L'Année psychologique, 2005, *105*, 669-696

NOTE MÉTHODOLOGIQUE

Laboratoire de Psychologie Sociale et de Psychologie Cognitive, LAPSCO/CNRS, Université Blaise Pascal[1]

FAUT-IL PRÉFÉRER L'ANALYSE DE VARIANCE À L'ANALYSE DE RÉGRESSION DANS LES EXPÉRIENCES UTILISANT DES VI CONTINUES ?

Alain MÉOT, Patrick BONIN

SUMMARY : Should we use analysis of variance rather than regression analysis for experiences using continuous independent variables ?

In this paper, we discuss the two possible statistical approaches when IV are continuous : analysis of variance and multiple linear regression. The statistical model common to these two approaches, and analysis of covariance is first briefly reminded. Using general arguments, followed by a psycholinguistic illustration, we discuss the advantages and drawbacks of using an experimental control of confoundings and random variability, which leads to analysis of variance, or a statistical control, which leads to multiple regression analysis. The problems to avoid for insuring the validity of the chosen approach are also underlined.

Key words : analysis of variance, multiple linear regression, experimental control, statistical control, power, validity.

INTRODUCTION

Selon Winer, Brown et Michels (1991), « la nature est caractérisée par la variation. L'idéal expérimental est la construction d'une situation dans laquelle *toute* la variation de la variable dépendante (VD) est systématiquement reliée aux manipulations expérimentales d'une ou plusieurs variables

1. 34, avenue Carnot, 63037 Clermont-Ferrand. Adresse e-mail : meot@srvpsy.univ-bpclermont.fr.

indépendantes (VI). Comme cet idéal ne peut être atteint, un plan expérimental vise deux objectifs : 1 / conduire à des résultats pour lesquels la variation systématique peut être attribuée seulement aux effets des traitements ; 2 / réduire l'erreur aléatoire, non systématique, le plus possible afin d'obtenir la meilleure estimation possible de ces effets. Les effets expérimentaux sont compréhensibles seulement lorsqu'ils sont relativement importants eu égard à leur arrière-fond bruité ». Les propositions de Winer *et al.* (1991) indiquent donc ce vers quoi doit tendre tout plan expérimental afin que les inférences statistiques soient fiables et valides.

En psychologie expérimentale, cet idéal a longtemps conduit au recours à des expériences factorielles, pour lesquelles les traitements statistiques relèvent de l'analyse de variance (ANOVA). Cependant, en raison de la multiplicité des facteurs qui peuvent potentiellement exercer un impact sur le comportement étudié, l'approche factorielle soulève des problèmes qui relèvent parfois de la gageure. L'expérimentateur se trouve souvent confronté à la difficulté de tenir compte d'un vaste ensemble de variables potentiellement influentes eu égard aux caractéristiques mesurées. De plus, les échelles de mesure des VI sont soit hétérogènes, c'est-à-dire que des variables continues et catégorielles sont mêlées, soit toutes continues, ce qui, dans les deux cas, soulève des problèmes quant à leur intégration dans une ANOVA (voir plus loin).

Certaines thématiques sont plus particulièrement concernées par ce type de difficultés. Par exemple, un nombre élevé d'études en psychologie sociale ne peuvent se contenter du caractère « frustre » de l'ANOVA, en raison du croisement important de facteurs sociaux et individuels qu'elles prennent en compte. Également, certaines problématiques de psycholinguistique, qui nous serviront d'illustration, doivent nécessairement intégrer de multiples caractéristiques. Ces caractéristiques sont destinées à opérationaliser les processus sous-jacents et à prendre en compte l'hétérogénéïté des dimensions (orthographique, phonologique, sémantique, etc.) du matériel utilisé. Ces caractéristiques sont en général opérationalisées par des VI continues.

Dans les situations qui intègrent des VI continues, deux procédures d'analyses statistiques alternatives à l'ANOVA sont utilisables :

1 / l'analyse de covariance (ANCOVA) lorsque interviennent aussi des VI catégorielles ;
2 / la régression linéaire multiple (RLM) lorsque toutes les VI sont continues.

Dans les deux cas, il s'agit de remplacer le contrôle expérimental des confusions d'effets, et de la variabilité associée à l'erreur expérimentale, par un contrôle statistique.

Si l'utilisation de l'ANCOVA ne soulève guère de protestations au sein de la communauté des chercheurs en psychologie, il en va tout autrement de la RLM. Cette dernière est souvent suspectée de n'être pas réellement efficace dans sa capacité à détecter les effets attendus ou, autrement dit, de manquer de puissance. Même si les critiques sur la régression multiple sont

rarement formulées explicitement dans des publications en psychologie, elles transparaissent dans le processus d'expertise par pairs d'articles. En effet, très souvent, les experts conseillent soit de compléter une approche en RLM par une autre en ANOVA, soit d'éviter d'utiliser la RLM et de préférer une approche en ANOVA.

La suspicion à l'égard de la RLM semble tenir essentiellement à deux raisons.

Premièrement, l'histoire et les pratiques des champs considérés. La RLM provient de la biologie et du versant « prédictif » des sciences du comportement, comme la prédiction d'une performance en fonction des résultats à des tests. Il s'agissait d'étudier la covariation entre des caractéristiques invoquées qui proviennent d'échantillons issus de populations naturelles (c'est-à-dire où les unités d'observation ne sont pas affectées aléatoirement aux conditions expérimentales). L'ANOVA et l'ANCOVA sont apparues plus tardivement dans le champ agronomique, avec comme objectif, le contrôle de la variabilité dans des expériences où des VI sont manipulées. Les deux approches se sont développées en parallèle et largement indépendamment. Cependant, la RLM, étant donné son association initiale avec des données observationnelles (non expérimentales), a très rapidement été perçue comme moins « noble » que l'ANOVA.

Deuxièmement, le contrôle statistique des confusions d'effets et de l'erreur expérimentale en RLM, pose un certain nombre de problèmes. Ces problèmes sont la contrepartie statistique des caractéristiques interreliées dont les effets sont étudiés. Le problème le plus souvent soulevé est relatif à une perte de puissance de la quantification des effets. En effet, le processus de partiallisation utilisé en RLM consiste à quantifier la part de variance expliquée par une VI comme la différence entre le R2 obtenu lorsque celle-ci figure dans l'équation, et celui obtenu, lorsqu'elle n'est pas dans l'équation. Est ainsi éliminée de la variance expliquée par une VI, toute la part d'explication commune à celle-ci et aux autres VI. Lorsque cette VI est fortement corrélée avec d'autres VI, la mesure de la part de variance qu'elle explique, à elle seule, est donc « pessimiste » et les tests sont plus conservateurs. L'énoncé souvent critique de ce problème, et la généralisation parfois un peu rapide de propriétés négatives de certaines variantes de RLM (comme la régression « pas à pas » fortement critiquée par Morris en 1981), a conduit à renforcer la suspicion à l'égard de la RLM.

La méfiance vis-à-vis de la RLM est sans doute aussi due au fait que ces deux approches sont en général présentées de façon séparée, alors qu'elles peuvent se formaliser dans un même modèle statistique : le modèle linéaire général. La mise en regard de l'ANOVA avec la RLM, par l'intermédiaire du modèle linéaire, permet de mieux saisir la proximité élevée des deux approches et, également, leurs différences. Nous rappelons, dans une première partie, en quoi consiste le modèle linéaire et la manière dont l'ANOVA peut s'y insérer. Ce faisant, nous montrons qu'en dehors des problèmes liés à la puissance, la suspicion manifestée par certains psychologues à l'égard du recours

à la RLM pour le traitement statistique des données expérimentales n'est pas fondée. Dans une deuxième partie, nous argumentons qu'opérer un contrôle expérimental des confusions d'effets et de l'erreur soulève, dans de nombreux cas, aussi bien des problèmes de puissance que de validité, et que la perte de puissance due au processus de partiallisation, est un handicap qui est souvent très relatif en RLM. Nous y discutons aussi de questions que soulèvent plus spécifiquement l'une ou l'autre méthode. Enfin, dans une dernière partie, nous montrons, en nous appuyant sur des études de psycholinguistique, que les problèmes rencontrés dans l'une des approches font écho à ceux rencontrés dans l'autre. Nous tenterons de convaincre les lecteurs qu'aucune des deux méthodes ne peut prétendre à une quelconque suprématie. Seule la nature des observations peut permettre de trancher sur la question de la pertinence de celle à utiliser pour une question expérimentale donnée.

I. LE MODÈLE LINÉAIRE GÉNÉRAL : UN CADRE STATISTIQUE COMMUN AUX ANALYSES DE VARIANCE, DE COVARIANCE ET DE RÉGRESSION LINÉAIRE MULTIPLE

I.1. MODÈLE LINÉAIRE GÉNÉRAL / RÉGRESSION LINÉAIRE MULTIPLE[1]

Le modèle linéaire général stipule qu'il existe une relation linéaire entre une variable dépendante Y quantitative et p variables indépendantes, elles-mêmes quantitatives X_1, ..., X_p :

$$Y = \alpha + \beta_1 X_1 + \ldots + \beta_p X_p$$

où α, β_1, ..., β_p sont les coefficients partiels de l'équation. L'interprétation de ceux-ci est la suivante : puisque pour un ensemble donné de valeurs x_1, ... x_p des variables indépendantes, nous avons :

$$y/x_1, \ldots, x_p = \alpha + \beta_1 x_1 + \ldots + \beta_p x_p$$

Si nous augmentons d'une unité la valeur d'une des variables indépendantes, par exemple en fixant X_1 à la valeur $x_1 + 1$, les autres valeurs restant constantes, nous obtenons :

$$\begin{aligned} y/x_1 + 1, \ldots, x_p &= \alpha + \beta_1(x_1 + 1) + \ldots + \beta_p x_p \\ &= \alpha + \beta_1 x_1 + \ldots + \beta_p x_p + \beta_1 \\ &= y/x_1, \ldots, x_p + \beta_1 : \end{aligned}$$

1. Modèle linéaire et RLM sont en fait identiques. Cependant différencier les deux permet de conserver une spécificité à la RLM, qui peut être conçue alors comme un modèle linéaire dans le seul cas où l'échelle des VI est continue. Notons par ailleurs que nous ne discutons que du modèle fixe de RLM.

