Puissance statistique des tests et date de publication : une corrélation trompeuse ?

Adrien Bernard Bonache

Puissance statistique des tests et date de publication : une corrélation trompeuse ?

Statistical power of tests and publication date: A spurious correlation?

Adrien Bernard BONACHE*

Résumé

.

Cet article synthétise une étude de la puissance statistique des tests des articles publiés dans la revue *Comptabilité – Contrôle – Audit* depuis vingt ans et de ses déterminants. Pour ce faire, nous avons réalisé des analyses de puissance, bivariée et multivariée. A l'instar des études antérieures, les analyses de puissance et bivariée montrent un accroissement significatif de la puissance statistique des tests des articles. Mais l'analyse multivariée révèle que la corrélation entre la date de publication et la puissance des tests est trompeuse. Cette corrélation pourrait être due à un biais de publication lié à l'utilisation de la signification statistique comme critère de publication.

Abstract

.

This article synthesizes a study of the statistical power of tests of the studies published in Comptabilité – Contrôle – Audit during the past twenty years and its determinants. To this end, we performed a power analysis, and bivariate and multivariate analyses. Similar to previous studies, the power and bivariate analyses demonstrated a statistically significant increase in the statistical power of tests of the articles studied. However, the multivariate analysis uncovered the spuriousness of the correlation between the publication date and the power of tests of these articles. This spurious relation could stem from a publication bias due to the use of statistical significance as a criterion for publication.

MOTS CLES : PUISSANCE – TAILLE D'ECHANTILLON – CORRELATION TROMPEUSE – BIAIS DE PUBLICATION

KEYWORDS: POWER – SAMPLE SIZE – SPURIOUS RELATION – PUBLICATION BIAS

* Maître de conférences des Universités, Université de Bourgogne-Franche Comté, Institut d'Administration des Entreprises, CREGO EA 7317, PICCO.

Correspondance : Adrien Bernard Bonache
Institut d'Administration des Entreprises
2 Bd Gabriel
21 000 DIJON
Adrien.BONACHE@u-bourgogne.fr

Remerciements : L'auteur tient à remercier la rédactrice en chef de la revue *Comptabilité-Contrôle-Audit*, Professeure Aude Deville, et les deux réviseurs anonymes pour leurs précieux conseils et judicieuses remarques. Véronique Collange et Jimmy Lopez par leurs lectures attentives du papier de travail et de la première révision respectivement ont contribué à ce projet, ainsi que les membres de l'axe PICCO par leurs commentaires lors d'une présentation de la première révision. Enfin, l'auteur est très reconnaissant envers les organisateurs et participants de l'*English Summer School 2016*. Notamment, Mary Bouley a corrigé des coquilles dans la version anglaise du résumé et Professeur Philippe Desbrières a suggéré l'utilisation d'une régression non-linéaire. Mais l'auteur assume seul les imperfections résiduelles de ce travail.

Introduction

Suite à une étude statistique sur un échantillon représentatif, imaginons que nous n'observons pas d'association significative entre une variable supposée explicative et la variable expliquée. Pouvons-nous en conclure qu'il n'y a aucune association entre ces deux variables dans la population étudiée ? Il est possible qu'il n'y ait aucun lien entre ces deux variables dans la population. Mais l'échantillon peut aussi ne pas être assez important pour détecter une petite association. L'utilisation d'un échantillon issu d'une population pour tester une hypothèse nulle (la nullité de l'association dans la population, ici) risque donc de nous faire conclure à tort que cette hypothèse n'est pas rejetée. Ce risque est l'erreur de seconde espèce.

La probabilité de cette erreur est égale à β. La puissance d'un test statistique est la probabilité de rejeter l'hypothèse nulle, sachant qu'elle est fausse. La puissance statistique d'un test vaut donc 1 – β. En sciences sociales, une étude utilisant des tests statistiques cherche généralement à montrer l'existence d'un effet. Pour ce faire, elle a besoin de tests d'autant plus puissants que l'effet est faible.

Pour expliquer ce lien entre la puissance statistique d'un test et la taille de l'effet étudié, on peut comparer la puissance statistique d'un test au pouvoir grossissant d'un microscope. Un microscope avec un pouvoir grossissant important permet d'observer des choses invisibles en utilisant un microscope au pouvoir grossissant plus faible. Il en est de même avec un test statistique réalisé sur un échantillon issu d'une population pour tester l'absence d'un effet. Une puissance statistique suffisante est nécessaire pour montrer l'existence de l'effet. Plus cet effet est petit, plus la puissance statistique du test devra être importante. La taille d'un effet est souvent définie comme l'écart entre la valeur de la statistique de test sous l'hypothèse nulle et sa valeur dans la population (Cohen 1988, p. 9). La taille d'un effet est parfois définie comme l'importance de cet effet en termes pratiques (Kirk 1996). En somme, la taille d'un effet peut se définir comme toute mesure reflétant une quantité d'intérêt (association, différence…) soit en termes absolus, soit relativement à une valeur spécifiée (Preacher et Kelley 2011).

Depuis maintenant plus de cinquante ans (Cohen 1962), des articles en sciences sociales recommandent aux chercheurs de considérer la puissance des tests, mais peinent à montrer un accroissement de la puissance statistique des tests (Sedlmeier et Gigerenzer 1989). En comptabilité, contrôle et audit, Lindsay (1993) montre aussi une faible puissance statistique des tests des études publiées. Cependant, Borkowski *et al.* (2001) et McSwain (2004) détectent la présence d'une association entre la date de publication et la puissance des tests des articles publiés en comptabilité, contrôle et audit. Cette association est mise en avant tant en considérant plusieurs disciplines (Borkowski *et al.* 2001), qu'en considérant un champ particulier (Helmuth *et al.* 2013 ; Helmuth *et al.* 2015). Helmuth *et al.* (2015) interprètent cette corrélation comme le résultat d'une maturation d'un champ, une prise en compte de la puissance des tests. Mais cette corrélation peut être trompeuse. Une corrélation trompeuse résulte non d'une relation entre deux variables, mais d'une troisième variable non prise en compte influençant ces deux variables corrélées (de Vaus 2001, p. 207).

Comme il existe dans le champ de la comptabilité, du contrôle et de l'audit un biais de publication dû à l'utilisation de la signification statistique comme critère de publication (Lindsay 1994 ; Dyckman 2016), les études d'effets faibles requièrent plus d'effort de collecte de données pour donner lieu à publication (Jennions et Møller 2002). Ainsi, la taille de l'échantillon influencerait à la fois la puissance des tests et la date de publication de l'article.

La corrélation, entre la date de publication et la puissance des tests des articles, observée dans les études précédentes serait donc trompeuse.

Montrer que cette corrélation est trompeuse avec un design longitudinal rétrospectif requiert l'utilisation de contrôles statistiques. Un premier contrôle peut aussi être effectué avant l'analyse en limitant l'étude à une seule revue. Ainsi, le présent article a pour but de répondre à la question : *La corrélation entre la puissance des tests et la date de publication des articles publiés dans la revue Comptabilité – Contrôle – Audit est-elle trompeuse ?* Pour répondre à cette question, nous avons d'abord synthétisé la littérature pertinente. Cela nous a permis d'identifier le rôle de la taille de l'échantillon. Puis, une analyse de puissance et des calculs de corrélation ont permis de retrouver l'association entre la date de publication et la puissance des tests des articles publiés dans *Comptabilité – Contrôle – Audit*. Enfin, une analyse multivariée a révélé l'effet du contrôle de la taille de l'échantillon sur la relation entre la date de publication et la puissance des tests. Le contrôle de la taille de l'échantillon dans notre modèle rend non significative l'association entre la date de publication et la puissance des tests des articles ; la taille de l'échantillon est associée positivement et significativement à la puissance des tests et à la date de publication. Enfin, puisque seuls des arguments théoriques permettent d'interpréter ces résultats, nous les discutons au regard de la littérature.

Par cette étude, nous souhaitions combler un manque dans la littérature. En testant l'association entre la date et la puissance statistique des tests d'une publication avec des variables de contrôle, nous espérions montrer les déterminants d'un accroissement de la puissance des tests des articles publiés en comptabilité, contrôle et audit. Les études antérieures se sont bornées à comparer les moyennes de la puissance statistique des tests par année en ne retirant pas l'effet de facteurs pouvant expliquer l'évolution observée. En contrôlant l'effet de la taille de l'échantillon, notre étude révèle que la corrélation date de publication-puissance des tests est trompeuse. De plus, notre recherche permet d'évaluer les moyens utilisés par les chercheurs en comptabilité, contrôle et audit pour accroître la taille de leur échantillon et indirectement la puissance des tests. Il s'agit de l'utilisation de designs longitudinaux, de tests unilatéraux et de l'entreprise comme unité d'analyse. L'utilisation de base de données secondaires ne permet pas, à elle seule, d'accroître la taille de l'échantillon et la puissance des tests. L'utilisation de données internationales est associée positivement à la taille de l'échantillon et seulement indirectement à la puissance des tests.

Pour montrer le bien-fondé de ces contributions, leurs fondements théoriques et conceptuels sont d'abord exposés (1). Puis, les bases méthodologiques et les résultats, à l'origine de nos contributions, sont présentés (2). Enfin, une conclusion et une discussion de ces résultats permettent d'esquisser des pistes de recherche.

1 Fondements théoriques et conceptuels

Ce premier développement est une lecture critique des précédentes analyses de puissance. Cela permet de motiver la question de recherche liminaire. Pour ce faire, nous résumons les précédentes analyses de puissance réalisées dans le champ de la comptabilité, du contrôle et de l'audit (1.1). Puis, nous justifions l'hypothèse de la taille de l'échantillon comme variable omise dans les précédentes études (1.2).

