de contourner une difficulté. La question de l'interprétation n'est certes pas une question simple (elle peut même devenir parfois très compliquée) et ce n'est probablement pas pour rien que les chercheurs reportent fréquemment des mesures sans leur fournir d'interprétation ni les inclure dans les discussions (Funder & Ozer, 2019; Thompson & Snyder, 1997). Elle n'en reste pas moins une question très importante, et réfléchir à des pistes, en vue de faciliter l'interprétation de l'estimateur g^* de Cohen, pourrait augmenter l'intérêt et la portée pédagogique de notre travail.

Lorsque nous avons construit le plan des simulations présentées dans le quatrième chapitre de cette thèse, nous avons accordé plus d'importance que précédemment au réalisme des distributions générées. Il reste cependant des pistes à explorer pour se rapprocher encore plus de la réalité des données. Par exemple, une voie intéressante serait de créer des variables dépendantes construites au départ d'échelles de Likert lors de la réalisation de simulations Monte Carlo, ces dernières étant très fréquemment utilisées dans la recherche en Psychologie (Croasmun & Ostrom, 2011; Joshi et al., 2015). A titre d'illustration, 9 des 10 articles publiés dans l'*International Journal of Clinical and Health Psychology* (volume 13, numéro 3, Septembre 2013) décrivaient l'usage ou le développement d'échelles de Likert (Hartley, 2014).

Les échelles de Likert sont constituées d'un ensemble d'affirmations (ou items) liées les unes aux autres. Le principe de base est de demander à un participant de statuer sur son degré d'accord pour chaque affirmation, de lui attribuer un score par affirmation en fonction de la réponse donnée (par exemple, 1 = "tout à fait d'accord", 2 = "d'accord"...), et de combiner ensuite l'ensemble des scores obtenus. Le score composite résultant est une manière de refléter la position du participant par rapport à un construit donné (Joshi et al., 2015).[51]

Dans la mesure où rien ne garantit que les différentes modalités d'une échelle de Likert soient

équidistantes (par exemple, rien ne garantit que l'écart entre les première et deuxième modalités "pas du tout d'accord" et "pas d'accord" soit le même que l'écart entre les deuxième et troisième modalités "pas d'accord" et "neutre"), plusieurs auteurs s'accordent à penser que les échelles de Likert sont de nature ordinale (voir par exemple H. N. Boone & D. A. Boone, 2012; Jamieson, 2004; Joshi et al., 2015). Cependant, il est également fréquemment admis que les scores composites résultant de la combinaison de plusieurs items de type Likert puissent être traités comme étant mesurés sur des échelles d'intervalle. D'aucuns considèrent que ce critère suffit à justifier l'usage d'outils reposant sur le calcul des moyennes et des écart-types (Joshi et al., 2015; Subedi, 2016). Si l'on partage cette conception, il fait tout à fait sens de tenter d'établir la robustesse des tests de Welch et de la mesure g^* de Cohen lorsque l'on simule des échelles de Likert. Soulignons qu'un tel projet serait de grande ampleur et qu'il impliquerait de se positionner par rapport à de nombreux degrés de liberté, tant ce qui concerne la manière de créer les items de type Likert (par exemple, combien de modalités à l'échelle? inclusion ou non d'une position neutre?) qu'en ce qui concerne la manière de les combiner pour former les échelles (par exemple, combien d'items forment l'échelle? Ont-ils tous une distribution de forme homogène ou pas?).

Par ailleurs, pour l'ensemble des études présentées, nous nous sommes exclusivement focalisés sur l'usage de tests lorsqu'on postule que les erreurs sont toutes indépendantes les unes des autres. Autrement dit, nous nous sommes concentrés sur l'étude de plans expérimentaux inter-sujets. Cependant, dans certains domaines de recherche en psychologie tel qu'en psychophysiologie, les ANOVAs sont fréquemment utilisées dans le cadre de plans à mesures répétées, c'est-à-dire lorsque les mêmes individus sont exposés à plusieurs conditions expérimentales (Vasey & Thayer, 1987).

