Nouveaux cahiers de la recherche en éducation

Mise sur pied d'une mesure de la compréhension de textes en langue seconde
Creating an instrument for measuring second language reading comprehension
Puesta en marcha de una prueba para medir la comprensión de textos en una segunda lengua

François Pichette, Linda de Serres et Marc Lafontaine

Volume 16, numéro 2, 2013

URI : https://id.erudit.org/iderudit/1029140ar
DOI : https://doi.org/10.7202/1029140ar

Aller au sommaire du numéro

Éditeur(s)

Faculté d'éducation, Université de Sherbrooke

ISSN

1911-8805 (numérique)

Découvrir la revue

Citer cet article

Pichette, F., de Serres, L. & Lafontaine, M. (2013). Mise sur pied d'une mesure de la compréhension de textes en langue seconde. *Nouveaux cahiers de la recherche en éducation*, *16*(2), 1–28. https://doi.org/10.7202/1029140ar

Résumé de l'article

La technique de vérification de phrases (TVP) (Royer, Hastings et Hook, 1979) est utilisée par des enseignants pour vérifier la lisibilité de textes destinés à leurs étudiants. Après la lecture de courts textes, les participants doivent indiquer si le sens d'énoncés qui leur sont présentés correspond à ce qu'ils ont lu. Ces énoncés consistent en des modifications de phrases du texte (paraphrase, changement de sens), en des phrases intactes ou en l'ajout de leurres. Cet article présente la mise sur pied d'un premier test TVP pour mesurer l'habileté en compréhension en lecture de l'anglais langue seconde. Quatre textes sur des sujets distincts d'intérêt général, de 12 phrases chacun, ont été soumis à deux échelles de lisibilité. À la lumière de recherches antérieures, l'article présente les étapes de l'élaboration de cet instrument, suivi de la mise à l'essai auprès de plus de 200 universitaires francophones apprenant l'anglais comme langue seconde. Pour les chercheurs, les tests standardisés actuels ne sont pas en accès libre ce qui, de ce fait, occasionne des coûts; par ailleurs, ils sont longs à compléter. Compte tenu des résultats de cette mise à l'essai, notre instrument semble être une alternative prometteuse à des tests standardisés pour mesurer l'habileté de compréhension en lecture de l'anglais L2.

Nouveaux c@hiers de la recherche en éducation
Vol. 16, nº 2, 2013

Mise sur pied d'une mesure de la compréhension de textes en langue seconde

François Pichette
Université du Québec – TÉLUQ
Linda de Serres
Université du Québec à Trois-Rivières
Marc Lafontaine
Université Laval

Résumé

La technique de vérification de phrases (TVP) (Royer, Hastings et Hook, 1979) est utilisée par des enseignants pour vérifier la lisibilité de textes destinés à leurs étudiants. Après la lecture de courts textes, les participants doivent indiquer si le sens d'énoncés qui leur sont présentés correspond à ce qu'ils ont lu. Ces énoncés consistent en des modifications de phrases du texte (paraphrase, changement de sens), en des phrases intactes ou en l'ajout de leurres. Cet article présente la mise sur pied d'un premier test TVP pour mesurer l'habileté en compréhension en lecture de l'anglais langue seconde. Quatre textes sur des sujets distincts d'intérêt général, de 12 phrases chacun, ont été soumis à deux échelles de lisibilité. À la lumière de recherches antérieures, l'article présente les étapes de l'élaboration de cet instrument, suivi de la mise à l'essai auprès de plus de 200 universitaires francophones apprenant l'anglais comme langue seconde. Pour les chercheurs, les tests standardisés actuels ne sont pas en accès libre ce qui, de ce fait, occasionne des coûts; par ailleurs, ils sont longs à compléter. Compte tenu des résultats de cette mise à l'essai, notre instrument semble être une alternative prometteuse à des tests standardisés pour mesurer l'habileté de compréhension en lecture de l'anglais L2.