À l'augmentation d'une unité pour l'une quelconque des VI (ici la première), les autres VI étant fixées à des niveaux constants quelconques, correspond donc une augmentation de β (ici β_1) unités de la variable dépendante. Les coefficients β peuvent donc être considérés comme les effets des VI.

Cette interprétation des coefficients peut expliquer le succès que rencontre la RLM. Notamment, en psychologie expérimentale, la question de la qualité structurale du modèle (est-ce la bonne relation VD/VI qui est employée ?) se pose peu au regard des questions relatives aux erreurs d'opérationalisation ou de mesure.

Ces premières formules permettent de comprendre le problème essentiel à l'origine de la suspicion à l'égard de la RLM : la dépendance ou multicolinéarité *partielle* entre les VI.

La multicolinéarité *stricte* correspond au fait qu'une quelconque des VI, par exemple X_j, s'écrit comme modèle linéaire de l'ensemble des autres VI, c'est-à-dire qu'il existe p coefficients ($\delta, \gamma_1, \ldots, \gamma_{j-1}, \gamma_{j+1}, \ldots, \gamma_p$) tels que :

$$X_j = \delta + \gamma_1 X_1 + \ldots + \gamma_{j-1} X_{j-1} + \gamma_{j+1} X_{j+1} + \ldots + \gamma_p X_p$$

La conséquence d'une telle relation est qu'il est alors impossible de considérer l'effet d'une VI lorsque les autres VI sont fixées à des niveaux quelconques. En effet, si on inverse la relation, alors lorsque X_j augmente d'une unité, la valeur de toute autre VI augmente obligatoirement de l'inverse du coefficient lui étant affecté, de sorte qu'il n'est pas possible de faire varier les valeurs d'une variable sans que les valeurs des autres n'en soient affectées. Sur le plan technique, une relation de multicollinéarité stricte conduit à une infinité de solutions pour les coefficients de la relation linéaire entre la VD et l'ensemble des VI.

En pratique, la multicolinéarité *stricte* est très rarement vérifiée. Et même lorsqu'elle existe au niveau populationnel, la variabilité d'échantillonnage fait que ce type de relation peut difficilement être observé au niveau des échantillons. La dépendance ou la multicolinéarité est donc le plus souvent *« partielle »* : sans que la dépendance soit parfaite, une relation linéaire se trouve être un modèle satisfaisant des relations qui existent entre une des VI et certaines des autres VI. Dans ce cas, le processus de partialisation utilisé en RLM conduit à une vision conservatrice de la variance de la VD expliquée par cette VI. La puissance des tests utilisés s'en trouve donc diminuée.

Il faut cependant noter que, sauf dans les cas exceptionnels où l'on s'approche d'une stricte multicolinéarité, la RLM peut toujours être utilisée. En effet, les propriétés statistiques du modèle restent les mêmes, et sont optimales en ce qui concerne l'efficacité des estimateurs (dans la terminologie anglo-saxonne, ceux-ci sont « BLUE », pour *Best Linear Unbiaised Estimator*).

Le modèle présenté ci-avant est supposé correspondre aux vraies relations entre VD et VI. Celui-ci étant cependant par nature inconnu, l'utilisation d'échantillons pour en assurer l'estimation nécessite d'y ajouter un « étage aléatoire », à savoir que les mesures réalisées à partir d'échantillons se conforment au modèle à une quantité aléatoire additive près, le résidu (ou l'erreur), égale à un effet résultant d'un grand nombre de causes non identifiées et non systématiques. La relation linéaire est alors supposée vraie en moyenne, sur un grand nombre d'unités d'observation présentant les mêmes valeurs sur les VI, ce qui revient à supposer qu'en moyenne (en espérance), les perturbations aléatoires sont nulles. Ce qui donne l'expression classique du modèle de RLM :

$$E(Y / x_1, \ldots, x_p) = \alpha + \beta_1 x_1 + \ldots + \beta_p x_p$$

qui signifie qu'en moyenne (espérance), sur un grand nombre de mesures sur des unités d'observation présentant le même jeu de valeurs des VI, la relation linéaire est respectée. Cette relation constitue le « cœur logique » du modèle statistique utilisé. Enfin, pour assurer de « bonnes » qualités aux estimateurs obtenus à partir de l'ajustement des moindres carrés, deux hypothèses techniques doivent être ajoutées : 1 / l'égalité des variances des résidus et l'indépendance de ceux-ci et 2 / une distribution gaussienne des résidus qui permet d'utiliser des tests d'hypothèses et de construire des intervalles de confiance des coefficients.

I.2. L'ANALYSE DE VARIANCE

Pour rester simple, nous ne considérons que le cas de plans fixes à deux variables indépendantes A à p modalités indicées par un i et B à q modalités indicées par un j. L'extension à des plans plus complexes est relativement évidente dans le cas de plans fixes. Elle nécessite des adaptations dans le cas de plans mixtes ou totalement aléatoires. L'expression la plus commune du modèle d'ANOVA pour un plan à deux VI fixes avec interaction est :

$$E(Y_{ijk}) = \mu(ij) = \mu + \alpha(i) + \beta(j) + \tau(ij)$$

où $\mu(ij)$ correspond à la moyenne dans les conditions i et j et k correspond à la k-ième unité d'observation de la condition (i, j).

Ce modèle spécifie donc que la moyenne, dans une des conditions, est égale à la somme des divers effets. Il recouvre le même type de relation que le modèle de la régression multiple, c'est-à-dire une relation linéaire entre la VD et les VI. Autrement dit, en moyenne, le score de la VD, dans une combinaison de deux niveaux des VI, est le résultat d'une combinaison additive d'un effet constant, des deux effets principaux et d'un effet « conjoint » des deux VI.

L'explicitation de cette relation dans les termes RLM se formalise dans la relation :

$$Y = \mu * C + \alpha(1)\, A_1 + \ldots + \alpha(p)\, A_p + \beta(1)\, B_1 + \ldots + \beta(q)\, B_q + \tau(11)\, AB_{11} + \ldots + \tau(pq)\, AB_{pq} :$$

où :

— C est une « VI » composée uniquement de 1.

— A_i (B_j) est une VI composée de 0 et de 1 où le 1 traduit le fait qu'une mesure provient d'une unité d'observation appartenant à la *i*-ème (*j*-ème) condition définie par A (B), et 0 est associé à une mesure provenant d'une unité d'observation n'appartenant pas à cette condition.

— AB_{ij} est une VI composée de 1 et de 0 où le 1 traduit le fait qu'une mesure provient d'une unité d'observation appartenant à la fois à la la *i*-ème condition définie par A et à la *j*-ème condition définie par B, et 0 est associé à une mesure provenant d'une unité d'observation n'appartenant pas à cette combinaison de conditions[1].

Il apparaît donc que l'ANOVA est strictement équivalente à la définition d'une relation linéaire entre la VD Y et 1 (pour C) + *p* (pour les A_i ; $i = 1, \ldots p$) + *q* (pour les B_j ; $j = 1, \ldots q$) + *pq* (pour les AB_{ij} ; $i = 1, \ldots, p$; $j = 1, \ldots, q$) VI.

Cependant, ce modèle linéaire est problématique du fait de la présence de strictes multicolinéarités entre les VI (par exemple, $A_1 + \ldots + A_p = C$). Comme une VI peut s'écrire comme modèle linéaire des autres, il est impossible d'en estimer l'effet. Ce problème est en général résolu en imposant des restrictions, c'est-à-dire en imposant des relations sur certains des paramètres, rendant ainsi les autres paramètres libres. La conséquence en est l'élimination pure et simple de certaines des VI. Les deux restrictions les plus courantes sont les suivantes (mais voir Cohen et Cohen, 1983, pour d'autres exemples) :

(1) le codage factice (*« dummy coding »*) revient à éliminer $p + q + 1$ des VI de la manière suivante. Des conditions *a* de A et *b* de B étant choisies (classiquement les numéros *p* et *q*), les VI A_a et B_b sont éliminées. Les VI traduisant l'interaction correspondent aux produits terme à terme des VI A_i et B_j restantes prises deux à deux.

(2) le codage d'effet (*« effect coding »*) opère de la même manière jusqu'à élimination des VI A_a et B_b. De plus, pour chaque VI restante A_i (B_j), les observations relevant de la condition *a* (*b*) sont codées – 1. Enfin, ici aussi, les VI traduisant l'interaction sont obtenues par les produits terme à terme des VI A_i et B_j restantes prises deux à deux.