Avec ce type de plan expérimental, l'existence d'une corrélation entre les mesures répétées implique que l'ANOVA, traditionnellement utilisée, recquiert une condition supplémentaire par rapport aux plans inter-sujets pour être valide : le respect de la condition de sphéricité. Il s'agit d'une condition d'après laquelle les différences de scores entre toutes les combinaisons de conditions possibles doivent être extraites de populations ayant la même variance (Lane, 2016).[52] Une violation de cette condition entraîne une augmentation du taux d'erreur de type I de l'ANOVA (Vasey & Thayer, 1987) qui sera d'autant plus forte que la violation est importante

(Lane, 2016).

De manière peu surprenante, compte tenu de ce que nous avons pu constater pour les plans expérimentaux inter-sujets, de nombreux chercheurs semblent ne pas avoir conscience de ces limites de l'approche classique. A titre d'illustration, Vasey et Thayer (1987) soulignaient que dans les volumes 21 et 22 de la revue *Psychophysiology*, parus en 1984 et 1985, plus de la moitié des études impliquant une ANOVA à mesure répétée ne discutaient pas de la condition de sphéricité (Vasey & Thayer, 1987).[53]

Lorsqu'on se penche sur les solutions proposées par les méthodologistes, celles-ci consistent souvent en des procédures en plusieurs étapes : dans un premier temps, on tente de détecter une violation de la condition de sphéricité et dans un deuxième temps, on prend une décision en conséquence.[54] A l'instar de nos articles présentés au sein des chapitres 2 à 4, un article pédagogique montrant les limites d'une telle approche sur base de simulations Monte Carlo, et proposant des recommandations concrètes et facilement applicables par la majorité des psychologues pourrait s'avérer utile. Cela s'avèrait d'autant plus nécessaire que par le passé, plusieurs références ont déjà souligné non seulement les limites des tests existants pour détecter les écarts à la condition de sphéricité (Vasey & Thayer, 1987)[55], mais également le caractère peu réaliste de la condition de sphéricité (Keselman et al., 2001; Keselman & Rogan, 1980; McCall & Appelbaum, 1973; O'Brien & Kaiser, 1985).

Pour finir, nous nous sommes exclusivement limités à l'étude de plans expérimentaux impliquant que les conditions expérimentales soient délimitées par un et un seul facteur catégoriel. Or, il arrive fréquemment que les chercheurs soient amenés à considérer des plans expérimentaux plus complexes, tels que des plans incluant plusieurs prédicteurs catégoriels, un (ou plusieurs) prédicteur(s) continu(s) ou encore simultanément des prédicteurs catégoriels et continus (cela arrive notamment lorsqu'on décide de prendre une covariable continue en compte dans notre plan d'étude; Judd et al., 2018).

Afin de faciliter la transition vers ces modèles plus élaborés, il nous semblerait intéressant d'écrire un court article expliquant comment le test de Welch peut être utilisé comme un outil servant à définir un modèle particulier de prédiction, conformément à l'approche proposée par Judd et al. (2010). Il s'agit d'une approche plus intégrée selon laquelle on se demande si un modèle

incluant le groupe d'appartenance d'un sujet particulier (appelé le *modèle augmenté*) permet de prédire le score de ce sujet avec moins d'erreur qu'un modèle n'incluant pas cette information (appelé le *modèle compact*), plutôt que de se demander s'il existe des différences significatives entre les moyennes des différents groupes étudiés comme on le fait traditionnellement.[56]

Conclusion générale

J'ai entamé cette thèse avec le désir d'aider les chercheurs à améliorer certaines de leurs pratiques d'analyses de données. Lorsque j'ai commencé à enseigner les statistiques, à l'ULB puis à l'UCLouvain, j'ai souvent observé que les étudiants appliquaient aveuglément les stratégies apprises durant leurs cours, et qu'ils ne se souciaient pas des conditions dans lesquelles ils les utilisaient. In fine, leur attention était bien souvent portée essentiellement sur la p-valeur du test utilisé plutôt que sur les statistiques descriptives et l'allure générale des données. J'ai plus tard eu le plaisir de côtoyer quelques chercheurs très compétents et sérieux et, pourtant, même parmi ceux-ci, il m'est arrivé de constater cette tendance.