Mots-clés: technique de vérification de phrases, compréhension en lecture, test de lecture, lisibilité, langue seconde

Creating an instrument for measuring second language reading comprehension

Abstract

The Sentence Verification Technique (SVT) (Royer Hastings and Hook, 1979) has been used by teachers to assess the readability of texts intended for their students. After reading short texts, participants indicate whether or not isolated sentences presented to them correspond to what they just read. These sentences are modifications of sentences from the text (paraphrases, meaning changes), intact sentences, or added distracters. This paper reports on the creation of the first SVT test developed for measuring reading comprehension ability in English as a second language. Four English texts of general interest with 12 sentences each were submitted to two readability scales. Readers are guided through the various steps in the creation of this instrument, in light of previous research, followed by the testing of the instrument with more than 200 French-speaking university students learning English as a second language. For researchers, current standardized tests are not open access, involve costs, and take time to complete. Results from this experiment suggest that our instrument is a promising alternative to standardized tests for measuring reading comprehension ability in English as a second language.

Keywords: sentence verification technique, reading comprehension, reading test, readability, second language

Puesta en marcha de una prueba para medir la comprensión de textos en una segunda lengua

Resumen

La técnica de verificación de frases (TVF) (Royer, Hastings y Hook, 1979) es empleada por los docentes para verificar la legibilidad de los textos que son utilizados con sus estudiantes. Tras la lectura de textos cortos, los participantes tienen que indicar si el sentido de los enunciados que se les presentan corresponde a lo que leyeron. Estos enunciados consisten en modificaciones de algunas frases del texto (paráfrasis, cambio de significación), frases que no presentan alteraciones o en la incorporación de distractores en las frases. Este artículo presenta la puesta en marcha de una primera prueba TVF para medir la habilidad de comprensión de lectura del inglés como segunda lengua. Cuatro textos sobre diferentes temas de interés general, de 12 frases cada uno, fueron sometidos a dos escalas de legibilidad. A la luz de investigaciones anteriores, el artículo presenta las etapas de la elaboración de este instrumento, y la aplicación, a modo de prueba, realizada a más de 200 universitarios francófonos que estudian el inglés como segunda lengua. Para los investigadores, las pruebas estandarizadas vigentes no son de libre acceso, lo que, de hecho, ocasiona costos adicionales, además de tomar mucho tiempo para ser contestadas. Tomando en cuenta los resultados de esta primera aplicación, nuestro instrumento parece ser una alternativa prometedora frente a las pruebas estandarizadas que se utilizan para medir la habilidad lectora del inglés como segunda lengua.

Palabras clave: técnica de verificación de frases, comprensión de lectura, prueba de lectura, legibilidad, segunda lengua

1. Introduction

La compréhension de textes repose sur des habiletés nombreuses et complexes (Giasson, 2007; Pulido, 2007) qui s'exercent à plusieurs niveaux: le décodage des mots, la sélection du sens approprié au contexte, l'interprétation des marqueurs référentiels et

discursifs, les inférences logiques, etc. La lecture est un processus interactif qui repose sur des variables relevant à la fois du texte, du lecteur et du contexte (Berhnardt, 1991; Grabe, 2009; Stanovich, 1980). Le nombre de variables en jeu est tel qu'il rend difficile de définir la compréhension en lecture et qu'il pourrait exister plusieurs types de compréhension (Brantmeier, 2003). Pour cette étude, nous retenons comme définition que comprendre consiste à attribuer à un texte une signification adéquate (Nuttall, 1996). Par conséquent, tel que souligné par Day et Park (2005), l'évaluation de la compréhension exige que l'on tienne compte de cette complexité. Quiconque choisit d'élaborer un test de compréhension pourrait envisager d'interroger le lecteur sur les principaux acteurs (le «qui»), l'idée véhiculée par le texte (le «quoi») ou, encore, sur les raisons qui motivent diverses actions posées (le «pourquoi»). Toutefois, de telles questions risqueraient de ne requérir, de la part du lecteur, que des habiletés de repérage.