1. Howell (1998) donne une présentation beaucoup plus explicite de cet aspect. Une image « tableur » permet de bien se représenter la relation : Y correspond à la colonne « VD » de celui-ci ; C est une colonne de 1, A_i, B_j et AB_{ij} des colonnes composées de 0 et 1 où les valeurs 0 et 1 sont définies comme précisé dans le texte.

Il faut souligner que la question des restrictions sur les paramètres du modèle utilisé en RLM n'est que l'écho du surnombre de paramètres présents dans le modèle d'ANOVA. La présence de restrictions est en fait commune aux deux approches et ne rend donc pas le modèle utilisé en RLM plus arbitraire que celui d'ANOVA.

Notons aussi qu'un autre type de restrictions, particulièrement utile, consiste à utiliser des ensembles de contrastes. Ces contrastes sont traduits dans la RLM sur la base de VI dont les coefficients correspondent à ceux des contrastes (voir Cohen et Cohen, 1983). À cet endroit, la RLM prouve tout son intérêt par rapport à l'ANOVA. En effet, dans son expression classique, cette dernière ne peut prétendre, si ce n'est au prix d'une grande complexité, traiter autre chose que des contrastes orthogonaux. En effet, elle ne permet pas de contrôler ce qui, dans la part de variance expliquée par un contraste particulier, provient en fait de ses relations avec d'autres contrastes. Utiliser des contrastes non orthogonaux dans le cadre ANOVA donnera donc en général une vision « optimiste » des différences reflétées par ceux-ci. Comme la RLM a pour objectif de partialliser de l'effet d'une des VI ce qui en fait peut être attribué à d'autres VI, elle conduit « naturellement » à un tel contrôle. Cette possibilité offerte de contrôler les confusions d'effets se retrouve d'ailleurs aussi au niveau des tests d'effets (principaux, d'interaction) lorsque les plans ne sont pas équilibrés. Cet aspect explique d'ailleurs pourquoi la RLM est utilisée en fait en arrière fond des modules d'ANOVA à plus d'un facteur des logiciels statistiques.

Il est important de préciser à cet endroit que la réalisation d'une ANOVA par l'intermédiaire d'une procédure de régression multiple, bien que tout à fait possible, n'est pas particulièrement pratique, notamment lorsque certaines VI sont intra-participants. Pour se convaincre de cet aspect, on pourra consulter l'ouvrage de Keppel et Zedeck (1989) qui décrit comment aborder une vaste variété de plans factoriels, ainsi que les contrastes, par ce moyen.

Rappelons pour finir que les suppositions aléatoires sont strictement les mêmes qu'en RLM : indépendance, égalité des variances des résidus et distribution normale de ceux-ci centrée en 0.

I.3. L'ANALYSE DE COVARIANCE OU LE LIEN ÉVIDENT DE L'ANOVA AVEC LA RÉGRESSION MULTIPLE

La forme la plus simple, qui correspond aussi à la définition classique de l'analyse de covariance (ANCOVA), met en jeu une VI qualitative A et une VI quantitative X à travers la relation présupposée :

$$E(Y / ik, x) = \mu + \alpha(i) + \beta x$$

où l'indice i désigne la i-ème condition de A, x est une valeur fixée de X et k est la k-ième mesure dans la condition i.

Nous retrouvons clairement un modèle linéaire, cette fois entre la VD et les deux VI. Cependant, celui-ci présente à la fois un aspect ANOVA et un

aspect RLM. Si on combine les deux présentations précédentes, on se ramène au modèle de régression multiple suivant :

$$Y = \mu * C + \alpha(1)\, A_1 + \ldots + \alpha(p)\, A_p + \beta X :$$

Un problème de surnombre de paramètres à estimer est aussi présent, problème qui se traduit, par exemple, par la présence d'une multicolinéarité stricte entre les VI associées à A et à la variable constante. Ce problème peut à nouveau être résolu, soit par la définition d'une restriction sur les paramètres, soit par l'utilisation de $p - 1$ contrastes en lieu et place de ces VI. Notons à titre anecdotique qu'une approche plus générale, basée sur des inverses généralisés, peut aussi être utilisée (*e.g.* Graybill, 1976).

Le modèle linéaire est ici beaucoup plus explicite qu'en ANOVA. En effet, cette dernière est en général présentée à travers des décompositions de la variance, puis de comparaison de rapports de carrés moyens ayant des espérances communes sous H0. Autrement dit, la présentation de l'ANOVA fait généralement abstraction du « côté » régression. Cela n'est guère possible pour l'ANCOVA. D'une part cette dernière fait clairement apparaître la régression sur X ; d'autre part, la résolution du problème de l'estimation des paramètres peut difficilement faire abstraction d'une présentation par les moindres carrés, ce qui la plonge directement dans la RLM.

I.4. EN RÉSUMÉ : UN MODÈLE STATISTIQUE UNIQUE RECOUVRANT RÉGRESSION LINÉAIRE MULTIPLE, ANALYSES DE VARIANCE ET DE COVARIANCE

Face à des VI à la fois catégorielles et continues, de nombreux travaux cherchent à se rapprocher le plus possible du cadre de l'ANOVA. Pour ce faire, divers procédés sont utilisés avant ou après l'expérience pour transformer les VI continues en VI catégorielles. Cependant, comme développé ci-avant, ANOVA, ANCOVA et RLM peuvent être formalisées à l'aide du modèle linéaire, ce qui conduit à utiliser les mêmes méthodes d'estimation et suppositions statistiques. L'existence d'un modèle commun souligne donc que la plupart des questions rencontrées dans l'une des analyses se rencontre aussi dans les autres. Ces problèmes sont généralement de nature pratique. Nombre d'entre eux conduisent cependant à une inadéquation entre certaines des propriétés du modèle postulé (ANOVA ou RLM) et les observations, ce qui contribue à réduire la validité statistique des résultats obtenus. Pour différencier ces deux aspects dans la suite du texte, nous utiliserons pour chaque problème soulevé la notation (*p*) pour désigner son aspect pratique, et celle (*t*) pour ses éventuelles conséquences statistiques. En ce qui concerne les problèmes communs que peuvent rencontrer les deux approches, peuvent notamment apparaître : 1 / Des questions de spécification du modèle : (*p*) n'a-t-on pas oublié de faire figurer dans le modèle des VI importantes qui peuvent avoir un effet potentiel sur la VD ? (*t*) Dans

ce cas, les estimations des divers coefficients sont biaisées, et manquent donc de validité. (*p*) Au contraire, n'a-t-on pas inclus dans le modèle un ensemble trop important de VI, c'est-à-dire des VI qui n'ont, en réalité, pas d'effets sur la VD ? (*t*) Dans ce cas, les relations entre ces VI vont conduire à un manque de précision des estimations pour les VI qui ont un intérêt véritable dans la problématique considérée. Cela se traduit par une inflation de la variabilité des estimations, et donc, par des tests beaucoup plus conservateurs. 2 / Des questions de validité structurale du modèle : (*p*) une relation linéaire est-elle bien justifiée ? La supposition d'additivité a-t-elle un sens ? Certains termes du modèle ne devraient-ils pas faire apparaître des fonctionnelles différentes des VI ? (*t*) Ces aspects peuvent avoir les mêmes incidences qu'une erreur de spécification par omission de VI (par exemple la présence d'estimations biaisées du fait de l'omission d'un terme quadratique), ou, plus grave encore (*p*) par une interprétation totalement inconsistante de certains termes (par exemple constance d'un effet en cas d'oubli d'un terme d'interaction) ; 3 / des questions de validité statistique qui tiennent à l'estimation par les moindres carrés et qui conduisent aux épineux problèmes concernant l'étude des valeurs influentes (*e.g.* Chatterjee et Hadi, 1986), (*t*) ces valeurs « tirant à elles » les estimations, certaines de celles-ci peuvent s'avérer parfaitement farfelues ; 4 / des questions de validité statistique dues à une mauvaise adéquation entre les suppositions nécessaires à la validité des tests d'hypothèses (ou des estimations par intervalles de confiance) et les observations réalisées : problèmes d'hétérodéscadisticité, de non-indépendance qui conduisent (*t*) essentiellement à une estimation incorrecte de la variabilité aléatoire ; problème de non-normalité des résidus qui (*t*) pour les petits échantillons conduit à utiliser des distributions d'échantillonnage inadéquates pour réaliser les tests d'hypothèses ou construire des intervalles de confiance.

II. CONTRÔLE DES CONFUSIONS D'EFFETS ET DE LA VARIABILITÉ ALÉATOIRE : EXPÉRIMENTAL OU STATISTIQUE ?

Souvenons-nous des recommandations de Winer *et al.* (1991) selon lesquelles la mise en évidence d'effets nécessite de réaliser des expériences tendant au mieux à ce que 1 / la variation systématique soit attribuée seulement aux effets des traitements et 2 / l'erreur aléatoire, non systématique, soit la plus réduite possible. Parler d'une erreur aléatoire, non systématique, signifie que toute source de variation systématique qui affecte la VD doit être intégrée explicitement dans l'analyse. Le fait de ne pas prendre en compte une telle source de variation conduit sur le plan statistique, à la fois à des estimations biaisées, c'est-à-dire incorrectes en moyenne sur un grand nombre de répétitions de l'expérience, et à un manque de puissance de

l'expérience réalisée. Afin d'obtenir une mesure de la variation systématique qui ne tienne qu'aux *seuls* effets des traitements, il faut prendre en compte dans l'expérience toute mesure dont les effets pourraient être confondus avec ceux des VI d'intérêt. En résumé, *les VI pertinentes se doivent d'être intégrées dans l'analyse.* Dans ce qui suit, nous appelons VI d'intérêt, celles qui sont relatives à la problématique de recherche, et VI secondaires, celles qui sont intégrées uniquement pour éviter les confusions d'effets ou/et avoir une meilleure estimation de la variabilité aléatoire[1].