La réalisation du travail de recherche dont cette thèse est l'aboutissement m'a aidée à réaliser que cette tendance ne se limitait pas aux étudiants à qui j'ai enseigné ou aux chercheurs avec qui j'ai collaboré, mais qu'elle reflétait une réalité à plus large échelle. Par ailleurs, elle n'était pas le résultat de la négligence de la part des chercheurs, mais plutôt d'un fossé qui existe encore aujourd'hui entre le monde des méthodologistes et celui des chercheurs pour qui la statistique n'est qu'un outil servant à répondre à des questions de recherche.

J'espère que je parviendrai, à travers cette thèse, et la lecture des articles qui la composent, à pousser mes lecteurs à prendre conscience du fait que la question des conditions d'application est bien plus qu'une considération intellectuelle, et qu'elle ne concerne pas que les experts en statistiques. La prise en compte de ces conditions a des implications réelles sur les résultats dans tout notre champ de recherche.

A travers cette thèse, nous avons expliqué que l'hétérogénéité des variances, très fréquemment présente dans la recherche en psychologie, pouvait conduire à une perte de puissance, ou à l'inverse à des inférences beaucoup trop audacieuses lorsqu'on utilise les tests paramétriques traditionnellement d'usage. Pour donner aux lecteurs des exemples concrets de conséquences non désirables que peut avoir l'hétérogénéité des variances, nous pouvons citer le fait d'obtenir des résultats non réplicables dont la perception d'importance a été amplifiée, que ce soit par la présence de faux positifs ou par une surestimation de la taille d'effet. Le contraire est également possible, à savoir le fait de passer à côté de résultats prometteurs suite à l'obtention de résultats non significatifs ou dont la perception d'importance a été dévaluée par une sous-estimation

de la taille d'effet. Dans un cas comme dans l'autre, nous pourrions même imaginer que cela influence négativement certaines décisions cliniques ou politiques, pour peu qu'elles se fondent sur des résultats non fiables. Pour paraphraser Wilcox (1998), il est temps de cesser de laisser des années de traditions nous pousser à utiliser des tests qui ont pour effet de remplir la littérature scientifique de résultats peu fiables.

Au cours de ce travail de recherche, mes collaborateurs et moi-même avons illustré la problématique des conditions d'application des tests à travers les cas des tests t et ANOVAs à un facteur pour échantillons indépendants. Il va de soi que cette problématique ne se limite pas à ces cas particuliers, et j'espère que nous sommes parvenus à faire prendre conscience à nos potentiels lecteurs de l'importance de bien considérer les enjeux des conditions d'application, quel que soit le plan expérimental étudié : par exemple, si la présence de deux conditions expérimentales suffit à induire de l'hétérogénéité des variances, comment concevoir qu'il en soit autrement lorsqu'on augmente le nombre de variables considérées, et dès lors, potentiellement, le nombre de sources d'hétérogénéité des variances possibles?

Toujours dans le cadre de ces illustrations, au-delà de la mise en lumière de la problématique, nous avons accordé beaucoup d'importance au fait de fournir des recommandations très précises, afin d'aider les chercheurs en psychologie à améliorer la fiabilité des résultats obtenus lorsqu'ils sont amenés à utiliser des tests t ou des ANOVAs à un facteur pour échantillons indépendants. Ce faisant, nous nous sommes efforcés de fournir des recommandations peu coûteuses, tant en termes de temps qu'en termes de flexibilité demandée. Les solutions proposées n'impliquent en effet pas de se former à de nouvelles techniques statistiques ou à l'usage de logiciels de traitement de données complexes.

Ces recommandations ne sont cependant pas exemptes de dérives possibles. Une première dérive serait de remplacer une stratégie, anciennement utilisée par défaut, par une nouvelle stratégie qui deviendrait le défaut à son tour, sans critiquer la pertinence de ce nouveau choix. Comme nous avons tenté de le souligner au sein des trois premiers articles, les tests de comparaison de moyennes dont font partie les tests de Welch ne sont pas toujours une solution adéquate. Ils ne conviennent par exemple pas lorsque les données présentent une forte asymétrie et que les échantillons sont de petite taille.