L'évaluation de la compréhension correspond à une mesure indirecte des habiletés mises en œuvre en lecture. De fait, il est impossible d'observer la compréhension de manière explicite. Seule la production langagière procure des éléments tangibles, lesquels donnent lieu à une mesure directe de ce processus. Au fil du temps, les techniques pour évaluer la compréhension se sont améliorées et diversifiées de façon à refléter le mieux possible la compréhension chez le lecteur. Par exemple, avec la technique des questions à choix multiples (QCM), puisque des leurres aberrants seraient souvent écartés sans même qu'une véritable lecture du texte ait été faite, il est souhaitable de proposer des leurres plausibles (Tagliante, 2005). On recommande aussi de formuler la réponse attendue à l'aide de synonymes et non à partir des mots du texte source. De fait, si l'un des choix de réponses est repris textuellement, la personne évaluée peut, sans comprendre, s'appliquer à cibler les unités lexicales communes entre le texte et les questions et réussir cet item. Un énoncé dont le sens équivaut aux idées exprimées dans le texte et qui serait rédigé à partir d'autres mots permettrait, en revanche, d'inférer qu'un individu a saisi le message contenu dans le texte.

Outre la technique des QCM, on trouve celle des questions ouvertes. Ces dernières, plus exigeantes sur le plan de la production notamment, offrent l'avantage d'écarter l'effet de hasard, mais présentent le désavantage de requérir une procédure de correction plus longue et élaborée. Le recours à ce type de questions impose également une difficulté liée

aux habiletés de production même. Par ailleurs, la formulation des questions se doit d'être judicieuse pour obtenir une mesure véritable de la compréhension. À titre d'exemple, une question ouverte à réponse courte consisterait à demander au répondant de reprendre, dans ses mots, un certain passage du texte ou, encore, d'attribuer un titre à un texte. De telles questions comportent toutefois des limites certaines: la difficulté d'embrasser l'ensemble des informations transmises par le texte; le danger de cibler certaines informations au détriment de l'idée principale du texte (In'nami et Koizumi, 2009; Wolf, 1993) et, enfin, l'impact possible de la formulation des questions – rédigées, par exemple, à la négative – sur la compréhension du lecteur (Gorin, 2005).

Une autre technique pour évaluer la compréhension de textes réside dans le test de closure (Taylor, 1953). Ce test aux variantes multiples peut, entre autres, consister à remplacer des mots à intervalles réguliers par un espace, aussi appelé lacune, que le participant doit combler. Pour la correction, il est ici avisé d'accepter des synonymes pertinents au sens premier du texte et d'éviter de comptabiliser les fautes d'orthographe ou d'accord (de Serres, 2003). Les réponses au test de closure témoigneraient d'une compréhension d'informations textuelles tronquées fournies au répondant. Toutefois, bien que ce type de test ait longtemps été considéré comme une mesure fiable de la compréhension en lecture (De Landsheere, 1978; Greene, 2001; Jonz, 1987; McKamey, 2006), d'aucuns l'utilisent plutôt en guise de mesure de la compétence dans la langue lue (Brown, 1988; Chapelle et Abraham, 1990; Oller, 1973; Oller et Conrad, 1971; voir Watanabe et Koyama, 2008). Le test de closure fait toujours l'objet d'un débat à savoir s'il mesure l'habileté en lecture ou bien la compétence dans la langue lue, et ce, étant donné que ces deux phénomènes sont fortement interreliés (voir Pichette, Segalowitz et Connors, 2003).[1]

En plus de ces moyens relativement répandus, il existe une autre technique qui, après trois décennies d'utilisation ponctuelle pour estimer le degré de difficulté de textes, peut