Étant donné qu'ANOVA et RLM relèvent du même modèle statistique, on peut se demander, lorsque les VI étudiées sont continues, si une des deux approches doit être privilégiée ? Si oui laquelle ? Et pour quelles raisons ? Le choix doit être dicté par ce qu'un type de contrôle – expérimental ou statistique – entraîne quant à la puissance des tests et à la validité des résultats obtenus.

II.1. LE CONTRÔLE EXPÉRIMENTAL ET L'ANOVA

Le contrôle expérimental est attractif car il permet d'utiliser l'ANOVA, c'est-à-dire une approche familière pour les psychologues, relativement simple, balisée, et qui fonctionne parfaitement lorsque les VI pertinentes sont utilisées pour réaliser les contrôles.

L'intégration des VI pertinentes pour assurer un contrôle efficace des confusions d'effets et de l'erreur emprunte classiquement quatre voies.

1 / L'appel à l'une des nombreuses procédures de « randomisation » destinées à affecter les unités d'observation à des blocs homogènes, blocs non définis sur la base de variables explicites si ce n'est l'information tenant au bloc lui-même. Ces blocs n'étant utilisés qu'à la seule fin d'obtenir une standardisation des conditions expérimentales, l'étude de leurs effets n'est en général pas au centre du questionnement. Du fait que ce type d'approche est très largement débattu dans de nombreux travaux sur les plans expérimentaux (voir Winer *et al.*, 1991), nous ne nous étendrons pas sur celles-ci.

2 / L'intégration dans l'étude de VI secondaires pour sélectionner des unités d'observation qui sont homogènes au regard de ces variables. Cette intégration a lieu en amont des analyses statistiques concernant les VI d'intérêt. Ce type de procédure pose essentiellement trois types de problèmes. (*p*) Le problème le plus fréquemment cité tient au degré de généralisation possible des résultats à des unités d'observation présentant d'autres caractéristiques sur les VI secondaires. Utiliser cette voie nécessite donc d'être clairement conscient du domaine dans lequel les résultats pourront

1. Cette distinction n'est bien sûr destinée qu'à servir notre propos, et il est clair que les VI d'intérêt servent aussi à atteindre ces objectifs et qu'il n'est jamais interdit de s'intéresser aux effets d'une VI secondaire.

être généralisés. La plupart des recherches étant peu explicites quant aux caractéristiques des VI secondaires, cette question rend extrêmement difficile les appréciations de validité externe. Un exemple de ce type de problème est donné par Zevin et Seidenberg (2002) sur les effets de fréquence des mots en lecture. Ces auteurs ont montré que de nombreuses expériences dans lesquelles était réalisé un appariement sur la fréquence objective des mots se révélaient peu fiables selon le type de mesure de fréquence utilisée. L'utilisation d'échantillons de tailles réduites conduit à un appariement correct uniquement si les VI secondaires sont suffisamment fiables et valides. (*p*) Le second problème concerne la détermination du degré d'homogénéïté que les conditions expérimentales doivent présenter vis-à-vis des VI secondaires pour que l'on puisse considérer que la mesure de la variation systématique puisse être attribuée aux seuls effets des VI d'intérêt. La réponse à cette question est généralement obtenue à travers des tests *t* ou d'ANOVA sur les VI secondaires dans lesquels un non-rejet de l'hypothèse nulle est considéré comme démonstratif de l'homogénéité. Cette procédure est loin de constituer une démonstration, et ce d'autant plus que les tailles d'échantillons sont généralement faibles. Au-delà, se pose aussi la question de savoir si l'homogénéité attendue doit être de nature uni ou multivariée, question rarement abordée dans les expériences. (*p*) Les deux points précédents posent enfin le problème des comparaisons entre études. D'une part le matériel utilisable par d'autres chercheurs n'a généralement pas exactement les mêmes caractéristiques que le matériel initialement sélectionné. D'autre part, les mesures des VI secondaires utilisées doivent être dotées d'une forte fiabilité.

3 / L'intégration de VI secondaires sous forme de facteurs, c'est-à-dire une démarche basée, comme la précédente, sur la définition de blocs d'unités d'observation homogènes mais, cette fois-ci, eu égard à des variables qui sont directement intégrées dans les ANOVA.

Lorsque les VI secondaires sont catégorielles, ce sont essentiellement (*p*) des questions d'orthogonalisation du plan d'expérience qui se posent. La multiplication du nombre de variables à contrôler rend évidemment cette orthogonalisation difficile. Le résultat peut être de deux ordres non exclusifs : l'orthogonalisation est obtenue au prix d'une perte d'unités d'observation et/ou il n'est pas possible d'obtenir une telle orthogonalisation. (*t*) Dans le premier cas, il y a une perte de puissance due à la diminution du nombre d'unités. Dans le second cas, les ANOVAS réalisées correspondent en fait à de la RLM dans laquelle existent des redondances partielles entre VI. Cela a pour conséquence que les estimations obtenues sont plus variables, et donc, les tests plus conservateurs.

Lorsque les VI secondaires sont en partie continues (*p*), des questions soulevées par la nécessaire catégorisation de ces VI pour les faire entrer dans le cadre de l'ANOVA se posent. (*t*) Cette catégorisation conduit à un ajustement sur la VD qui est moins efficace que si cette catégorisation n'avait pas eu lieu. En effet, la part de variance expliquée par la relation VD/VI supplé-

mentaires est plus faible lorsque celles-ci sont catégorisées que lorsqu'elles ne le sont pas (*e.g.* Cohen, 1983). Toutefois, du strict point de vue du contrôle de l'erreur expérimentale, il n'y a pas d'avantage à ne pas catégoriser les VI secondaires (voir Winer *et al.*, 1991). (*p*) Par ailleurs, la question des possibilités de réplication que soulèvent ces pratiques se trouve particulièrement exacerbée du fait que d'un échantillon à l'autre, les quantiles des VI ne sont pas les mêmes.

4 / L'intégration de VI secondaires continues sous forme de covariables. Dans ce cas, nous retrouvons directement les problèmes qui peuvent se poser en RLM, que nous envisageons plus loin.

II.2. AUTRES CRITIQUES DE L'ANOVA

Au-delà des problèmes du contrôle direct des confusions d'effet et de l'erreur lorsque les VI sont nombreuses, et au moins partiellement continues, cinq inconvénients potentiels de l'ANOVA méritent d'être soulignés.

1 / Lorsque les VI d'intérêt sont tout ou partie continues (*p*) un problème essentiel en ANOVA est de les catégoriser pour pouvoir les intégrer à l'analyse. Une procédure classique consiste à utiliser une dichotomisation par rapport aux médianes des variables d'intérêt. Elle peut être généralisée en définissant les groupes de sujets sur la base de divers quantiles. Par-delà la (*p*) nature souvent artificielle des catégories ainsi définies (*t*) Cohen (1983) a montré que, dans le cas d'une distribution normale d'une VI, la corrélation entre la VD et la représentation dichotomisée de cette VI sur la base de la médiane était de 64 % : la corrélation existant entre la VI non dichotomisée et la VD. Autrement dit, cette pratique conduit à une importante perte de puissance de l'ANOVA par rapport à la RLM. Moins le nombre de catégories est important, et plus ce problème est exacerbé, car la perte d'information engendrée par la catégorisation est inversement proportionnelle au nombre de catégories utilisées.

2 / Outre la question de puissance, on trouve les mêmes problèmes que lorsque la catégorisation concerne des VI secondaires : (*p*) les possibilités de réplication et/ou de comparaisons inter-études sont rendues difficiles du fait que d'une étude à l'autre le matériel et les règles de catégorisation peuvent varier.

3 / De nombreuses études utilisent un nombre extrêmement restreint de catégories. Il y a même une forte tendance à la réalisation d'expériences qui utilisent seulement deux catégories. Cette tendance peut se justifier parce qu'elle conduit à des expériences moins coûteuses en termes de participants ou d'analyse des données et que les contrastes importants entre conditions permettent d'augmenter la puissance des tests, mais elle pose elle aussi le problème de la généralisation des résultats. En effet, en dehors du cas de dichotomisation de VI continues (*p*) cette pratique conduit sou-

vent à opposer deux conditions « extrêmes » et à extrapoler implicitement (et généralement linéairement) ce qu'il se passe entre elles.

4 / (*p*) Avec l'ANOVA les unités d'observation sont souvent peu nombreuses. (*t*) Une nouvelle fois se trouvent ainsi soulevées des questions de puissance et de généralité des résultats obtenus.

5 / (*p*) Enfin, le dernier reproche pouvant être fait à l'ANOVA par rapport à la RLM est qu'elle ne favorise pas l'utilisation d'indice de tailles d'effets, dont l'interprétation est moins intuitive qu'en RLM.