1.1 Littérature pertinente

Dans le champ de la comptabilité, du contrôle et de l'audit, la puissance statistique des tests est un concept relativement peu présent. Pourtant, depuis les années 1980, des articles présentent le concept de puissance statistique et son opérationnalisation (Duke *et al.* 1982 ; Kinney 1986 ; Bailey *et al.* 1999). Mais, peu d'études évaluent a posteriori la puissance des tests des travaux de recherche publiés en comptabilité, contrôle et audit (Lindsay 1993 ; Borkowski *et al.* 2001 ; McSwain 2004). Le tableau 1 résume ces trois études et les compare à la nôtre. Ces trois études sont des analyses de puissance. Une analyse de puissance consiste à calculer la puissance de chaque test présenté dans des articles empiriques en supposant a priori des tailles d'effet petite, moyenne et grande. Pour chaque type de taille d'effet, la moyenne arithmétique simple des puissances statistiques des tests est calculée par article. Enfin, les effectifs sont rapprochés pour voir le nombre d'articles considérés comme ayant des tests d'une puissance suffisante pour détecter des effets petit, moyen et grand.

L'analyse de puissance de Lindsay (1993) plaide en faveur de la prise en compte de la puissance des tests dans les articles empiriques en comptabilité en soulignant les conséquences de l'absence de considération pour ce concept. Lindsay (1993) illustre ce manque de considération en réalisant une analyse de la puissance des tests de 43 articles

empiriques étudiant statistiquement le contrôle et la planification budgétaires. Cette analyse de puissance est donc limitée à un seul champ de recherche.

Borkowski *et al.* (2001) étudient l'évolution de la puissance des tests en comptabilité comportementale suite à la publication de Lindsay. La puissance des tests moyenne était alors plus élevée que dans l'étude de Lindsay (1993). Ces auteurs abordent aussi la problématique de l'évolution de la puissance des tests en sciences sociales. Pour cela, ils réalisent des analyses graphique et multivariée basées sur les résultats des précédentes analyses de puissance. Cela suppose implicitement que la diffusion du concept de puissance est la même dans différentes disciplines. De plus, les analyses de puissance utilisées par Borkowski *et al.* ne retiennent pas toutes les mêmes définitions opérationnelles d'un effet petit, moyen ou grand pour certaines statistiques de test. Cohen a révisé leurs définitions en 1969 (Sedlmeier et Gigerenzer 1989). Ainsi, l'accroissement de la puissance des tests des articles reste problématique. McSwain (2004) arrive à cette conclusion après une analyse de puissance de tests de publications sur les systèmes d'information comptables.

Sur ce thème, l'analyse de puissance de McSwain (2004) présente des résultats proches de ceux de Borkowski *et al.* (2001). Par ailleurs, McSwain a exploré statistiquement l'association entre le type de tests, le support de publication, l'année de parution, d'une part, et la puissance des tests, d'autre part. Sur cette base, McSwain discute la plus grande puissance des tests des articles publiés en 2000 dans le *Journal of Information Systems*. Cette plus haute puissance des tests pourrait être indicative d'une plus grande prise en considération de la puissance statistique par les chercheurs, éditeurs et relecteurs. Mais l'auteur ne parvient pas à montrer une tendance temporelle nette.

En somme, les trois articles se rejoignent sur le fait que les tests des études publiées en comptabilité manquent de puissance statistique, du moins pour détecter des effets petits et moyens tels qu'opérationnalisés par Cohen (1988). Les analyses de Borkowski *et al.* (2001) et McSwain (2004) suggèrent un accroissement de la puissance des tests des articles publiés en comptabilité, contrôle et audit. Mais Borkowski *et al.* (2001) et McSwain (2004) ne donnent aucune explication de cet accroissement suggéré. Ces résultats nous ont conduit à poser une première proposition :

Proposition 1 : La puissance des tests d'un article est associée positivement à sa date de publication.

Tableau 1

Comparaison de notre étude avec les précédentes analyses de puissance en comptabilité, contrôle et audit

Auteur(s) (année)	**Présente étude**	**McSwain (2004)**	**Borkowski *et al.* (2001)**	**Lindsay (1993)**
Revues étudiées	*Comptabilité – Contrôle – Audit*	*Journal of Information Systems, Journal of Management Information System*	*Issues in Accounting Education, Behavioral Research in Accounting, Journal of Management Accounting Research*	*Journal of Accounting Research, Accounting Review, Accounting Organizations, and Society*
Années de publication	1995 – 2016	1996 – 2000	1993 – 1997	1970 – 1987
Nombre d'articles	108	45	96	43
Champ des articles	Comptabilité, contrôle et audit	Système d'information comptable	Comptabilité comportementale	Contrôle et planification budgétaires
Moyenne de la puissance statistique des tests en supposant des tailles d'effet petite, moyenne et grande	0,4 ; 0,8 et 0,93	0,22 ; 0,74 et 0,92	0,23 ; 0,71 et 0,93	0,16 ; 0,59 et 0,83
Méthodes	Analyses de puissance, bivariée et régression bêta	Analyse de puissance, ANOVA	Analyse de puissance, analyse de puissance pour des sous-groupes d'articles, régression linéaire multivariée	Analyse de puissance, analyse de puissance pour des sous-groupes d'articles
Résultats	La puissance des tests est élevée, particulièrement en supposant une taille d'effet moyenne ou grande. La corrélation entre la date de publication et la puissance des tests est trompeuse du fait d'un biais de publication dû à l'utilisation de la signification statistique comme critère de publication. La taille de l'échantillon est la variable expliquant cette corrélation.	La puissance des tests est faible, surtout en supposant une taille d'effet petite. La puissance statistique des tests est associée au test utilisé, à l'année de publication et à la revue.	La puissance des tests est restée faible après l'analyse de puissance de Lindsay. La puissance des tests des études utilisant des étudiants ne diffère pas significativement de celle d'études faites sur des professionnels. L'année de publication des articles est associée significativement à la puissance moyenne des tests en supposant des tailles d'effet moyenne et grande.	La puissance des tests est faible. Celle des études en laboratoire est plus faible que celle des études basées sur des sondages.

Dans d'autres champs et disciplines, on n'observe pas forcément un accroissement de la puissance des tests et des explications sont avancées. Dans le champ du management de la chaîne de valeur, Helmuth *et al.* (2015) montrent une corrélation entre l'année de publication et la puissance des tests des études publiées. Selon ces auteurs, cette corrélation serait le signe d'une maturation de ce champ, prenant plus en considération la puissance des tests. En psychologie sociale, Sedlmeier et Gigerenzer (1989) mettent en avant une absence d'accroissement de la puissance des tests dans la revue où Cohen (1962) avait publié la première analyse de puissance. Sedlmeier et Gigerenzer suggèrent que le concept de puissance statistique des tests était largement ignoré vingt-deux ans après la première analyse de puissance de Cohen (1962). Plus précisément dans leur étude, l'absence d'accroissement de la puissance des tests viendrait de l'utilisation de procédures contrôlant le taux de faux positifs.

1.2 La taille de l'échantillon comme covariable

Pour augmenter la puissance des tests d'une étude, il est notamment possible d'accroître la taille de l'échantillon (Levin 2004). Un échantillon plus important permet de diminuer l'erreur d'échantillonnage et donc d'accroître la puissance des tests. Cela diminue aussi la signification statistique ou *p-value* (Hoenig et Heisey 2001).

Or, il existe des biais de publication dus à l'utilisation de la signification statistique comme critère de publication en comptabilité (Lindsay 1994 ; Dyckman 2016). Dyckman met en avant que les papiers empiriques publiés dans *The Accounting Review*, *Journal of Accounting Research* et *Journal of Accounting and Economics* de 2011 à 2014 s'appuient tous sur une approche fréquentiste. Cette approche s'appuie sur le rejet d'une hypothèse nulle en utilisant la signification statistique. Dyckman (2016, p. 324) souligne une erreur classique dans l'interprétation de la signification statistique ou *p-value*. Certains manuels inciteraient à lire cette probabilité conditionnelle comme étant la probabilité que l'hypothèse nulle soit vraie sachant la valeur de la statistique de test. Cette lecture erronée pourrait expliquer pourquoi « le test de signification est devenu synonyme de rigueur scientifique et la pierre angulaire de la connaissance scientifique » (Lindsay, 1994, p. 33). Lindsay a montré sur les mêmes articles que ceux de son analyse de puissance de 1993 qu'il y avait seulement 16 % d'articles pour lesquels la majorité (plus de 50 %) des hypothèses nulles centrales n'étaient pas rejetées.

Comme l'étude de Lindsay (1994) n'a pas été reconduite dans d'autres champs de la littérature, il convient d'évaluer la plausibilité de l'existence d'un tel biais de publication dus à l'utilisation de la *p-value* comme critère de publication et donc de la proposition :

Proposition 2 : Il existe un biais de publication dû à l'utilisation de la signification statistique comme critère de publication.

Lorsqu'il existe un tel biais de publication, les études portant sur des petits effets pourraient être publiées plus tardivement (Jennions et Møller 2002). L'utilisation de la signification statistique comme critère de publication privilégierait les études étudiant des effets plus importants. Ces études seraient donc publiées plus tôt. En revanche, les études portant sur des effets de petites tailles devraient avoir une taille d'échantillon suffisante pour satisfaire le critère de signification statistique et donc être publiées.

La taille de l'échantillon est donc une covariable. Face à un résultat non significatif, les chercheurs seraient incités soit à abandonner l'étude portant sur un effet petit, soit à accroître la taille de l'échantillon. En incitant les chercheurs à accroître la taille de leur échantillon, la faiblesse d'un effet retarderait la publication de l'étude. Il y aurait donc une corrélation positive entre la taille de l'échantillon et la date de publication. Mais la puissance statistique des tests dépendant positivement de la taille de l'échantillon, le contrôle statistique de la taille de l'échantillon pourrait diminuer l'association observée dans les études antérieures entre la puissance des tests et la date de publication des articles. Ces arguments permettent de proposer :

Proposition 3 : La taille de l'échantillon est associée positivement à la date de publication et à la puissance des tests d'un article.

Pour accroître la taille de l'échantillon d'une étude, il est possible d'élargir la population dont est tiré l'échantillon étudié. Cela peut être fait grâce à l'utilisation d'un échantillon international ou d'un plus faible niveau d'analyse (Helmuth *et al.* 2013 ; Zhan 2013). Il est possible aussi d'accroître le nombre d'observations avec une même population en utilisant un design longitudinal, une base de données secondaires (Zhan 2013) ou un échantillon d'étudiants (Borkowski *et al.*, 2001).