Une deuxième dérive possible serait de se limiter à l'étude de plans expérimentaux très simples, par crainte de se retrouver confronté à des situations plus complexes pour lesquelles on ne connait pas de solution robuste. Comme le soulignent judicieusement Judd et ses collaborateurs (2018), toutes les personnes confrontées à l'analyse de données devraient être capables *"d'interroger les données en posant les questions qui les intéressent au lieu de poser les questions que l'auteur d'un manuel ou d'un logiciel statistique supposait qu'ils souhaitent poser"* (p.12). Comme nous l'avons souligné dans la section dédiée aux perspectives, cette dérive pourrait être évitée en optant pour une approche plus intégrée, par exemple, en repensant le test de Welch selon l'approche par comparaison de modèles.

Enfin, ces techniques plus robustes à la présence d'hétérogénéité des variances, que nous avons illustrées au cours de cette thèse, pourraient être utilisées non seulement pour tester des hypothèses nulles d'absence de différence entre les moyennes mais également pour tester des hypothèses nulles de différence donnée de moyennes, tel qu'on le fait lorsqu'on utilise les tests d'équivalence, par exemple.

Le cœur du message de cette conclusion, et la réflexion à laquelle nous désirons amener le lecteur à travers celle-ci, est qu'il est toujours important de faire preuve d'esprit critique face à l'utilisation d'outils statistiques et de garder à l'esprit qu'un test statistique n'est pas une fin en soi. Il ne s'agit que d'un outil servant à tenter de répondre à une question de recherche. Pourtant, connaître ces tests statistiques, et être capable de choisir celui qui est le plus adapté au message que l'on souhaite faire passer, peut améliorer nos chances de parvenir à nos fins.

Notes de fin de chapitre

[40]Dans ce cas, les sujets ne sont pas répartis aléatoirement entre les groupes. Les variances inégales entre les groupes sont dès lors le résultats de la violation de la condition méthodologique d'indépendance des résidus.

[41]L'ANOVA F de Fisher peut être utilisée lorsqu'on compare deux ou plus de deux échantillons indépendants sur base de leur moyenne. Lorsqu'on compare exactement deux groupes, le test t de Student et l'ANOVA F de Fisher sont strictement équivalents. En effet, ils entretiennent la relation mathématique suivante : $F(1, x) = t^2(x)$.

[42]Dans Jamovi, par exemple, lorsqu'on réalise un test t de Welch tout en demandant une mesure de taille d'effet, c'est la mesure d^* de Cohen sans correction du biais qui est proposée.

[43]Ce package n'a pas été sousmis sur le CRAN de R. Pour pouvoir l'utiliser via la console R, vous devez appliquer le code suivant :
install.packages("devtools")
library(devtools)
install_github("mdelacre/deffectsize").

[44]L'application est disponible à l'adresse suivante : https://effectsize.shinyapps.io/deffsize/

[45]Le draft de l'article ainsi que l'ensemble des scripts et outputs générés sont disponibles via le lien suivant : https://github.com/mdelacre/Effect-sizes.

[46]Aaron Caldwell est un chercheur qui a obtenu un doctorat en Sciences de la Santé, du Sport et de l'Exercice à l'université d'Arkansas et qui réalise actuellement un post-doctorat axé sur la performance humaine dans les environnements extrêmes (chaleur, froid et altitude). Voici sa page : https://aaroncaldwell.us/

[47]Le package amélioré n'est pas encore disponible sur le CRAN et dans Jamovi, mais il est possible d'en avoir un aperçu via ce fil d'actualité Twitter : https://twitter.com/ExPhysStudent/status/1400861069048958981

[48]Mattan S. Ben-Shachar est un chercheur qui réalise actuellement un doctorat, au sein du laboratoire d'ERP neurocognitif développemental, à l'université Ben Gourion du Néguev (Israël). Voici sa page : https://sites.google.com/view/mattansb/

[49]Les modifications qu'il a apporté de la sorte sont les suivantes :
1) utilisation de la correction gamma exacte, plutôt qu'une approximation, en vue de supprimer le biais de l'estimateur g de Hedges;
2) utilisation de la méthode basée sur les distributions t non centrales pour définir les bornes de l'intervalle de confiance autour du g de Glass, alors que celles-ci étaient précédemment définies via la méthode du bootstrapping;
3) correction d'un bug dans le calcul de l'intervalle de confiance autour des traditionnels d de Cohen et g de Hedges (lié à une erreur dans la définition de la relation mathématique unissant le t de Student et le d de Cohen).