II.3. LE CONTRÔLE STATISTIQUE ET LA RLM

Les VI continues sont utilisées ici sous leur forme brute, soit à travers de l'ANCOVA lorsque certaines d'entre elles sont catégorielles, soit sur la base de RLM dans le cas où il n'y a que des VI continues. (*t*) Comme précisé, cette pratique conduit à une perte de puissance des tests réalisés. L'origine de celle-ci est donnée à travers des expressions comme « absence d'orthogonalité entre les VI utilisées », « colinéarité entre VI », « redondance entre VI », « non-indépendance entre VI » ou encore « corrélations existant entre ces VI ». Deux points de vue étroitement dépendants peuvent être empruntés pour caractériser les conséquences de cette propriété :

1 / (*t*) Celui de la variance expliquée par une VI particulière. Comme déjà évoqué, le contrôle statistique des confusions d'effets et de la variabilité due à l'erreur expérimentale s'opère à travers le processus de partialisation. Celui-ci revient à mesurer la part de variance expliquée uniquement par une VI comme la différence entre le R2 associé à l'équation incluant cette VI et celui résultant de l'équation qui ne l'inclut pas. La quantification de la capacité explicative d'une VI repose ainsi sur l'« élimination » de la part d'explication qui est commune à celle-ci et aux autres VI. Cette procédure est donc conservatrice et conduit à des tests moins puissants.

2 / (*t*) Celui de l'inflation de la variance des estimateurs des coefficients de la régression. Cette expression technique signifie qu'étant donné l'existence de relations entre les VI utilisées, les estimations des coefficients sont rendues plus variables d'un échantillon à l'autre.

Au regard de ces propriétés, l'utilisation de la RLM doit-elle alors être évitée ? La réponse est évidemment non univoque et dépend de plusieurs facteurs.

Le premier est celui de la présence effective de redondances importantes entre VI. L'étude des colinéarités (redondances, corrélations, etc.) entre VI s'appuie en général sur deux approches :

a) Les carrés des coefficients de corrélation multiple entre les VI ou des indices dérivés : les tolérances qui sont égales aux complémentaires (1 –) de ces carrés ; les facteurs d'inflation de la variance qui sont égaux à l'inverse (1 /) des tolérances. Il n'existe aucun test d'hypothèses ni de procédure univoque pour déterminer à partir de quel niveau l'un de ces indices

reflète une (trop) forte redondance entre une VI et les autres. Notre propre expérience de la RLM nous conduit à penser qu'il est bien rare que des VI affectées de carrés de corrélation multiple avec les autres VI inférieurs à 0,5 posent des problèmes (si tant est que d'autres problèmes, comme de faibles tailles d'échantillons, ne soient pas aussi présents). Les pertes de puissance occasionnée par ce type de variables sont souvent très relatives (cf. *infra*).

b) Les indices de conditionnement et les proportions de variance associées qui sont destinés à repérer quelles sont les VI concernées par des problèmes de multicolinéarité (pour plus de détails sur ces procédures, voir Belsey, Kuh et Welsch, 1980).

(*t*) Le deuxième facteur est celui des conséquences « pathologiques » de l'existence de telles redondances. Il faut entendre par « pathologique » le fait que certaines des estimations obtenues deviennent farfelues et/ou qu'il devient quasi impossible d'obtenir des tests significatifs. Là encore, il n'existe pas de règles univoques et un faisceau d'indicateurs doit être utilisé : *a)* Alors qu'un certain nombre de tests des coefficients individuels sont attendus significatifs, aucun ou très peu le sont. De plus le test pour le modèle complet est parfaitement significatif. *b)* Certaines des estimations des coefficients apparaissent comme parfaitement farfelues. Par exemple un signe négatif apparaît alors que la prédiction est celle d'un signe positif. Si, de plus, ces coefficients sont élevés, un réel problème de colinéarité peut être suspecté. *c)* Certaines des estimations changent énormément soit lorsqu'on utilise un autre échantillon, soit lorsqu'une des VI montrant de fortes relations avec les autres VI est retirée de l'équation. Ici aussi, un exemple caricatural de ce type est le changement de signe des estimations. Dans ces cas, le problème est certainement important et demande donc à être étudié avec soin.

Le troisième facteur consiste à savoir si les problèmes rencontrés ne peuvent être résolus soit en retirant certaines des VI de l'équation, soit en en combinant certaines, soit encore en utilisant des tests portant sur les variances expliquées par des sous-ensembles de VI ? Il arrive par exemple relativement souvent que l'origine du problème (*p*) réside dans l'utilisation d'indicateurs multiples pour un même construct théorique. Par exemple, recourir au nombre de phonèmes et au nombre de lettres pour opérationaliser la longueur d'un mot dans une expérience sur la reconnaissance visuelle de mots conduit souvent à des tests peu puissants et des estimations suspectes des effets de ces deux variables. (*p*) Un autre cas de figure courant est l'utilisation d'indicateurs composites dépendant de multiples dimensions déjà intégrées à l'analyse. En psycholinguistique, par exemple, de nombreuses VI correspondent à des estimations subjectives de certaines des caractéristiques du matériel utilisé. Pour établir celles-ci, les participants font appel à un faisceau d'autres propriétés de ce matériel. Intégré à une RLM, ce type d'indicateurs introduit ainsi une forte redondance avec ceux déjà présents. Il est donc nécessaire de les éviter le plus possible, et de les remplacer autant que faire se peut par les dimensions qui les sous-tendent (voir Bonin, Barry, Méot et Chalard (2004) pour une illustra-

tion de ce type de problème dans le cadre de l'âge d'acquisition des mots et son impact sur le traitement lexical chez des adultes).

En quatrième lieu, il faut se demander s'il existerait un avantage à se ramener à de l'analyse de variance ou de covariance ? (*t*) Est-ce que les pertes de puissance dues aux redondances entre les VI seraient compensées par la transformation de tout ou partie de celles-ci en VI catégorielles ? Comme abordé, l'ANOVA réalisée sur des VI continues qui ont été catégorisées pose, elle aussi, des questions, qui relèvent de la puissance et de la validité. Quant à l'ANCOVA, elle soulève exactement le même type de critiques que la RLM et l'observation d'une redondance souvent plus faible tient essentiellement à ce que les corrélations entre VI continues catégorisées et VI continues sont généralement plus faibles que celles entre VI « laissées » continues.

(*p*) Enfin, de manière moins générale, on pourra se demander si l'expérience est plutôt de type exploratoire ou confirmatoire. Dans le premier cas, l'aspect conservateur des tests pour certaines VI particulièrement redondantes autorise, en cas de significativité de ceux-ci (et sous réserve que les signes des coefficients ne soient pas opposés aux prédictions) de suspecter des effets particulièrement forts de ces VI. En effet, malgré le « handicap » dû à la redondance, les tests atteignent la significativité. Le phénomène est plus préjudiciable dans le cas où l'expérience a une visée confirmatoire puisqu'il peut conduire à ne pas obtenir un effet pourtant fortement anticipé.

II.4. AUTRES CRITIQUES DE LA RLM

(*p*) La première critique adressée à la RLM, comme aux études corrélationnelles, concerne bien entendu son manque de validité interne. Il est clair que l'absence d'affectation aléatoire des unités d'observation aux conditions expérimentales n'autorise pas les affirmations de causalité. Bien que non contestable, cette critique demande cependant à être relativisée. En effet, l'inférence causale dépend de la manière dont les observations sont produites, et pas de la méthode utilisée pour les analyser. Ainsi, l'ANOVA n'est qu'un cas particulier de la RLM et l'utilisation de VI catégorielles n'est pas une garantie de validité interne. En effet, nombre d'approches factorielles utilisent des plans quasi expérimentaux, dans lesquels l'affectation des sujets aux conditions n'est pas aléatoire et il n'est pas rare que, même dans les cas où il semble y avoir randomisation, celle-ci relève plus de la supposition que de la réalité (*e.g.* sujets volontaires s'inscrivant selon des heures possibles de passation). Enfin, comme exposé à propos du contrôle expérimental, validités externe et interne sont souvent obtenues en partie au détriment l'une de l'autre.

(*p*) Une seconde critique récurrente est qu'il ne serait pas possible d'étudier des interactions entre VI continues. Cohen et Cohen (1983) et

Aiken et West (1991) montrent le caractère infondé de cette critique et la possibilité de réaliser de manière routinière de telles analyses, ainsi que l'étude d'effets simples. (*t*) Cependant, notre propre pratique de la RLM nous amène à pondérer quelque peu cet optimisme du fait que les problèmes de valeurs extrêmes et/ou influentes sont plus aigus en RLM qu'en ANOVA. En effet, les termes d'interaction utilisés, en ANOVA comme en RLM, correspondent à des produits terme à terme de VI, VI généralement standardisées dans la RLM. Ces produits tendent à exacerber l'extrémisme des valeurs extrêmes en RLM, et du même coup, à les rendre particulièrement influentes lors de la procédure d'estimation. L'étude de telles valeurs doit donc être beaucoup plus fouillée en RLM qu'en ANOVA. Étant donné la complexité de cette tâche, notamment quand plusieurs valeurs « posent problème », la RLM peut alors parfois s'avérer peu fiable.

(*p*) Un troisième problème souvent rencontré en RLM est la tendance de l'utilisateur à vouloir intégrer un grand nombre de VI. Celle-ci s'explique par le désir de réaliser un contrôle efficace de toute source de variation pouvant conduire à des confusions d'effet ou à une sur-estimation de l'erreur aléatoire et par la perception plus exploratoire de la RLM que de l'ANOVA qui conduit à une certaine « pêche aux effets ». Étant donné les relations entre ANOVA et RLM, il doit être clair que, hors analyse purement exploratoire, il n'y a pas de raison d'intégrer plus de VI à la RLM qu'à l'ANOVA. Ce problème tient donc à une plus grande facilité pratique d'intégration des VI en RLM qu'en ANOVA, mais il n'a pas de justification théorique.