En somme, ce premier développement a permis de résumer les précédents travaux motivant la question liminaire. Ce développement a permis d'avancer une proposition de

réponse. *Du fait d'un biais de publication dû à l'utilisation de la signification statistique comme critère de publication, la taille de l'échantillon est associée positivement à la date de publication et à la puissance des tests d'un article. En présence d'un tel biais de publication, si l'on contrôle la taille de l'échantillon, l'association entre la puissance des tests et la date de publication d'un article est nulle.* Mais une proposition n'étant pas directement testable (Lee & Lings, 2008, p. 128), le développement suivant expose la méthode et les résultats permettant de répondre à la question liminaire.

2. Étude empirique

Le but de ce second développement est de présenter les éléments à la base de notre réponse à la question liminaire. Pour ce faire, nous avons opérationnalisé les construits présentés dans la première partie, retenu des contrôles statistiques pour améliorer la validité interne de nos conclusions et choisi des méthodes de collecte et d'analyse. Ces éléments méthodologiques sont présentés (2.1), avant d'exposer les résultats (2.2).

2.1 Méthode

2.2.1 Design de l'étude et échantillon

Pour faire des inférences plausibles et avoir des conclusions ayant une forte validité interne, il convient de choisir un design (de Vaus 2001). Pour répondre à notre question liminaire, l'utilisation d'un design en coupe instantanée ne semble pas envisageable. L'utilisation d'un tel design ne permet logiquement de voir un changement (Bono et McNamara 2011). De plus, il ne semble pas possible de répondre à une telle question avec un design expérimental du fait de problèmes méthodologiques, pratiques et éthiques (de Vaus 2001). Ainsi compte tenu de la question traitée et des problèmes associés au design expérimental, nous avons retenu un design longitudinal rétrospectif.

Ainsi, pour répondre à notre question liminaire, nous nous basons sur les articles empiriques utilisant des tests statistiques publiés dans *Comptabilité – Contrôle – Audit* entre janvier 1995 et avril 2016. Nous avons collecté les articles publiés après juin 2009 sur le site de l'Association Française de Comptabilité (http://www.afc-

cca.com/contenu.php?rub=22&srub=32, accédé le 20 mai 2016). Nous avons collecté les articles parus avant juin 2009 sur la base de données CAIRN (http://www.cairn.info/revue-comptabilite-controle-audit.htm, accédé le 17 septembre 2016). L'utilisation des publications d'une seule revue pose des problèmes de validité externe. Mais la finalité de notre étude est de montrer que la taille des échantillons utilisés explique la corrélation observée dans la littérature entre la puissance des tests et la date de publication des articles. Nous ne cherchons pas à généraliser les résultats à l'ensemble des revues du champ de la comptabilité, du contrôle et de l'audit. Ainsi, nous avons préféré contrôler l'effet du support de publication, montré par McSwain (2004), en limitant notre analyse à une seule revue pour améliorer la validité interne de nos conclusions.

Sur les 431 articles publiés dans les numéros 1(1) à 22(1) de *Comptabilité – Contrôle – Audit*, 118 articles contiennent des tests statistiques de signification (tableau 2). Parmi ceux-là, dix ont dû être retirés car les détails de ces articles ne permettent pas de calculer la puissance des tests. Ainsi, la taille de notre échantillon est de 108 articles.

Tableau 2
Description de l'échantillon

Années de publication considérées : 1995 – 2016	**Articles**
Articles publiés dans la revue (hors commentaires libres, découvertes des lecteurs, éditoriaux, hommages et revues des livres et des thèses)	431
– Articles non empiriques et/ou ne présentant de test de signification	– 313
Articles utilisant des tests de signification	118
– Articles ne présentant assez de détails pour calculer la puissance des tests	– 10
Echantillon d'articles utilisé	**108**

2.2.2 Mesures des variables

Le but de notre article est de montrer que l'accroissement de la puissance des tests est dû à une augmentation de la taille des échantillons utilisés dans les études publiées en comptabilité, contrôle et audit. Ainsi, la variable à expliquer est la puissance des tests ; les variables supposées explicatives sont la date de publication de l'article et la taille de l'échantillon.

Pour calculer la moyenne de la puissance des tests par article, nous avons fixé l'erreur de première espèce à 5 %. Avec le paquet *pwr* sur le logiciel R, nous avons calculé la puissance des tests usuels en retenant les tailles d'effet petite, moyenne et grande définies par Cohen (1988). En utilisant le logiciel G*Power 3.1.9.2, les résultats obtenus sont identiques. L'annexe 1 détaille le calcul de la puissance des tests des articles utilisant des régressions logistiques, équations structurelles ou tests de Wilcoxon-Mann-Whitney. Pour les articles formulant explicitement des hypothèses et/ou un modèle, nous n'avons évalué la puissance moyenne qu'à partir de la puissance des tests statistiques relatifs aux hypothèses centrales. Nous n'avons pas pris en compte les tests de robustesse et les tests des hypothèses auxiliaires pour calculer la puissance moyenne des tests de ces articles. Lorsque les hypothèses précisent le sens de l'effet attendu, nous avons calculé la puissance des tests associés en considérant ces tests comme unilatéraux. Autrement, nous avons supposé les tests bilatéraux. Nous n'avons pas considéré tous les tests comme étant bilatéraux pour ne pas diminuer la validité interne de nos conclusions.

La définition opérationnelle de la date de l'article ne semble pas poser de problème a priori. Retenir la date de publication semble pertinent vu la finalité de notre étude. Certes, la date de parution d'un article peut être influencée par des choix éditoriaux, par la rapidité de lecture des épreuves ou des problèmes techniques d'édition. Mais la date d'acceptation de l'article n'est pas disponible pour tous les papiers publiés avant 2005 dans la revue *Comptabilité – Contrôle – Audit*. La date de réception de l'article ne semble pas plus pertinente. Dans le papier original, la formulation des hypothèses, la méthode utilisée, voire l'échantillon peuvent être très différents de ceux de l'article publié. De plus, la date de réception n'est pas disponible pour les articles publiés avant 2005. Nous avons donc retenu la date de publication de l'article. La définition opérationnelle est la différence en nombre de mois entre la parution du numéro dans lequel est publié l'article et janvier 1995 (premier mois de l'année du premier numéro de *Comptabilité – Contrôle – Audit*).

Concernant la taille de l'échantillon, la définition opérationnelle est la moyenne arithmétique simple par article du nombre d'observations utilisées pour chaque test pour les articles utilisant plusieurs tests. Pour les rares articles présentant le résultat d'un seul test, la taille d'échantillon retenue est le nombre d'observations utilisées pour ce test seul. Nous

n'avons pas utilisé le nombre d'observations totales utilisables, car certains tests n'utilisent pas toutes ces observations.

Des variables omises peuvent diminuer la validité interne des conclusions issues de résultats d'une analyse multivariée. Il convient donc d'utiliser des contrôles statistiques. Pour ce faire, nous avons recensé dans la littérature les facteurs pouvant jouer sur les relations entre la puissance des tests et la date de publication d'un article. Les précédentes analyses de puissance en comptabilité ont identifié les covariables méthode et revue (McSwain 2004) et l'utilisation d'échantillon d'étudiants (Lindsay 1993 ; Borkowski *et al.* 2001). Au-delà de celles-là, les récentes études de Zhan (2013) et Helmuth *et al.* (2013) ont mis en avant d'autres covariables : le caractère international de l'échantillon, le design de l'étude et l'origine des données. Le tableau 3 résume les contrôles statistiques utilisés, leurs définitions opérationnelles et le signe attendu. L'annexe 2 précise les fondements du signe attendu.

Tableau 3

Description et opérationnalisation des variables de contrôle

Nom de la variable	Identifiant	Définition opérationnelle	Signe attendu
Utilisation d'équation(s) structurelle(s)	SEM	Si utilisation 1, sinon 0	?
Utilisation de régression(s) logistique(s)	LOGIT	Si utilisation 1, sinon 0	?
Échantillon international	INTER	Si plusieurs pays 1, sinon 0	?
Étude longitudinale	LONGI	Si longitudinale 1, sinon 0	+
Échantillon d'étudiants	STUDENT	Si étudiants 1, sinon 0	?
Utilisation de données secondaires	DATA	Si secondaires 1, sinon 0	?
Niveau d'analyse	UNIT_ANAL	Si les cas sont des individus 2, des entreprises 1, sinon 0	?
Part des tests unilatéraux	TEST_UNILAT	% d'hypothèses indiquant le signe de la relation testée	+

2.2.4 Modèles et hypothèses

Maintenant définis opérationnellement nos construits, nous pouvons présenter le modèle (1) testé avec une régression bêta (Ferrari et Cribari-Neto 2004) et les hypothèses associées aux propositions issues de la littérature.

$$PUISSANCE_i = \beta_0 + \beta_1\, DATE_i + \beta_2\, N_i + \beta_3\, SEM_i + \beta_4\, LOGIT_i + \beta_5\, INTER_i + \beta_6\, LONGI_i + \beta_7\, STUDENT_i + \beta_8\, DATA_i + \beta_9\, UNIT\_ANAL_i + \beta_{10}\, TEST\_UNILAT_i + \varepsilon_i \quad (1)$$

$PUISSANCE_i$ est la moyenne de la puissance statistique des tests, $DATE_i$ la date de publication et N_i la taille de l'échantillon de l'article i. L'unité d'analyse est donc l'article utilisant des tests statistiques de signification. La puissance des tests peut être estimée pour une taille d'effet petite, moyenne ou grande. Mais notre analyse de puissance a montré une trop faible variance de la moyenne de la puissance des tests par article en supposant un effet moyen ou grand (cf. 2.2.1). Ainsi, $PUISSANCE_i$ est estimée en supposant un petit effet. Les autres variables sont définies dans le tableau 3. La première hypothèse nulle testée est $H1_0 : \beta_1 \leq 0$ contre l'hypothèse alternative $H1_a : \beta_1 > 0$. Nous avons testé cette hypothèse avec et sans l'inclusion dans le modèle de notre seconde variable d'intérêt, la taille de l'échantillon N_i. La seconde hypothèse nulle testée est $H2_0 : \beta_2 \leq 0$ contre l'hypothèse alternative $H2_a : \beta_2 > 0$. L'utilisation d'un modèle avec et sans la seconde variable d'intérêt a permis d'apprécier l'effet de la prise en compte de la taille de l'échantillon sur le lien entre la date de publication et la puissance de tests d'un article. Pour nous assurer que la taille de l'échantillon est associée à la date de publication et interpréter nos résultats, nous avons aussi estimé le modèle (2) avec la méthode des moindres carrés ordinaires.