[1] Les tests de closure restent toutefois controversés comme outil de mesure de compétence dû au fait que la performance varie pour une même personne selon le texte choisi et le nombre de lacunes insérées (Douglas et Selinker, 1985), et selon la nature et la fonction des mots supprimés (Bachman, 1985; Rankin et Thomas, 1980). Leur fiabilité varierait aussi selon le niveau de compétence des personnes testées (Alderson, 1979; Heilenman, 1983). Alderson (1980) constate que le test de closure ne permet pas de différencier des locuteurs natifs et des locuteurs non natifs, ce qui ne serait pas le cas si l'on mesurait la compétence langagière, et son explication est que le test de closure mesurerait des habiletés de bas niveau (*lower-order language skills*). Cette situation s'explique toutefois si l'on considère davantage le test de closure comme une mesure de la compréhension en lecture.

s'avérer pertinente à plus d'un égard: la technique de vérification de phrases (*Sentence Verification Technique,* Royer, Hastings et Hook, 1979). Pour bien saisir de quoi il s'agit, le lecteur est invité à se soumettre dès à présent à une courte application. Pour en assurer la réussite, le lecteur est prié de ne pas relire le texte des pages précédentes. La consigne est brève: *est-ce que les énoncés ci-dessous correspondent à du contenu lu précédemment? Indiquez OUI ou NON, sans revenir sur le texte lu*.

1. La lecture est un processus interactif qui repose sur des variables relevant à la fois du texte, du lecteur et du contexte. (OUI/NON)
2. La façon de déterminer si un lecteur a compris consiste à mesurer indirectement les habiletés auxquelles il a eu recours en lisant. (OUI/NON)
3. Par ailleurs, la formulation des passages se doit d'être judicieuse pour obtenir une mesure véritable de la compréhension. (OUI/NON)
4. Un test de closure peut être développé à partir des paroles d'une chanson. (OUI/NON)

Ces quatre énoncés reprenaient certaines informations contenues dans le texte. Voyons maintenant ce qui les caractérise. Le premier énoncé est tiré tel quel du texte source; il s'agit de la deuxième phrase du tout premier paragraphe. La réponse attendue était «oui». Le deuxième énoncé consiste en une paraphrase, où l'on a modifié le plus de mots possible de la première phrase du deuxième paragraphe («L'évaluation de la compréhension correspond à une mesure indirecte des habiletés mises en œuvre en lecture»). À cette tournure syntaxique de longueur et de complexité équivalant à la phrase du texte initial, il fallait répondre «oui». À la différence des deux premiers énoncés, les deux derniers commandaient un «non». Ainsi, pour le troisième énoncé, le mot «questions» a été remplacé par «passages» afin que l'idée exprimée ne corresponde plus à celle du texte initial. Enfin, le quatrième énoncé, formulé de façon à coïncider avec le style d'écriture du texte, renvoie à un nouveau contenu. Il s'agit d'un leurre plausible par rapport au sujet traité dans le texte. La personne évaluée devrait donc y répondre par la négative puisque l'idée y étant exprimée ne correspond à aucun élément mentionné dans le texte.

Cet exercice commenté montre que la technique de vérification de phrases repose sur la création d'énoncés de natures diverses, à partir de textes élaborés selon certains critères précis. Les tests TVP sont habituellement basés sur des extraits de textes de 12 phrases chacun. Chaque extrait est accompagné d'un total de 16 énoncés répartis comme suit: quatre phrases laissées intactes, quatre phrases paraphrasées, quatre phrases avec un sens transformé, et quatre phrases ajoutées en guise de leurres.

Les tests TVP reposent sur l'idée selon laquelle la compréhension implique la préservation du sens plus que de la forme, tout en tenant compte du rôle majeur joué par la mémoire en lecture (Baddeley, Logie, Nimmo-Smith et Brereton, 1985; De Jonge et de Jong, 1996). Un survol exhaustif d'études effectuées à partir de cette technique (Royer, 2004) dévoile d'une part que les scores obtenus grâce à des tests TVP offrent des corrélations positives élevées avec d'autres tests de compréhension en lecture et, d'autre part, que la fiabilité de tests basés sur quatre textes atteint .7 à .8. Il convient de noter que comme les indices de fiabilité dépendent du nombre d'items, un éventail plus large de tests sera éventuellement à entreprendre pour statuer sur la fiabilité des tests TVP. Ces tests ont été utilisés dans diverses langues, dont le tchèque (Zdenka, 1986, dans Royer, 2004), l'espagnol (Carlo, n.d., dans Royer, 2004) et le français (Pichette, 2002). Qui plus est, les scores reflètent les jugements d'habileté en lecture émis par les enseignants (Pichette, 2005; Royer et Carlo, 1991; Royer, Carlo, Carlisle et Furman, 1991).