(*p*) Enfin, l'intégration de nombreuses VI conduit souvent l'utilisateur à faire appel à des procédures de sélection automatique de variables (voir Morris, 1981). Si cela peut se justifier dans un cadre prédictif, il vaut mieux s'abstenir de telles pratiques dans un cadre confirmatoire. En effet, la totalité des VI n'intervenant pas à un moment donné de l'algorithme, ces pratiques apparaissent être directement contradictoires avec le contrôle statistique des confusions d'effets.

III. UNE ILLUSTRATION : ÉTUDE DES EFFETS D'ÂGE D'ACQUISITION ET DE FRÉQUENCE DANS LA RECONNAISSANCE VISUELLE ET LA PRODUCTION VERBALE ORALE DE MOTS

Des recherches importantes ont été conduites sur la lecture et la production verbale orale de mots. En lecture, une tâche fréquemment utilisée est la lecture rapide de mots présentés de façon isolée. La tâche de décision lexicale est aussi très souvent utilisée. Dans cette dernière, des participants doivent décider le plus rapidement possible si la chaîne de caractères présentée sur un écran d'ordinateur renvoie à un mot de la langue ou non. Les

variables dépendantes analysées sont le temps de réaction, c'est-à-dire le temps écoulé entre la présentation du mot et le début de son énonciation ou de la décision, ainsi que divers types d'erreurs possible. Les illustrations qui suivent ont recours aux temps de réaction.

Dans certaines expériences, les caractéristiques des stimuli sont manipulées en raison de leur potentialité à opérationaliser des niveaux de traitement en jeu dans la lecture. À un niveau macroscopique, il est admis que la lecture met en jeu les niveaux de traitement suivants : sous-lexical, lexical et sémantique.

Parmi les variables supposées indexer le niveau lexical se trouvent la fréquence objective des mots, qui correspond à une mesure de la fréquence de rencontre de tel ou tel mot dans la langue. Elle est dérivée à partir d'un corpus écrit ou oral de textes de la langue en question. Toujours pour le niveau lexical, on trouve le nombre de lettres, la fréquence des bigrammes et le nombre de voisins orthographiques. Le niveau sémantique est indexé par la concrétude ; la valeur d'imagerie, et la familiarité conceptuelle. Le niveau sous-lexical est indexé par la consistance « graphie-phonie » (voir Bonin, 2003, p. 18, pour une définition de la consistance).

Certaines études intègrent des mesures de familiarité lexicale ou de fréquence subjective. Initialement destinées à pallier les insuffisances des estimations de fréquence (Gernsbascher, 1984), ces mesures se révèlent composites (Brown et Watson, 1987 ; Forster, 2000 ; Bonin, Chalard, Méot et Fayol, 2001), ce qui rend difficile leur mise en relation avec des niveaux de traitement bien spécifiés.

L'intervention de processus articulatoires en lecture à voix haute nécessite d'intégrer des indices de difficulté articulatoire. Les différences de niveaux énergétiques nécessaires à la production des phonèmes initiaux conduisent à intégrer leurs caractéristiques. Par exemple, Treiman, Mullenix, Bijeljac-Babic et Richmond (1995) ont proposé d'intégrer ces caractéristiques sous la forme de 12 traits qualitatifs différents. La longueur, exprimée en nombre de phonèmes, est aussi souvent intégrée pour prendre en compte cet aspect.

Une variable importante est l'âge d'acquisition des mots (AoA). Elle est supposée indexer le niveau lexical. L'AoA des mots peut être une mesure subjective établie par des adultes auxquels on demande d'évaluer pour un ensemble de mots la période à laquelle ils les ont acquis sous leur forme orale ou écrite (par exemple : 0-2 ans, 3-5 ans, etc.). Elle peut être une mesure dite « objective » qui est dérivée à partir d'un pourcentage de références d'enfants de telle classe d'âge étant capables de produire par exemple oralement le mot en question à partir d'un dessin (voir Chalard, Bonin, Boyer, Méot et Fayol, 2003, pour des détails).

Caroll et White (1973) ont les premiers rapporté un effet de l'AoA en dénomination d'objets. Un débat s'est alors développé entre les chercheurs pour lesquels la fréquence objective est le déterminant essentiel des vitesses de production et ceux pour lesquels il s'agit de l'AoA. D'autres ont aussi

considéré que les deux variables jouaient un rôle important (voir Bonin, 2003 pour une synthèse).

Une des difficultés essentielles des études réalisées dans le cadre de ce débat tient à ce que : 1 / AoA et fréquence objective ne sont pas indépendantes car les mots les plus fréquents sont en général appris tôt et inversement ; 2 / Fréquence et AoA sont elles-mêmes corrélées, et ce assez fortement, à d'autres caractéristiques, comme la valeur d'imagerie ou la familiarité conceptuelle.

Les chercheurs qui participent à ce débat ont utilisé des approches en ANOVA et en RLM. Nous allons illustrer à partir de travaux récents certaines des insuffisances de chacune des deux approches. Dans la suite du texte, AoA et fréquence sont considérées comme les VI d'intérêt, et donc les autres variables, comme des VI secondaires. Sauf cas particulier, nous ne nous étendrons pas sur les débats afférents aux VI qu'il est nécessaire de contrôler dans l'analyse.

Nous illustrons l'approche factorielle ou semi-factorielle à travers les travaux de Morrison et Ellis (1995), Gerhand et Barry (1998), Monhagan et Ellis (2002) pour la lecture à voix haute et Turner, Valentine et Ellis (1998) pour la décision lexicale. Nous avons restreint notre analyse à quatre études car il est difficile de les comparer – du point de vue ANOVA – étant donné la diversité des mesures et échelles utilisées. Quelques-uns des questionnements soulevés par la RLM sont illustrés à travers l'étude de Brown et Watson (1987) et celles récentes de Morrison et Ellis (2000), Bonin *et al.* (2004) en lecture à voix haute ainsi que trois ré-analyses de temps de latence réalisées par Zevin et Seidenberg (2002).

Le tableau 1 décrit les variables prises en considération pour réaliser les appariements (ANOVA) ou les contrôles statistiques (RLM). Seules les variables ayant été utilisées dans au moins deux études y figurent[1].

(*p*) La différence la plus apparente entre les deux types d'approche est la forte tendance à prendre en compte un nombre plus élevé de variables en RLM qu'en ANOVA. Deux raisons essentielles semblent en être à l'origine. (*p*) La première est la perception de l'ANOVA comme une méthode confirmatoire et celle de la RLM comme méthode plutôt exploratoire. Comme souligné, il n'y a pas de raisons théoriques statistiques pour lesquelles l'une serait plus confirmatoire que l'autre. Les modèles causaux utilisent par exemple abondamment la RLM, ou des approches dérivées, en tant qu'approches confirmatoires. Les variables secondaires qui doivent être considérées, sont celles suspectées de produire certaines confusions d'effets avec les VI d'intérêt ou/et d'avoir un effet sur la VD. (*t*) Dans ce dernier cas, il est nécessaire de les intégrer pour obtenir une estimation plus correcte de la variabi-

1. Les variables indépendantes utilisées dans une étude unique sont le point d'unicité (Turner *et al.*, 1998), une mesure de régularité (Gerhand et Barry 1998), une mesure d'ambiguïté (Brown et Watson, 1987) ainsi que l'indication d'un bloc de présentation des mots (Brown et Watson, 1987).

TABLEAU 1. — *Variables indépendantes et contrôlées dans au moins deux des travaux cités*

Independent variables and variables which were controlled in at least two of the studies

	ANOVA							RLM			
	Turner 1	Turner 2	Morrison 1	Morrison 2	Monhagan 1	Monhagan 2	Gerhand	Brown	Ellis	Bonin	Zevin
Fréquence	X	O	O	X	X	O	X	O	O	O	O
AoA	O	X	X	O	O	X	X	O	O	O	O
Longueur	O	O	O	O	O	O	O	O	O	O	O
Imagerie	O	O	O	O	O	O	O	O	O	O	O
Voisinage	O	O			O	O	O		O	O	O
Concrétude							O	O		O	O
Familiarité								O (*s*)	O (*c*)	O (*c*)	O (*s*)
Bigrammes								O		O	
Phonèmes								O	O	O	
Tailles	2 × 32	2 × 33	2 × 24	2 × 24	4 × 20	4 × 20	4 × 16	416	220	190	528

Notes. Turner = Turner, Valentine et Ellis (1998) ; Morrison = Morrison et Ellis (1995) ; Monhagan = Monhagan et Ellis (2002) ; Gerhand = Gerhand et Barry (1998) ; Brown = Brown et Watson (1987) ; Ellis = Morrison et Ellis (2000) ; Bonin = Bonin, Barry, Méot et Chalard (2004) ; Zevin = Zevin et Seidenberg (2002). 1 et 2 = expériences 1 et 2. AoA = âge d'acquisition ; Longueur = longueur (nombre de lettres ou/et de phonèmes) ; Imagerie = valeur d'imagerie ; Voisinage = densité du voisinage orthographique ; Familiarité = familiarité conceptuelle (*c*) ou lexicale (*s*) ; Bigrammes = fréquence des bigrammes ; Phonèmes = caractéristiques du phonème initial, mesures de Treiman *et al.* (1995) pour Morrison et Ellis (2000) et Bonin *et al.* (2004), autre mesure pour Brown et Watson (1987). Tailles = tailles des échantillons utilisés ; O = contrôle sur la variable en question ; X = VI servant à contraster les conditions.

lité aléatoire. Il est étonnant que certains auteurs utilisent plus de VI en RLM qu'en ANOVA pour aborder les mêmes questions de recherche. (*p*) La seconde tient à ce que le faible nombre d'unités d'observation (ici des mots) généralement utilisés en ANOVA peut forcer l'utilisateur à restreindre le nombre de VI utilisées à des effets déjà attestés. Ainsi, en dehors de la fréquence des mots et de l'AoA, seuls deux critères communs sont utilisés dans toutes les études pour réaliser les appariements entre conditions expérimentales : la longueur, mesurée en nombre de phonèmes ou/et de lettres et la valeur d'imagerie. De plus, en dehors de Monhagan et Ellis (2002), la densité du voisinage orthographique est aussi commune à toutes les études. *A contrario,* la RLM, utilisant beaucoup plus d'unités d'observation, autorise l'intégration de plus de dimensions, ce qui, selon certains chercheurs, en plus de confirmer la présence de certains effets, permet d'explorer d'autres effets potentiels.