$$N_i = \alpha_0 + \alpha_1\, INTER_i + \alpha_2\, LONGI_i + \alpha_3\, STUDENT_i + \alpha_4\, UNIT\_ANAL_i + \alpha_5\, DATA_i + \alpha_6\, DATE_i + \mu_i \quad (2)$$

Suivant Baron et Kenny (1986) et de Vaus (2014), pour s'assurer que l'association PUISSANCE-DATE est une corrélation trompeuse, il est nécessaire de rejeter $H1_0$ dans le modèle (1) sans N_i (modèle 1.1 désormais) et de ne pas la rejeter dans le modèle (1) avec N_i (modèle 1.2 désormais) et de rejeter $H2_0$. De plus, pour évaluer s'il s'agit d'une corrélation trompeuse, il convient de rejeter l'hypothèse nulle $H3_0 : \alpha_6 \leq 0$ au profit de l'hypothèse alternative $H3_a : \alpha_6 > 0$. Enfin, des arguments théoriques et logiques doivent permettre d'interpréter, comme une corrélation trompeuse et non une relation indirecte, nos résultats.

Pour évaluer la plausibilité de la deuxième proposition, portant sur l'existence d'un biais de publication dû à l'utilisation de la *p-value* comme critère de publication, nous avons utilisé une analyse proche de celle de Lindsay (1994 :42-43) – le codage des articles est précisé en

annexe 3. Nous avons ensuite calculé la proportion d'articles dont la majorité des hypothèses nulles centrales ne sont pas rejetées et utilisé un test non paramétrique binomial de l'hypothèse nulle $H4_0 : \pi \geq 0{,}5$ contre l'hypothèse alternative $H4_a : \pi < 0{,}5$. π est la proportion d'articles, dont la majorité des résultats des tests des hypothèses centrales sont négatifs, estimée par $p = u/k$. k est le nombre total d'articles empiriques formulant explicitement des hypothèses et u le nombre d'articles dont la majorité des hypothèses nulles centrales ne sont pas rejetées.

2.2 Résultats

Le but de ce point 2.2 est d'exposer le fondement empirique de nos discussions et conclusions relatives à la question liminaire. Pour ce faire, avant de montrer le résultat de notre régression (2.2.3), nous allons présenter les statistiques descriptives (2.2.2) et notre analyse de puissance des tests des articles publiés dans la revue *Comptabilité – Contrôle – Audit* depuis vingt ans (2.2.1 et tableau 4).

2.2.1 Analyse de puissance

Tableau 4

Description de la puissance moyenne des tests des 108 articles étudiés

Taille d'effet	Petite		Moyenne		Grande	
Puissance	**Effectif**	**% cumulé**	**Effectif**	**% cumulé**	**Effectif**	**% cumulé**
[0,99;1[	7	6,54	36	33,64	65	60,19
[0,95;0,99[	1	7,48	15	47,22	16	75,00
[0,90;0,95[	1	8,41	6	52,78	6	80,56
[0,80;0,90[	7	14,95	12	63,89	10	89,81
[0,70;0,80[	7	21,50	7	70,37	1	90,74
[0,60;0,70[	3	24,30	8	77,78	2	92,59
[0,50;0,60[	5	28,97	8	85,19	3	95,37
[0,40;0,50[	11	38,89	6	90,74	2	97,22
[0,30;0,40[	13	50,93	5	95,37	2	99,07
[0,20;0,30[	19	68,52	2	97,22		99,07
[0,10;0,20[	23	89,81	3	100,00	1	100,00
[0,05;0,10[	11	100,00		100,00		100,00
Total	108		108		108	
Moyenne	0,40		0,80		0,93	
Médiane	0,31		0,92		1,00	
Mode	0,15		0,995		0,995	
Écart-type	0,29		0,24		0,16	
Q1	0,16		0,64		0,95	
Q3	0,56		1,00		1,00	
Min	0,05		0,11		0,19	
Max	0,99998		1,00		1,00	
Skewness	0,82***		-1,14***		-2,80***	

Notes : Q1 et Q3 rendent compte du premier et troisième quartiles. La « skewness » ou coefficient de dissymétrie est le moment d'ordre trois centré et réduit de l'estimation de la puissance des tests par article. *** dénote une valeur de p inférieure à 1 %.

Le tableau 4 résume les effectifs et les fréquences cumulatives et les caractéristiques de forme et de dispersion de la moyenne de la puissance statistique des tests des articles. En moyenne, les études synthétisées ici ont près de quatre chances sur dix de détecter un effet petit. Plus de deux tiers des articles étudiés ont moins d'une chance sur deux d'observer une relation significative, si l'effet étudié est petit. Seulement seize de ces articles, soit près de 15 %, présentent des tests atteignant ou dépassant le seuil conventionnel de puissance de 0,8 (Cohen 1988). La distribution est étalée à droite (coefficient d'asymétrie de Pearson = 0,82, $p < 0,01$). Cette asymétrie et la présence de valeurs proches de l'unité (Max = 0,99998) légitiment l'utilisation d'une régression bêta (Cribari-Neto et Zeileis 2010).

Lorsqu'on suppose que l'effet est moyen dans la population, la puissance moyenne des tests est de 0,798, soit très proche du seuil conventionnel de puissance (0,8). De plus, près de deux tiers des études dépassent ce seuil. Lorsqu'on suppose un effet grand dans la population, près de 90 % des articles dépassent le seuil conventionnel de puissance. Un chercheur ayant publié une étude empirique dans la revue *Comptabilité–Contrôle–Audit* et étudié un effet grand avait plus de neuf chances sur dix (0,93) de rejeter l'hypothèse nulle. Ces résultats traduisent une faible variance de la puissance des tests estimée en supposant des effets moyens ou grands. Cette faible variance légitime l'utilisation de la puissance des tests estimée en supposant un effet petit dans les analyses bivariée et multivariée.

2.2.2 Statistique descriptive et analyse bivariée

La partie A du tableau 5 expose les statistiques descriptives des variables supposées explicatives et des variables de contrôle. On note que peu de papiers empiriques de *Comptabilité – Contrôle – Audit* ont utilisé une modélisation par équations structurelles (5,56 %) ou une régression logistique (12,96 %). Bien souvent, l'unité d'analyse est l'entreprise, les données sont secondaires et les tests unilatéraux.

Tableau 5 Statistiques descriptives

Partie A	**Caractéristiques de dispersion et de forme**						
Statistiques	Moyenne	Ecart-type	Min	Q1	Médiane	Q3	Max
DATE	137,69	64,04	2	94	143	188	255
N	370,99	1151,92	19	79,75	123	257,01	10523,23
SEM	0,06	0,23	0	0	0	0	1
LOGIT	0,13	0,34	0	0	0	0	1
INTER	0,10	0,30	0	0	0	0	1
LONGI	0,43	0,50	0	0	0	1	1
STUDENT	0,06	0,25	0	0	0	0	1
DATA	0,64	0,48	0	0	1	1	1
UNIT_ANAL	1,21	0,56	0	1	1	2	2
TEST_UNILAT	0,63	0,44	0	0	0,9	1	1

Partie B	**Matrice de corrélation de Pearson**									
Variables	1	2	3	4	5	6	7	8	9	10
1 PUISSANCE										
2 N	0,40***									
3 DATE	0,19**	0,15								
4 SEM	0,26***	-0,04	*0,20***							
5 LOGIT	-0,09	-0,05	0,13	-0,09						
6 INTER	0,11	*0,38****	-0,01	-0,08	-0,04					
7 LONGI	0,44***	*0,22***	*-0,17**	*-0,21***	0,00	0,08				
8 STUDENT	-0,09	-0,04	0,05	-0,06	0,01	-0,09	-0,07			
9 DATA	0,35***	*0,17**	0,01	*-0,24***	*0,17**	0,12	*0,56****	*-0,28****		
10 UNIT_ANAL	-0,35***	-0,11	-0,08	*0,20***	-0,10	-0,13	*-0,39****	*0,30****	*-0,72****	
11 TEST_UNILAT	0,20**	-0,08	0,12	0,04	*0,17**	-0,15	*0,20***	-0,08	*0,30****	-0,11

Tableau 5 (suite)

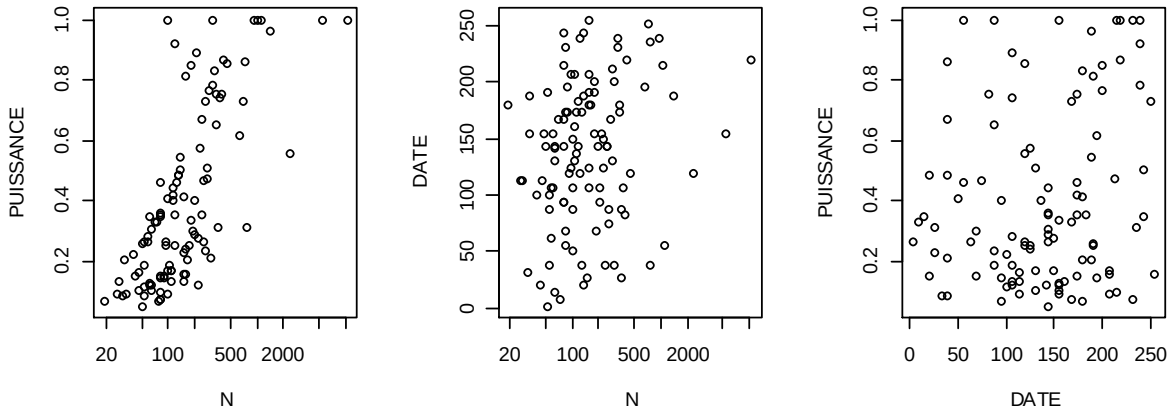
Notes : ***, ** et * indiquent que la valeur de p associée à l'hypothèse de nullité du coefficient de corrélation est inférieure à 1 %, 5 % et 10 %, respectivement, en supposant les tests bilatéraux. Nous avons corrigé ces valeurs de p avec la procédure de Benjamini et Yekutieli (2001) pour diminuer le taux de faux résultats positifs. Les caractères en italiques figurent les coefficients de corrélation indiquant la présence de colinéarité entre des variables de contrôle et/ou des variables supposées explicatives. N est la moyenne par article du nombre d'observations utilisé pour chaque test, DATE est la différence en nombre de mois entre la date de publication et janvier 1995 et PUISSANCE est la puissance moyenne des tests par article. Les autres variables sont définies dans le tableau 3.