[50]Ce feedback a donné lieu à un échange et ensuite à un blog post, écrit par Geoff Cumming et disponible à l'adresse suivante : https://thenewstatistics.com/itns/2021/06/17/which-standardised-effect-size-measure-is-best- when-variances-are-unequal/.

[51]A l'origine, il n'est donc pas question d'analyser chaque item séparément (H. N. Boone & D. A. Boone, 2012), bien qu'il arrive que des chercheurs soient amenés à le faire (H. N. Boone & D. A. Boone, 2012; Joshi et al., 2015; Subedi, 2016).

[52]Par exemple, lorsqu'on réalise une ANOVA à mesures répétées avec 3 conditions nommées "A", "B" et "C", la condition de sphéricité sera respectée comme les différences "A-B", "A-C" et "B-C" sont toutes extraites de populations ayant la même variance (Lane, 2016).

[53]Comme le soulignent Vasey et Thayer (1987), ce pourcentage a été calculé en ne prenant en compte que les études pour lesquelles le facteur intra-sujets contenait au moins 3 modalités, puisque le problème de la sphéricité ne se pose pas lorsqu'il n'y a que deux mesures répétées.

[54]Parmi les alternatives proposées, on distingue l'approche multivariée (MANOVA; Vasey & Thayer, 1987; Lane, 2016) et une correction de l'approche univariée, consistant à corriger les degrés de liberté de la statistique F en les multipliant par une estimation de ϵ, un paramètre qui reflète le degré auquel la sphéricité est violée dans la population. $\epsilon = 1$ lorsque la condition de sphéricité est respectée et $\epsilon < 1$ quand elle est violée. L'argument est que la distribution de la stat univariée F serait mieux approximée par une distribution F avec un nombre réduit de degrés de liberté, d'après Box (1954). Les estimations les plus connues de ϵ sont $\hat{\epsilon}$ de Greenhouse et Geisser (Greenhouse & Geisser, 1959) et $\tilde{\epsilon}$ de Huynh et Feld (Huynh & Feldt, 1976). $\hat{\epsilon}$ est négativement biaisée, ce qui implique que la correction l'utilisant est légèrement conservatrice, et le sera d'autant plus que la violation de la condition de sphéricité est faible : lorsque la condition de sphéricité est respectée, $\hat{\epsilon} < 1$, les degrés de liberté diminuent donc, ce qui a pour conséquence de faire augmenter le seuil critique de la statistique F (Quintana & Maxwell, 1994). A l'inverse, $\tilde{\epsilon}$ tend à être de plus en plus libérale à mesure que la violation de la condition de sphéricité augmente, au point d'observer une inflation du taux d'erreur de type I en cas de violation sévère (Quintana & Maxwell, 1994). Quelle que soit la correction retenue, dans la mesure où la même correction est apportée sur les degrés de liberté du numérateur et du dénominateur de la statistique F, cela impactera de manière proportionnelle les valeurs représentant la part de variabilité expliquée par le facteur ($CM_{Facteur}$, le numérateur dans le calcul de la F) et la part de variabilité résiduelle ($CM_{Résidus}$, le dénominateur dans le calcul de la F). In fine, la valeur de la statistique F est inchangée : seuls les degrés de liberté sont impactés!

[55]Le test de Mauchly, par exemple, est sensible aux violations de la condition de normalité. Entre autres, il manquera de puissance avec des distributions à forte densité au niveau des extrémités ainsi que de manière plus générale, avec des petits échantillons (O'Brien & Kaiser, 1985). Le test de Box semble également manquer de puissance (Vasey & Thayer, 1987).

[56]La question de l'équation de prédiction appropriée à utiliser en cas d'hétérogénéité des variances ainsi que de l'approche utile pour tester la significativité du (ou des) coefficient(s) associé(s) au prédicteur "groupe d'appartenance" est abordée dans le forum de discussion suivant : https://stats.stackexchange.com/questions/142685/equivalent-to-welchs-t-test-in-gls-framework.