La question de savoir si l'on perd de la puissance en utilisant une RLM à la place d'une ANOVA n'est pas simple du fait d'éléments contradictoires. Comme le montre le tableau 1 (*p*) les tailles d'échantillons sont beaucoup plus élevées en RLM qu'en ANOVA, ce qui rend plus puissante la première des approches. Par ailleurs, la ou les variables d'intérêt initialement continues, sont dichotomisées à partir de règles souvent purement techniques (cf. les textes originaux) pour être intégrées à l'ANOVA, ce qui, là aussi, favorise la RLM au regard de la (*t*) puissance mais aussi de la (*p*) validité. *A contrario,* le contrôle statistique, à travers le processus de partiallisation, conduit à (*t*) sous-estimer la variance expliquée par une VI particulière. C'est l'argument classique à l'encontre de la RLM, argument qui n'est pas contestable dans le cadre d'une ANOVA ou d'une RLM bien menée. Trois aspects contradictoires sont cependant à noter à décharge de la RLM en regard de nos illustrations :

1 / (*p* et *t*) La perte de puissance consécutive à l'utilisation de la RLM peut être renforcée par la perception plus exploratoire de celle-ci. En effet, cette perception qui conduit à intégrer plus de variables en RLM qu'en ANOVA, a pour conséquence que les tests sur les VI d'intérêt sont moins puissants, notamment dans le cas où les VI intégrées à titre exploratoire n'ont en fait pas d'effet réel sur la VD.

2 / *A contrario,* si ces VI « exploratoires » contribuent à certaines confusions d'effets avec les VI d'intérêt, alors la (*p* et *t*) puissance de l'ANOVA pour les tests des effets d'intérêt est obtenue en partie au détriment des confusions d'effets : le degré explicatif des variables d'intérêt est surévalué du fait du non-contrôle de nombre de VI secondaires, c'est-à-dire au détriment de sa validité.

3 / De la même manière, si les VI exploratoires ont un effet réel sur la VD sans pour cela contribuer à des confusions d'effets, ne pas les intégrer contribue à (*t*) l'inflation artificielle de l'estimation de la variance de l'erreur aléatoire, ce qui tend à une diminution de la puissance des tests en ANOVA par rapport à la RLM. À ce sujet, nos illustrations sont claires : les caractéristiques des phonèmes initiaux, qui expliquent en lecture à voix haute de 22 % (Treiman *et al.*,1995) à 40 % (Bonin *et al.*, 2004 ; Morrison et

Ellis, 2000) de la variance des temps de latence, sans pour cela être particulièrement corrélées aux autres caractéristiques[1], ne sont jamais intégrées dans les approches factorielles. La perte de puissance occasionnée par la non-intégration de telles VI est donc particulièrement importante.

(*p*) Un autre aspect concerne la manière dont les conditions expérimentales sont contrastées sur les dimensions d'intérêt (AoA et fréquence objective) dans les analyses factorielles. Le choix des mots devant se faire parmi un ensemble de mots normalisés sur ces dimensions, il apparaît que les contrastes obtenus sont conditionnés par l'échantillon de mots pour lequel des données normalisées existent, échantillon qui est de taille relativement réduite. Cette contrainte amène à une définition purement technique des conditions expérimentales, par exemple en utilisant la médiane ou une procédure automatique non explicitée de gestion de base de données (voir détails dans les articles originaux).

Le tableau 2 fournit des statistiques descriptives concernant l'AoA pour les études ayant manipulé cette dimension. Les moyennes inter-expériences sont comparées dans le tableau 3.

TABLEAU 2. — *Statistiques descriptives d'âge d'acquisition pour les études utilisant des approches factorielles et contrastant sur l'AoA*

Descriptive statistics of age of acquisition for studies using factorial approaches with AoA contrasted

		Turner	Morrison	Gerhand hf	Gerhand lf	Monhagan c	Monhagan i
ATo	m	2,39	2,22	2,67	2,71	2,75	2,58
	sd	0,53	0,20	0,21	0,23	0,64	0,62
	min-max	1,50-3,91	1,86-2,5	2,19-2,92	2,19-2,97	1,21-3,58	1,33-3,58
ATa	m	3,50	5,42	4,82	4,91	4,49	4,52
	sd	0,64	0,37	0,27	0,40	0,56	0,63
	min-max	2,36-4,83	5,03-6,36	4,50-5,39	4,42-5,54	3,65-5,71	3,60-5,46

Notes. Turner : Turner *et al.* (1998) ; Morrison : Morrison et Ellis (1995) ; Gerhand hf : Gerhand et Barry (1998), fréquences élevées ; Gerhand lf : Gerhand et Barry (1998), fréquences faibles ; Monhagan c : Monhagan et Ellis (2002), mots consistants ; Monhagan i : Monhagan et Ellis (2002), mots inconsistants. ATo = mots acquis tôt ; ATa = mots acquis tard. m = moyenne ; sd = écart type ; min-max = minimum-maximum.

1. Dans Bonin *et al.* (2004), le R2 le plus élevé entre une des VI d'intérêt et l'ensemble des caractéristiques des phonèmes initiaux est égal à 0,176.

TABLEAU 3. — ***Tests de comparaisons des moyennes inter-expériences dans les conditions expérimentales comprenant des mots acquis tôt (ATo, au-dessus de la diagonale) et acquis tard (ATa, sous la diagonale)***

Means comparison tests between experiments using early (ATo, above the diagonal) and late acquired words (ATa, below the diagonal)

ATa\ATo	Turner	Morrison	Gerhand hf	Gerhand lf	Monhagan c	Monhagan i
Turner			*	*	*	
Morrison	***		***	***	***	*
Gerhand hf	***	***		exp		
Gerhand lf	***	***	exp			
Monhagan c	***	***	*	*		exp
Monhagan i	***	***		*	exp	

Notes. Turner : Turner *et al.* (1998) ; Morrison : Morrison et Ellis (1995) ; Gerhand hf : Gerhand et Barry (1998), fréquences élevées ; Gerhand lf : Gerhand et Barry (1998), fréquences faibles ; Monhagan c : Monhagan et Ellis (2002), mots consistants ; Monhagan i : Monhagan et Ellis (2002), mots inconsistants. ATo = mots acquis tôt ; ATa = mots acquis tard ; *, $p < .05$; **, $p < .01$; ***, $p < .001$; exp : comparaison non significative de deux conditions ATo ou ATa définies dans la même expérience (expériences à deux facteurs).

La définition des mots acquis tôt versus acquis tardivement apparaît peu consensuelle. En effet, les tests de comparaison de moyennes inter-expériences sont majoritairement significatifs, pour les mots acquis tôt comme pour les mots acquis tard. Si les critères traditionnellement utilisés pour décider de l'existence de différences inter-conditions étaient utilisés, les expériences ne seraient pas considérées comme comparables. Par-delà les moyennes, les scores minimaux et maximaux d'AoA pour un mot acquis tôt s'étalent de 1,33 à 3,91, c'est-à-dire approximativement de 3 ans 8 mois à 6 ans et 10 mois. Même en faisant abstraction de l'étude de Turner *et al.* (1998) qui semble manifestement problématique puisque certains mots classés « acquis tôt » ont des dates d'acquisition plus élevées que les mots « acquis tard », l'écart reste important : 1,33-3,58 soit 3 ans 8 mois - 6 ans 2 mois. Cet aspect est encore plus évident pour les mots

acquis tard pour lesquels les valeurs minimales et maximales sont extrêmement différentes (de 6 ans 2 mois à 11 ans 9 mois).

L'acquis tôt et l'acquis tard n'étant pas clairement définis, on pourrait suspecter que ce qui importe n'est pas tant une définition précise de ce qu'est un mot acquis tôt ou un mot acquis tard, mais l'importance du contraste entre les conditions expérimentales. Ici aussi cependant, il n'y a pas de consensus. Les moyennes montrent un contraste acquis tôt/acquis tard allant de 3 ans 6 mois (Monhagan et Ellis : 1,74 points d'AoA) à 6 ans 5 mois (Morrison et Ellis : 3,2 points d'AoA). Quant au contraste sur la valeur maximale des acquis tôt et la valeur minimale des acquis tard il va de 0 mois (Monhagan et Ellis : 0,02 points d'AoA) à 5 ans et 1 mois (Morrison et Ellis : 2,53 points d'AoA).