La partie B du tableau 5 présente la matrice de corrélation pour voir la présence de colinéarité et apprécier, en l'absence de contrôle statistique, la plausibilité de nos hypothèses. En supposant une taille d'effet petite a priori, la PUISSANCE des tests est corrélée positivement et significativement à la DATE de publication des articles (0,19 ; $p < 0,05$) et à la taille de l'échantillon N (0,40 ; $p < 0,01$). Ces résultats permettent de rejeter les hypothèses nulles $H1_0$ et $H2_0$ et n'excluent pas la possibilité que la corrélation PUISSANCE-DATE soit trompeuse. En revanche, la taille de l'échantillon n'est pas corrélée significativement avec la date de publication (0,15 ; $p = 0,06$, *p-value* calculée en considérant que le test est unilatéral à droite). Cela ne permet pas de rejeter $H3_0$. Mais en l'absence de contrôle statistique, ces corrélations ne permettent pas de conclure.

Pour mieux visualiser ces relations, nous avons représenté les nuages de points entre la taille d'échantillon N, la DATE de publication de l'article et la PUISSANCE des tests (figure 1). La relation entre N et PUISSANCE est sigmoïdale. Du fait de cette forme en S (Long 1997), nous n'avons pas utilisé une régression linéaire pour estimer le modèle 1. Par ailleurs, une relation entre DATE et N semble perceptible, mais aucune relation n'apparait entre PUISSANCE et DATE.

Figure 1

Nuages de points des relations entre les variables supposées explicatives et la variable à expliquer



Notes : N est la moyenne par article du nombre d'observations utilisé pour chaque test, DATE est la différence en nombre de mois entre la date de publication et janvier 1995 et PUISSANCE est la puissance moyenne des tests par article.

Concernant la présence de colinéarité, certains coefficients de corrélation entre certaines variables explicatives et de contrôle sont significatifs (tableau 5, partie B, en italiques). Mais l'examen de la matrice de corrélation n'est pas suffisant pour détecter des colinéarités multiples. Ainsi, nous avons calculé les indices de conditionnement (Belsley *et al.* 1980) et facteurs d'inflation de la variance (Hair *et al.* 2010). Les indices de conditionnement étant tous inférieurs à trente (max(IC) = 3,38) et les facteurs d'inflation de la variance inférieurs à dix (tableau 6), la multicolinéarité ne semble pas réduire la validité des conclusions issues de notre analyse multivariée.

2.2.3 Analyse multivariée

Tableau 6

Résultats de l'analyse multivariée

Variables indépendantes	Signe attendu	Modèle 1.1		Modèle 1.2				
		$\hat{\beta}_j$	$\left(z_{\hat{\beta}_j}\right)$	$\hat{\beta}_j$	$\left(z_{\hat{\beta}_j}\right)$	VIF_j	f_j^2	$1 - \beta$
DATE	+	0,004***	(2,425)	0,002	(0,964)	1,223	0,012	0,303
N	+			0,001***	(8,530)	1,304	0,392	0,999
SEM	?	1,882***	(3,926)	2,134***	(4,781)	1,170	0,145	0,975
LOGIT	?	–0,075	(–0,242)	0,003	(0,012)	1,091	< 0,001	0,051
INTER	?	0,628*	(1,847)	–0,259	(–0,780)	1,217	0,001	0,068
LONGI	+	1,170***	(4,573)	0,827***	(3,540)	1,635	0,091	0,928
STUDENT	?	0,588	(1,346)	0,632	(1,572)	1,169	0,035	0,484
DATA	?	0,197	(0,551)	0,109	(0,334)	2,973	0,001	0,067
UNIT_ANAL	?	–0,287	(–1,075)	–0,448*	(–1,838)	2,278	0,018	0,282
TEST_UNILAT	+	0,209	(0,826)	0,337*	(1,444)	1,254	0,014	0,331
Constante	?	–1,324**	(–2,367)	–0,889	(–1,749)		0,020	0,307
Nombre d'observations		108		108				
Pseudo-R^2		32,82 %		51,74 %				
Chi-deux		60,577***		150,47***				
Variable expliquée		Moyenne de la puissance des tests exposés dans les articles publiés dans CCA de 1995 à 2016 en supposant une petite taille d'effet a priori						

Notes : *, ** et *** indiquent une probabilité critique inférieure à 10 %, 5 % et 1 % respectivement (+ tests unilatéraux à droite, ? tests bilatéraux). $\hat{\beta}_j$ est l'estimation du j-ème coefficient du modèle et $z_{\hat{\beta}_j}$ la *z-value* associée. VIF_j est le facteur d'inflation de la variance calculé pour la j-ème variable et f_j^2 la taille d'effet calculée a posteriori en rapportant la variance expliquée par cette variable à la variance résiduelle, pour le modèle 1.2. Enfin, $1 - \beta$ est la puissance ad hoc calculée à partir de la taille d'effet calculée a posteriori et avec un risque de

première espèce de 5 %. La fonction de lien utilisée est la fonction logit. N est la moyenne par article du nombre d'observations utilisé pour chaque test et DATE est la différence en nombre de mois entre la date de publication et janvier 1995 (année du premier numéro de CCA). Les autres variables sont définies dans le tableau 3.

Le tableau 6 présente les résultats d'une régression bêta sans et avec la taille de l'échantillon. Pour montrer que la corrélation entre DATE et PUISSANCE est trompeuse, nous avons d'abord montré une association entre ces deux variables sans contrôle statistique de la taille de l'échantillon N (modèle 1.1). Puis, nous avons étudié s'il y avait bien une association entre PUISSANCE et N et une disparition de l'association PUISSANCE–DATE en contrôlant statistiquement N (modèle 1.2). Enfin, pour montrer une relation trompeuse, nous avons examiné la présence d'une association entre la taille de l'échantillon et la date de publication des articles (tableau 7).

Le modèle 1.1 explique 32,82 % de la variance totale de la puissance des tests. L'association entre la date de publication et la puissance des tests est positive et significative ($\hat{\beta}_1 = 0{,}004$; $p = 0{,}008$). Cela permet de rejeter notre première hypothèse nulle au profit de la première hypothèse alternative selon laquelle l'association entre PUISSANCE et DATE est positive. Ainsi, la première condition d'une corrélation trompeuse est établie : association PUISSANCE–DATE en l'absence du contrôle statistique de la taille de l'échantillon N.

Le modèle 1.2 prenant en compte la taille de l'échantillon explique 51,74 % de la variance totale de la puissance des tests. Ce modèle montre une association positive et significative entre la puissance statistique des tests et la taille de l'échantillon de l'article ($\hat{\beta}_2 = 0{,}001$; $p < 0{,}001$). Cela permet de rejeter notre deuxième hypothèse nulle au profit de la deuxième hypothèse alternative selon laquelle l'association entre PUISSANCE et la taille de l'échantillon N est positive. En revanche, la prise en compte de la taille de l'échantillon ne permet plus de rejeter la première hypothèse nulle. L'inclusion dans le modèle de la taille de l'échantillon permet d'observer une association statistiquement non significative entre la date de publication et la puissance des tests des articles ($\hat{\beta}_1 = 0{,}002$; $p = 0{,}168$). Il y a bien une association entre PUISSANCE et N et une disparition de l'association PUISSANCE-DATE dans le modèle incluant N. Ces résultats mettent en avant le respect de la seconde condition d'une corrélation trompeuse.

En pratique, lorsqu'on prend en compte la taille de l'échantillon, l'estimation a posteriori de la taille de l'effet de la DATE de publication sur la PUISSANCE des tests est très faible ($f_1^2 = 0{,}012$), inférieure à une taille d'effet petite $f^2 = 0{,}02$ (Cohen 1988). En revanche, l'estimation a posteriori de la taille de l'effet de N (la taille de l'échantillon) sur la

PUISSANCE des tests est très importante ($f_2^2 = 0{,}392$), supérieure à la définition opérationnelle d'une taille d'effet grande $f^2 = 0{,}35$ (Cohen 1988). Remarquons que cela n'est pas le cas pour toutes les variables de contrôle.

L'utilisation d'un modèle d'équations structurelles ou d'un design longitudinal dans les articles étudiés est associée positivement et significativement avec la puissance statistique des tests. Les tailles d'effet estimées a posteriori des autres variables de contrôle sont inférieures à un effet moyen, $f^2 = 0{,}15$ (Cohen 1988). Nos observations ne peuvent indiquer si l'association est non nulle mais l'effet trop petit pour être détecté avec 108 articles ou si l'association est non significative car l'effet de ces variables de contrôle est nul.

Concernant la troisième condition d'existence d'une corrélation trompeuse, nous avons estimé le modèle 2 en utilisant la méthode des moindres carrés ordinaires (Tableau 7). Ce modèle 2 montre les moyens utilisés pour accroître la taille des échantillons par les chercheurs publiant dans la revue *Comptabilité – Contrôle – Audit*. Nous constatons que la taille de l'échantillon est associée significativement et positivement avec l'utilisation de données issues de plusieurs pays – INTER (1387,776 ; $p < 0{,}001$), de design longitudinaux – LONGI (531,061 ; $p = 0{,}038$) et avec la date de publication – DATE (3,586 ; $p = 0{,}031$). Ce dernier résultat permet de rejeter $H3_0$ au profit l'alternative $H3_a$ selon laquelle il existe une association positive entre la date de publication et la taille de l'échantillon des articles. Cela nous a permis de nous assurer du respect de la troisième condition d'une corrélation trompeuse. Cependant, seuls des arguments théoriques et logiques permettent d'interpréter le respect de ces trois conditions comme une corrélation trompeuse et non comme une relation indirecte.