L'aspect non consensuel de la définition de ce que sont des mots acquis tôt ou tard peut être généralisé, dans une plus ou moins grande mesure, à l'ensemble des autres dimensions sur lesquelles les conditions sont généralement contrastées (fréquence, consistance des mots) ou appariées (longueur, valeur d'imagerie, densité des voisinages orthographiques). Cela illustre l'une des difficultés majeures de l'approche factorielle qui est celle d'une forte spécificité du matériel utilisé dans la plupart des expériences. Cette difficulté est soulignée par Monhagan et Ellis (2002, p. 185) : « [...] the experimental items finally selected for inclusion in an experiment of the sort described here are far from being a random selection of all possible words. » En conséquence, les résultats sont fortement conditionnés par les unités d'observation utilisées et leurs qualités. Autrement dit, la validité externe de telles expériences est faible.

Si la problématique AoA / fréquence objective est propice à de telles critiques du fait de bases de données pour lesquelles on dispose de normes de tailles peu importantes, ces critiques s'appliquent à de nombreuses études qui comptent de nombreuses variables inter-reliées à apparier ou à contraster et utilisent pour ce faire des échantillons choisis parmi un nombre restreint d'unités d'observation. Notons que la critique opposée peut souvent être adressée à la RLM du fait que l'utilisateur s'autorise généralement plus qu'avec l'ANOVA un certain flou quant aux qualités exactes du matériel utilisé. Cependant, l'utilisation d'échantillons de grande taille, dont la sélection ne se base pas sur les dimensions qui sont elles-mêmes étudiées, semble plus en accord avec le modèle population - échantillon aléatoire à la base des approches statistiques utilisées.

CONCLUSION

Comme nous l'avons exposé, lorsque certaines ou toutes les VI d'intérêt sont continues, et qu'il est nécessaire de contrôler de multiples variables « secondaires », soit pour éviter des confusions d'effets, soit pour

obtenir une meilleure estimation de la variabilité aléatoire, il n'existe pas de critères non équivoques permettant de décider s'il faut préférer les analyses factorielles aux régressions multiples ou *vice versa*. En particulier, l'affirmation selon laquelle la régression multiple, de par la procédure même de contrôle statistique des confusions d'effets qui y est utilisée, est moins puissante que l'analyse de variance doit être relativisée au regard de différents critères : tailles d'échantillons généralement beaucoup plus élevées en régression ; intégration de variables secondaires moins problématique ; absence de catégorisation des variables d'intérêt. De plus, une partie des pertes de puissance classiquement attribuées à la technique de régression provient de sa perception plus exploratoire que celle de l'ANOVA. Même si la régression peut être utilisée comme technique exploratoire, une telle perception n'est pas fondée théoriquement. En ANOVA comme en régression, l'important est d'opérer un contrôle à partir des VI pertinentes ; VI, qui, pour une question expérimentale donnée, sont les mêmes quelle que soit l'approche utilisée. Par ailleurs, la plus grande puissance de l'analyse de variance est souvent obtenue au détriment de divers types de validité des expériences : externe du fait de sélections drastiques du matériel qui ne permettent pas toujours de savoir à quoi généraliser et comparer ; statistique, du fait que les nombres de variables de contrôle sont réduits le plus possible, ce qui peut conduire à une inflation artificielle des effets attribués aux variables d'intérêt et/ou à une surestimation de la variabilité due à l'erreur ; interne, du fait que les procédures destinées à assurer des contrastes suffisants entre conditions sont trop mécaniques pour assurer là aussi une définition claire des caractéristiques du matériel.

Le choix entre les deux approches ne peut être fondé sur la base de l'unique propriété de la RLM qui stipule que lorsque les VI adéquates sont inclues dans l'expérience, le processus de partiallisation rend les tests moins puissants en RLM. D'une part la puissance doit être appréciée au regard d'autres indicateurs (essentiellement la taille de l'échantillon, le contrôle plus important de la variabilité aléatoire et des confusions d'effets), d'autre part la validité, notamment externe, de l'approche utilisée doit être évaluée avec la plus grande attention.

RÉSUMÉ

Dans cet article, nous discutons des deux approches statistiques possibles lorsque les VI sont continues : l'analyse de variance et la régression linéaire multiple. Le modèle statistique commun à ces deux approches, ainsi qu'à l'analyse de covariance, est tout d'abord rappelé de manière succincte. À travers des arguments généraux, puis une illustration issue de la psycholinguistique, nous discutons ensuite des avantages et des inconvénients du recours à un contrôle expérimental des confusions d'effets et de la variabilité aléatoire, qui amène à privilégier l'analyse de variance, et à celui

d'un contrôle statistique, qui conduit à la régression linéaire multiple. Les écueils à éviter pour assurer une certaine validité à l'approche privilégiée sont aussi présentés.

Mots clés : analyse de variance, régression linéaire multiple, contrôle expérimental, contrôle statistique, puissance, validité.

BIBLIOGRAPHIE

Aiken L. S., West G. W. — (1991) *Multiple Regression : Testing and Interpreting Interactions*, Londres, Sage.

Belsey D. A., Kuh E., Welsch R. E. — (1980) *Regression Diagnostics : Identifying Influential Data and Sources of Collinearity*, New York, Wiley.

Bonin P. — (2003) *Production verbale de mots. Approche cognitive*, Bruxelles, DeBoeck Université.

Bonin P., Barry C., Méot A., Chalard M. — (2004) The influence of age of acquisition in word reading and other tasks : A never ending story ?, *Journal of Memory and Language, 50*, 456-476.

Bonin P., Chalard M., Méot A., Fayol M. — (2001) Age-of-acquisition and word frequency in the lexical decision task : Further evidence form the French language, *Current Psychology of Cognition, 20*, 401-443.

Brown G. D. A., Watson F. L. — (1987) First in, first out : Word learning age and spoken word frequency as predictors of word familiarity and word naming latency, *Memory and Cognition, 15*, 208-216.

Carroll J. B., White M. N. — (1973) Word frequency and age of acquisition as determiners of picture naming latency, *Quarterly Journal of Experimental Psychology, 25*, 85-95.

Chalard M., Bonin P., Méot A., Boyer B., Fayol M. — (2003) Objective age-of-acquisition (AoA) norms for a set of 230 object names in French : Relationships with other variables used in psycholinguistic experiments, the English data from Morrison *et al.* (1997) and naming latencies, *European Journal of Cognitive Psychology, 15*, 209-245.

Chatterjee S., Hadi A. S. — (1986) Influential observations, high leverage points, and outliers in linear regression, *Statistical Science, 1*, 379-416.

Cohen J. — (1983) The cost of dichotomisation, *Applied Psychological Measurement, 7*, 249-253.

Cohen J., Cohen P. — (1983) *Applied multiple regression/correlation analysis for the behavioral sciences*, Hillsdale, NJ, Erlbaum.

Forster K. I. — (2000) The potential for experimenter bias in word recognition experiments, *Memory and Cognition, 28*, 1109-1115.

Gerhand S., Barry C. — (1998) Word frequency effects in oral reading are not merely age-of-acquisition effects in disguise, *Journal of Experimental Psychology : Learning, Memory, and Cognition, 24*, 267-283.

Gernsbacher M. A. — (1984) Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concretness and polysemy, *Journal of Experimental Psychology : General, 113*, 256-281.

Graybill F. A. — (1976) *Theory and Application of the Linear Model*, Belmont, CA, Wadsworth Publishing Co.

Howell D. C. — (1998) *Méthodes statistiques en sciences humaines*, Bruxelles, De Boeck Université.

Keppel G., Zedeck S. — (1989) *Data Analysis for Research Designs*, New York, W. H. Freeman.

Monaghan J., Ellis A. W. — (2002) What exactly interacts with spelling-sound consistency in word naming, *Journal of Experimental Psychology : Learning, Memory and Cognition, 28,* 183-206.

Morris P. E. — (1981) Age of acquisition, imagery, recall, and the limitations of multiple regression analysis, *Memory and Cognition, 9,* 277-282.

Morrison C. M., Ellis A. W. — (2000) Real age of acquisition effects in word naming and lexical decision, *British Journal of Psychology, 91,* 167-180.

Morrison C. M., Ellis A. W. — (1995) The role of word frequency and age of acquisition in word naming and lexical decision, *Journal of Experimental Psychology : Learning, Memory, and Cognition, 21,* 116-133.

Treiman R., Mullenix J., Bijeljac-Babic R., Richmond-Welty E. D. — (1995) The special role of rimes in the description, use and acquisition of English orthography, *Journal of Experimental Psychology : General, 124,* 107-136.

Turner J. E., Valentine T., Ellis A. W. — (1998) Contrasting effects of age of acquisition and word frequency on auditory and visual lexical decision, *Memory and Cognition, 26,* 1282-1291.

Winer B. J., Brown D. R., Michels K. M. — (1991) *Statistical Principles in Experimental Design* (3e ed.), McGraw-Hill, Inc.

Zevin J. D., Seidenberg M. S. — (2002) Age of acquisition effects in word reading and other tasks, *Journal of Memory and Language, 47,* 1-29.