Tableau 7

Les déterminants de la taille d'échantillon des articles

	Signe attendu	Coefficient	T de Student
INTER	?	1387,776***	4,118
LONGI	+	531,061**	2,101
STUDENT	+	–50,635	–0,116
UNIT_ANAL	+	173,494	0,656
DATA	+	122,103	0,359
DATE	+	3,586**	2,194
Constante	?	–776,671	–1,38
Nombre d'observations		108	
R^2 ajusté		17,22 %	

Notes : ** et *** indiquent une probabilité critique inférieure à 5 % et 1 % respectivement (+ tests unilatéraux à droite, ? tests bilatéraux). DATE est la différence en nombre de mois entre la date de publication et janvier 1995 (premier mois de l'année de la première parution de CCA). Les autres variables sont définies dans le tableau 3.

Enfin, pour évaluer la validité interne de l'étude, nous avons évalué la plausibilité de la proposition 2 portant sur l'existence d'un biais de publication dû à l'utilisation de la signification statistique comme critère de publication. Le test binomial effectué permet de rejeter l'hypothèse $H4_0$ au profit de $H4_a$ selon laquelle la plupart des articles rejettent au moins 50 % de leurs hypothèses nulles. Sur les 92 articles étudiés ayant pu faire l'objet d'une telle analyse, 61 (soit près de 66 %) rejettent au moins 50 % de leurs hypothèses nulles ($p < 0{,}001$). Ce résultat montre la prépondérance des articles présentant une majorité de résultats positifs.

Par ailleurs, nous avons conduit des analyses de robustesse et de sensibilité (annexe 4). Celles-ci nous ont permis de voir si les résultats des analyses multivariées ne peuvent pas être attribués à des observations influentes, l'inclusion de variables de contrôle ou la fonction de lien utilisée. Ces analyses ne remettent en cause nos conclusions et montrent la validité interne de notre étude.

Conclusion et discussion

Résumé des résultats

Y a-t-il un accroissement de la puissance des tests des études en comptabilité, contrôle et audit ? Le tableau 4 montre que la réponse à cette question dépend de la taille de l'effet supposée a priori pour calculer la puissance des tests. Si l'on suppose que les effets que tentent de mettre en avant les articles sont moyens ou grands, la variance de la puissance moyenne des tests des articles est trop faible pour voir un accroissement. Ce résultat nous a poussé à nous focaliser sur les puissances des tests calculées en supposant que les effets recherchés étaient petits. Les analyses bivariée (tableau 5) et multivariée sans prendre en compte la taille de l'échantillon (tableau 6, modèle 1.1) permettent de rejeter l'hypothèse nulle d'une absence d'association entre la puissance des tests des articles publiés de 1995 à 2016 dans la revue *Comptabilité – Contrôle – Audit*, d'une part, et leurs dates de publication, d'autre part. Sur cette base, on pourrait faire l'inférence d'un accroissement de la puissance statistique des tests des études publiées en comptabilité, contrôle et audit. Mais, l'inclusion de la taille de l'échantillon des articles dans la modélisation (tableau 6, modèle 1.2) permet de

montrer que l'association entre la puissance des tests et la taille de l'échantillon est positive. Or, la date de publication est associée à la taille de l'échantillon de l'article (tableau 7) et lorsque l'on inclut la taille de l'échantillon dans le modèle (1) la date de publication n'est plus associée significativement avec la puissance des tests (tableau 6, modèle 1.2). Ainsi, les trois conditions d'une relation indirecte ou d'une corrélation trompeuse sont réunies.

Discussion

Les résultats de l'analyse de puissance (2.2.1) peuvent être discutés en les rapprochant de ceux d'analyses dans la même discipline (tableau 1). En moyenne, la puissance des tests des articles publiés dans *Comptabilité – Contrôle – Audit* de 1995 à 2016 est supérieure à celle des articles synthétisés dans les précédentes analyses du genre en comptabilité (Lindsay 1993 ; Borkowski *et al.* 2001 ; McSwain 2004), si l'on suppose des effets petits ou moyens. Lorsqu'on considère un effet grand, on obtient une puissance moyenne supérieure à celle obtenue par Lindsay (1993) et proche de celles observées par Borkowski *et al.* (2001) et McSwain (2004). Cependant, il faut se garder de conclure de cette comparaison qu'il y a un accroissement de la puissance des tests dans notre champ de recherche. D'une part, les études synthétisées examinent des sujets différents de la recherche en comptabilité dans des revues différentes. D'autre part, les trois analyses de puissance précédentes en comptabilité posaient l'hypothèse que tous les tests d'hypothèses étaient bilatéraux. Cette supposition a tendance à réduire la puissance de chaque test unilatéral. Pour cette raison, nous n'avons pas supposé que tous les tests étaient bilatéraux, mais cela nuit à la comparabilité de nos résultats.

Sans contrôle statistique de la taille de l'échantillon, l'association entre la date de publication et la puissance des tests des articles est positive et significative. Cela corrobore les résultats obtenus en comptabilité comportementale par Borkowski *et al.* (2001) et en management de la chaîne de valeur par Helmuth *et al.* (2015). Mais cette association peut être considérée comme trompeuse ou indirecte, c'est-à-dire due à une co-variable corrélée à la date de publication et à la puissance des tests. Les biais de publication dus à l'utilisation de la signification statistique comme critère de publication en comptabilité (Lindsay 1994 ; Dyckman 2016) peuvent retarder la publication d'articles portant sur des petits effets et partant pousser les chercheurs à accroître la taille de l'échantillon pour montrer un tel effet (Jennions et Møller 2002). Ainsi, la corrélation observée entre la date de publication et la puissance des tests est soit trompeuse, soit indirecte. Certes, les conditions de Baron et Kenny (1986) d'une médiation sont réunies : association date de publication – taille de l'échantillon,

association entre taille de l'échantillon et puissance statistique des tests et, association entre date de publication et puissance statistique des tests sans prise en compte de la taille de l'échantillon. Mais, la taille d'échantillon est choisie par les chercheurs avant la réalisation des tests et la décision d'acceptation du papier pour publication. Le choix de la taille d'échantillon influence donc la date de publication et la puissance des tests calculée a posteriori. Ainsi, la corrélation entre la date de publication et la puissance des tests doit être interprétée comme étant trompeuse (de Vaus 2014). Ce résultat n'infirme pas pour autant les résultats d'Helmuth *et al.* (2015) et Borkowski *et al.* (2001), mais les explique par un biais de publication particulier. De plus, notre étude illustre en comptabilité, contrôle et audit l'explication mise en avant par Jennions et Møller (2002) dans une autre discipline.

Le résultat du test binomial, visant à tester la prépondérance des articles montrant majoritairement des résultats positifs, renforce la validité externe de l'étude de Lindsay (1994). Ce résultat suggère l'existence d'un biais de publication dû à l'utilisation de la signification statistique comme critère de publication dans une revue francophone publiant des articles de comptabilité, contrôle et d'audit depuis 1995. Lindsay (1994) avait montré l'existence d'un tel biais en se basant sur les articles publiés de 1970 à 1987 dans *The Accounting Review*, *Journal of Accounting Research* et *Accounting, Organizations and Society* dans le champ du contrôle et de la planification budgétaires. Cependant, nos résultats montrent une moindre prépondérance des articles montrant une majorité de résultats positifs, 66 % contre 84 % dans l'étude de Lindsay (1994). Nos analyses de sensibilité suggèrent l'absence de contingence temporelle : le taux de 66 % d'articles montrant au moins 50 % de résultats positifs semble constant.

Concernant les variables de contrôle issues des précédentes études, nos résultats ne vont pas systématiquement dans le sens de nos prédictions basées sur la littérature. Concernant la méthode utilisée, l'utilisation d'un design longitudinal, d'un échantillon d'étudiants et de tests unilatéraux, nos résultats corroborent les conclusions des études précédentes. La puissance des tests est associée significativement avec l'utilisation d'équations structurelles, mais pas avec l'utilisation d'un modèle logit. Ces résultats confirment partiellement des résultats de McSwain (2004) et Zhan (2013) mettant en avant la méthode comme déterminant de la puissance des tests. L'association positive et significative entre l'utilisation d'un design longitudinal et la puissance des tests corrobore les résultats de Zhan (2013). A l'instar de Zhan, nous obtenons cette association en calculant la puissance des tests en supposant a priori une taille d'effet petite. Mais Zhan ne retrouvait pas cette association en supposant a priori

une taille d'effet moyenne. Il est possible que son résultat contradictoire soit dû à une faible variance de la variable dépendante ou à l'utilisation d'une régression linéaire avec une variable dépendante comprise entre 0 et 1. L'association non significative entre l'utilisation d'échantillon d'étudiants et la puissance des tests corrobore les résultats de Lindsay (1993) et Borkowski *et al.* (2001). Helmuth *et al.* (2013) n'observent aussi aucune différence significative en termes de puissance des tests entre les études expérimentales et celles n'utilisant pas des designs expérimentaux. Enfin, l'association positive et significative entre la part des hypothèses unilatérales et la puissance des tests n'est guère surprenante (Cohen 1988).

En revanche, notre étude aboutit à des résultats surprenants concernant l'utilisation de données secondaires et le niveau d'analyse de l'étude. Premièrement, l'association non significative entre l'utilisation de données secondaires et la puissance des tests corrobore le résultat de Zhan (2013), mais pas ceux d'Helmuth *et al.* (2013). Helmuth *et al.* (2013) montrent que les études utilisant des données d'archive réalisent des tests ayant une plus forte puissance. Mais, Helmuth *et al.* (2013) ne distinguent pas parmi ces études celles avec un design longitudinal. Or, nos résultats montrent que la seule utilisation de base de données secondaires ou d'archive n'améliore pas significativement la puissance des tests. Le résultat d'Helmuth et de ses collègues pourrait s'expliquer par les études utilisant des données d'archive avec des designs longitudinaux. Deuxièmement, l'association négative et non significative entre le niveau d'analyse et la puissance des tests ne corrobore ni les résultats d'Helmuth *et al.* (2013), ni ceux de Zhan (2013). Zhan met en avant que les niveaux d'analyse plus élevés sont associés à une puissance des tests plus faible. En revanche, Helmuth *et al.* révélaient une puissance des tests moyenne plus élevée pour des études adoptant un niveau d'analyse élevé. Nos résultats singuliers peuvent s'expliquer par une taille d'échantillon trop faible pour déceler un petit effet du niveau d'analyse sur la puissance des tests. Ils peuvent aussi être dus au fait que les études utilisant des niveaux d'analyse élevés (pays, industrie, réseau d'entreprises, chaîne de valeur…) sont plus rares dans *Comptabilité – Contrôle – Audit* qu'en commerce international ou en management des opérations.

Limites et perspectives de recherche future

En somme, le présent article donne une explication originale d'un phénomène observé dans la littérature : la corrélation entre la puissance des tests et la date de publication des articles serait due à un accroissement de la taille des échantillons. Cet accroissement de la taille des

articles pourrait s'expliquer par un biais de publication dû à l'utilisation du niveau de signification comme critère de publication en comptabilité, contrôle et audit. Cette conclusion est paradoxale, car l'utilisation des tests de signification est décriée par ceux qui tentent de diffuser le concept de puissance statistique des tests parmi les chercheurs en comptabilité, contrôle et audit (Lindsay, 1994 ; Dyckman, 2016). Or le biais de publication dû à l'utilisation du niveau de signification comme critère de publication pourrait paradoxalement améliorer la puissance des tests en poussant les chercheurs à accroître la taille des échantillons utilisés (Jennions et Møller, 2002). Par rapport aux articles précédents, l'utilisation d'une régression bêta est aussi une contribution méthodologique. Cette méthode pourrait être utilisée dans d'autres études en comptabilité, contrôle et audit dans lesquelles la variable à expliquer est comprise entre 0 et 1, hétéroscédastique et asymétrique.

Notre étude n'est cependant pas exempte de limites qui sont autant de sources de pistes de recherches futures. Premièrement, en nous cantonnant aux seules publications dans une revue francophone de comptabilité, contrôle et audit, notre étude a une validité externe faible. Pour améliorer sa validité externe, la contrainte principale est le temps de calcul de la moyenne de la puissance des tests d'un article. Pour dépasser cette contrainte, il conviendrait de faire un sondage en grappes stratifié à plusieurs niveaux des articles en comptabilité avec une probabilité proportionnelle à la taille. L'avantage d'un tel sondage est que l'on peut obtenir une très bonne représentation de la population d'articles avec moins d'observations (de Vaus 2014). Cette solution permettrait d'étendre nos résultats à une population de revues plus large.

Deuxièmement, pour améliorer la validité interne de l'étude, il conviendrait d'inclure dans de futures études la taille de l'effet déterminée a posteriori. Même s'il s'agit d'une mesure imparfaite de la taille d'effet dans la population du fait d'erreurs d'échantillonnage, cela permettrait de voir si nos résultats sont dus à un biais de publication ou à d'autres explications. Il est possible notamment que notre observation soit due à la prise en compte par les chercheurs du concept de puissance. Cette prise en compte passe par la détermination d'une taille d'échantillon avant l'étude (calcul de la puissance a priori) pour avoir des tests suffisamment puissants pour montrer un effet d'une certaine taille. Cette explication pour la revue *Comptabilité – Contrôle – Audit* semble cependant peu plausible. Parmi les 413 articles étudiés, seuls deux mentionnent le concept de puissance statistique des tests et aucune étude empirique n'utilise un calcul de puissance a priori pour déterminer la taille d'échantillon.

Bibliographie

Bailey, C. D., Hoffman, L. L., Sloan, A. (1999). *Divulging statistical power in auditing research*. Working Paper, University of Central Florida.

Baron, R. M., Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology* 51 (6) : 1173‑1182.

Belsley, D. A., Kuh, E., Welsch, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. Hoboken : John Wiley & Sons, Inc.

Benjamini, Y., Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* 29 (4) : 1165‑1188.

Bono, J. E., McNamara, G. (2011). Publishing in AMJ--Part 2: Research Design. *Academy of Management Journal* 54 (4) : 657‑660.

Borkowski, S. C., Welsh, M. J., Zhang, Q. M. (2001). An analysis of statistical power in behavioral accounting research. *Behavioral Research in Accounting* 13 (1) : 63‑84.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology* 65 (3) : 145‑153.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. 2e édition, Hillsdale : Erlbaum.

Cribari-Neto, F., Zeileis, A. (2010). Beta regression in *R*. *Journal of Statistical Software* 34 (2) : 1‑24.

de Vaus, D. A. (2001). *Research design in social research*. Thousand Oaks : SAGE.

de Vaus, D. A. (2014). *Surveys in social research.* 6e édition, Abingdon : Routledge.

Demidenko, E. (2007). Sample size determination for logistic regression revisited. *Statistics in Medicine* 26 (18) : 3385‑3397.

Duke, G. L., Neter, J., Leitch, R. A. (1982). Power characteristics of test statistics in the auditing environment: An empirical study. *Journal of Accounting Research* 20 (1) : 42‑67.

Dyckman, T. R. (2016). Significance testing: We can do better. *Abacus* 52 (2) : 319‑342.

Ferrari, S., Cribari-Neto, F. (2004). Beta Regression for Modelling Rates and Proportions. *Journal of Applied Statistics* 31 (7) : 799‑815.

G*Power 3.1 manual (2014). Consulté à l'adresse http://www.gpower.hhu.de/fileadmin/redaktion/Fakultaeten/Mathematisch-Naturwissenschaftliche_Fakultaet/Psychologie/AAP/gpower/GPowerManual.pdf

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E. (2010). *Multivariate data analysis*. 7e édition, Upper Saddle River : Prentice Hall.

Harlow, L. L. (2005). *The essence of multivariate thinking: basic themes and methods*. Mahwah : Lawrence Erlbaum Associates.

Helmuth, C. A., Connelly, B. L., Collier, D. Y., Hanna, J. B. (2013). Power and effect size in supply chain research. *Academy of Management Proceedings* 2013 (1) : 14748‑14748.

Helmuth, C. A., Craighead, C. W., Connelly, B. L., Collier, D. Y., Hanna, J. B. (2015). Supply chain management research: Key elements of

study design and statistical testing. *Journal of Operations Management* 36 : 178‑186.

Hoenig, J. M., Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician* 55 (1) : 19‑24.

Jennions, M. D., Møller, A. P. (2002). Relationships fade with time: A meta-analysis of temporal trends in publication in ecology and evolution. *Proceedings of the Royal Society B: Biological Sciences* 269 (1486) : 43‑48.

Kinney, W. R. J. (1986). Empirical accounting research design for Ph. D. students. *The Accounting Review* 61 (2) : 338‑350.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement* 56 (5) : 746‑759.

Lee, N., Lings, I. (2008). *Doing business research: A guide to theory and practice*. Los Angeles : SAGE.

Levin, I. P. (2004). Statistical power. In *The SAGE Encyclopedia of Social Science Research Methods 2* (Eds, Lewis-Beck, M. S., Bryman, A., Futing Liao, T.). Thousand Oaks : SAGE, 1080.

Lindsay, R. (1993). Incorporating statistical power into the test of significance procedure: A methodological and empirical inquiry. *Behavioral Research in Accounting* 5 (1) : 211‑236.

Lindsay, R. (1994). Publication System Biases Associated with the Statistical Testing Paradigm. *Contemporary Accounting Research* 11 (1) : 33‑57.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks: Sage.

MacCallum, R. C., Browne, M. W., Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods* 1 (2) : 130‑149.

McSwain, D. (2004). Assessment of statistical power in contemporary accounting information systems research. *Journal of Accounting and Finance Research* 12 (7) : 100‑108.

Menard, S. W. (2002). *Longitudinal research*. 2e édition, Thousand Oaks : Sage.

Preacher, K. J., Coffman, D. L. (2006). *Computing power and minimum sample size for RMSEA [Computer software]*. Consulté à l'adresse http://quantpsy.org/

Preacher, K. J., Kelley, K. (2011). Effect size measures for mediation models: Quantitative strategies for communicating indirect effects. *Psychological Methods* 16 (2) : 93‑115.

Rigdon, E. E. (1994). Calculating degrees of freedom for a structural equation model. *Structural Equation Modeling: A Multidisciplinary Journal* 1 (3) : 274‑278.

Sedlmeier, P., Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin* 105 (2) : 309‑316.

Zhan, G. (2013). Statistical power in international business research: Study levels and data types. *International Business Review* 22 (4) : 678‑686.

Annexes

Annexe 1 : Détails sur le calcul de la puissance de certains tests

Nous avons utilisé le logiciel G*Power 3.1.9.2 pour calculer la puissance des tests mobilisant des régressions logistiques avec l'algorithme de Demidenko (2007). Pour calculer la part de la variance de chaque variable explicative d'un article expliquée par les autres variables explicatives et de contrôle, nous avons utilisé la matrice de corrélation. Dans certains articles, cette matrice n'était pas disponible. Nous avons contacté les auteurs des articles. Mais ils ne nous ont renvoyé aucune matrice de corrélation. Cela ne nous a pas permis d'intégrer trois articles utilisant une ou des régressions logistiques. De plus, un autre article comprend une matrice de corrélation renvoyant des coefficients de détermination supérieurs à l'unité. Nous n'avons pas intégré cet article. Enfin, nous avons retenu comme distribution des variables explicatives pour les variables continues ou ordinales une loi normale, si et seulement si les statistiques descriptives et les tests de normalité de l'article permettaient de faire cette hypothèse, sinon une loi uniforme. Pour les variables dichotomiques, nous avons retenu une distribution binomiale. Nous avons utilisé les statistiques descriptives pour avoir des estimations des paramètres de ces lois.

Pour les articles utilisant des équations structurelles pour tester leurs hypothèses, nous avons utilisé l'algorithme de McCallum *et al.* (1996) programmé dans le logiciel R (Preacher et Coffman 2006). Nous avons calculé le degré de liberté suivant Ridgon (1994). Cela fut notamment utile pour un article pour lequel le degré de liberté indiqué n'était pas exact. Pour calculer la puissance des tests des articles utilisant des équations structurelles, nous avons fixé le RMSEA à 0,05, sous l'hypothèse nulle, et à 0,05, 0,08 et 0,1, pour respectivement un effet petit, moyen et grand (Harlow 2005, p. 242).

Enfin, pour les articles utilisant un ou des tests de Wilcoxon-Mann-Whitney, nous avons calculé pour chaque test la moyenne des résultats renvoyés par le logiciel G*Power 3.1.9.2. Les statistiques descriptives ne permettant pas privilégier une distribution, nous avons déterminé avec le logiciel G*Power 3.1.9.2 la puissance pour chaque test en retenant une distribution normale, logistique, de Laplace et « Min ARE ». Puis, pour chaque test et chaque taille d'effet a priori (petite, moyenne et grande), nous avons calculé la moyenne arithmétique simple des quatre résultats renvoyés par le programme G*Power 3.1.9.2.

Annexe 2 : Fondements théoriques et empiriques du signe attendu pour l'association entre les variables de contrôle et la puissance des tests

La méthode d'analyse retenue peut avoir une incidence sur la puissance calculée des tests (McSwain 2004), surtout s'il s'agit d'une méthode pour laquelle les algorithmes utilisés pour calculer la puissance des tests sont plus récents et peut-être moins fiables (car laissant plus de discrétion à l'utilisateur). Pour les équations structurelles, généralement exclues des analyses de puissance précédentes (Helmuth *et al.* 2015), il est possible d'utiliser l'algorithme de McCallum *et al.* (1996). Pour les régressions logistiques, il existe plusieurs algorithmes. Celui de Demidenko (2007) parait être le plus pertinent (G*Power 3.1 manual 2014). Cela justifie que l'on rajoute à notre modèle deux variables dichotomiques (SEM et LOGIT, tableau 3) pour identifier les articles ayant utilisé des équations structurelles ou des régressions logistiques.

Le caractère international, le design de l'étude et l'origine des données sont des facteurs jouant sur la taille de l'échantillon. Le fait de prendre un échantillon de cas sur une base internationale (INTER) permet d'avoir un plus grand nombre de cas potentiels (Zhan 2013). Mais les différences culturelles et institutionnelles et la disponibilité des données augmenteraient l'effort de collecte (Helmuth *et al.* 2013). Ainsi, l'effet de l'utilisation d'un échantillon international sur la puissance des tests est indéterminé.

Lorsqu'on utilise un design longitudinal (LONGI), on rassemble des données collectées sur les mêmes cas (panel) ou des cas comparables (quasi-longitudinal) sur plusieurs années (Menard 2002). La puissance des tests augmente lorsque les données collectées sur plusieurs années sont analysées simultanément (Menard 2002, p. 65). Nous nous attendions donc à trouver une association positive entre la puissance des tests et l'utilisation d'un design longitudinal, à l'instar de Zhan (2013).

Selon Borkowski *et al.* (2001), les designs expérimentaux ayant pour sujets des étudiants (STUDENT) permettraient a priori d'avoir des études avec plus de répondants du fait de la facilité d'accès à ces répondants pour le chercheur. Dans les faits, la puissance des tests est plus faible pour les études utilisant des designs expérimentaux (Lindsay 1993) et des échantillons d'étudiants (Borkowski *et al.* 2001). Helmuth *et al.* (2013, 2015) obtiennent des résultats similaires pour les études en management de la chaîne de valeur. Du fait de cette contradiction entre l'effet a priori et les résultats d'études empiriques, le signe de l'association

entre la variable dichotomique indiquant l'utilisation d'un échantillon d'étudiants et la puissance des tests est donc indéterminé.

Enfin, l'analyse de puissance de Zhan (2013) considère que l'origine des données peut être un facteur influençant la puissance moyenne des tests des articles. D'une part, les études utilisant des bases de données secondaires (DATA) donnent généralement lieu à des analyses avec plus d'observations. L'utilisation de données secondaires peut ainsi avoir un effet sur la puissance des tests. Eu égard aux résultats contradictoires à ce sujet (Helmuth *et al.* 2015 ; Zhan 2013), nous avons considéré indéterminé le signe de l'association entre l'utilisation de données secondaires et la puissance des tests. D'autre part, l'unité d'analyse (UNIT_ANAL) peut influencer la puissance des tests. En commerce international, Zhan (2013) met en avant que les études ayant des unités d'analyse dont la population est plus restreinte (pays, industrie) ont des tests moins puissants. Mais Helmuth *et al.* (2013, 2015) obtiennent des résultats contradictoires en management de la chaîne de valeur. Dans le champ de la comptabilité, du contrôle et de l'audit, nous ne pouvons donc prédire le signe de l'association entre la puissance des tests et le niveau d'analyse.

A l'instar de Zhan (2013), nous avons aussi choisi de ne pas calculer la puissance des tests de la même façon suivant le type de test. Les tests unilatéraux ayant plus de puissance que les tests bilatéraux (Cohen 1988), nous avons donc insérer une variable (TEST_UNILAT) prenant en compte la part des tests unilatéraux. Nous attendions donc une association positive entre la part des hypothèses unilatérales et la puissance des tests d'un article.

Annexe 3 : Codage des articles pour détecter un biais de publication dû à l'utilisation de la signification statistique comme critère de publication

Pour évaluer la plausibilité de la proposition 2 portant sur l'existence d'un biais de publication dû à l'utilisation de la signification statistique comme critère de publication, nous avons adapté la procédure de Lindsay (1994). Parmi les 118 articles empiriques utilisant la statistique inférentielle publiés dans *Comptabilité–Contrôle–Audit*, vingt-six papiers ont été retirés car ils ne posaient pas explicitement des hypothèses ou propositions. Ainsi, la procédure décrite n'a été utilisée que sur 92 articles, contre 38 articles dans une analyse proche conduite par Lindsay (1994).

1. Seules les hypothèses centrales, celles explicitement formulées d'après la littérature, ont été considérées. Les hypothèses auxiliaires et celles associées aux variables de contrôle n'ont donc pas été prises en compte. Pour être compté parmi les articles dont la majorité des hypothèses nulles principales sont rejetées, plus de la moitié des hypothèses nulles centrales de l'article devaient être rejetées au seuil de signification adopté par le (ou les) auteur(s).
2. Pour déterminer cela pour chaque article, un score total X a été calculé. Pour chaque article, nous avons d'abord étudié la discussion par l'auteur des résultats. Une hypothèse partiellement corroborée rajoutait 0,5 au score total de l'article. Mais lorsqu'environ un quart des résultats corroborait une hypothèse associée, nous n'augmentions que de 0,25 ce score total, contre 0,75 lorsque presque tous les résultats associées rejetaient l'hypothèse nulle. Nous ajoutons 1 au score total de l'article lorsqu'une hypothèse alternative était corroborée par tous les résultats associés (ou lorsqu'une hypothèse non directionnelle ou nulle était rejetée par tous les résultats associés).
3. Le score total X de l'article a ensuite été divisé par le nombre d'hypothèses explicitement formulées n dans cet article. Si le résultat de cette division X/n égalait ou dépassait 0,5 l'article était classé comme « positif ». Sinon, l'article était classé comme « négatif ». Les articles positifs, dont la majorité des hypothèses nulles centrales sont rejetées, ont été ensuite comptés pour faire un test binomial.

Annexe 4 : Analyse de sensibilité et test de robustesse

Premièrement, pour tester la présence d'observations influentes, nous avons calculé les distances de Cook. Après retrait de six observations influentes (en utilisant un seuil de 0,04), l'association entre la date de publication et la puissance des tests reste statistiquement non-significative au seuil de 5 % dans le modèle 1.2 (0,002 ; $p = 0,073$). Concernant les variables de contrôle, les associations entre la puissance des tests, et l'utilisation d'échantillon d'étudiants (0,881 ; $p = 0,016$) et de tests unilatéraux (0,347 ; $p = 0,044$) sont positives et statistiquement significatives au seuil de 5 %. Les autres résultats sont inchangés.

Deuxièmement, l'utilisation de variables de contrôle peut parfois impacter les résultats. L'estimation des modèles 1.1 et 1.2 présentés dans le tableau 6 en l'absence de variables de contrôle donnent des résultats davantage en faveur de notre première hypothèse alternative.

L’association entre la date de publication et la puissance des tests des articles est positive et significative sans (0,004 ; $p = 0,01$) et avec (0,003 ; $p = 0,05$) contrôle statistique de la taille de l’échantillon. Mais l’on observe toujours que l’inclusion dans le modèle (1) de la taille de l’échantillon réduit la signification statistique et l’importance pratique de l’association entre la date de publication et la puissance des tests des articles.

Troisièmement, la sélection d’une fonction de lien pertinente peut améliorer significativement l’ajustement (Cribari-Neto et Zeileis 2010). Cela est d’autant plus plausible dans notre étude que certaines estimations de la puissance des tests sont proches de l’unité. Nous avons donc estimé le modèle 1.2 en utilisant d’autres fonctions de lien disponibles dans le paquet *betareg* du logiciel R. Le critère du pseudo-R^2 indique que l’utilisation d’autres fonctions n’améliore pas la qualité de l’ajustement.

Enfin, concernant l’existence d’un biais de publication dû à l’utilisation de la signification statistique comme critère de publication, nous avons divisé la période en deux sous-périodes. Le taux de 66 % d’articles rejetant au moins 50 % de leurs hypothèses nulles centrales semble stable. Sur la sous-période allant de janvier 2006 à avril 2016, 36 (67 %) articles sur 54 rejettent la plupart de leurs hypothèses nulles centrales, contre 25 (66 %) sur 38 articles avant 2006.