Jusqu'à maintenant, ces tests tendent à être utilisés par des enseignants d'anglais comme langue maternelle et aussi comme langue seconde, surtout aux États-Unis, désireux de savoir si leur manuel ou un texte est de niveau approprié pour leurs élèves. Pour ce faire, il leur suffit de retenir des passages pour lesquels ils élaboreront 16 énoncés réponses comme nous l'avons vu plus tôt. Des scores entre 70 % et 80 % signifieraient que le texte ciblé est approprié aux besoins de leur clientèle. En ce qui nous concerne, nous avons voulu aller au-delà de l'utilisation ponctuelle de cette méthode pour vérifier si le niveau du matériel pédagogique convient. Nous souhaitions tester l'habileté en compréhension en lecture en L2.

Bien que la lecture repose sur de nombreux processus, son efficacité en L2 a principalement trait à deux facteurs: l'habileté en lecture – habituellement développée en langue maternelle – et la compétence en L2 (Alderson, 1984; voir Pichette, Segalowitz et Connors, 2003). Ainsi, l'apprenant dont le niveau de compétence dans la langue lue est très limité serait aux prises avec de trop grands efforts de décodage, ce qui lui laisserait insuffisamment de ressources cognitives pour effectuer une lecture adéquate et qui, en corollaire, ne permettrait pas l'application de stratégies de lecture efficaces (de Serres, 1998, 2003; Harrington et Sawyer, 1992; Miyake, Carpenter et Just, 1994, Pichette, 2005; Pulido, 2009). Un niveau-seuil (Clarke, 1980) de connaissances lexicales et syntaxiques, à modulation variable (Hudson, 1982), est donc nécessaire pour aspirer à lire avec aise en L2. Dans cette optique, le test ici présenté est destiné à celui qui possède un minimum de connaissances en anglais, à savoir quelque 200 heures de formation, c'est-à-dire approximativement le niveau A2 selon le Cadre européen commun de référence pour les langues (CERL; Council of Europe, 2011). En clair, ce test n'est pas à la portée d'apprenants qui en sont à leurs tout débuts en L2.

2. Buts

Le but de la recherche est la mise sur pied d'un premier test TVP pour mesurer la compréhension en lecture. Nous souhaitions proposer un test de compréhension en accès libre, sans frais, qui soit à la fois plus simple et plus court à compléter pour les utilisateurs que ne le sont les tests standardisés actuels disponibles sur le marché. Dans les prochains paragraphes, nous décrivons le processus de sélection des textes pour créer un test de compréhension selon la technique de vérification de phrases, les modifications à apporter aux textes et, enfin, les règles qui régissent la sélection et la rédaction des énoncés.

L'utilisation prévue pour cet instrument n'en serait pas une de classement ou de certification, mais serait à des fins de mesure de la compréhension dans le cadre de projets de recherche.

3. Méthodologie

La mise sur pied de ce test a couvert les huit premières des dix étapes suggérées par Crocker et Algina (1986). Certaines particularités de ce test demandaient une adaptation de ces étapes. Par exemple, les étapes 4, 5 et 6 qui concernent le passage d'une série d'items initiaux au nombre d'items désiré ont été réalisées en deux itérations, avec un premier cycle pour la sélection des textes et un second pour la sélection des items eux-mêmes. De plus, l'étape 7 de l'étude pilote a été certes réalisée, mais tel qu'il sera vu plus loin, il n'a pu être possible d'avoir un échantillon de population aussi important que celui recommandé par Crocker et Algina. Des tests supplémentaires seront nécessaires pour établir la validité de construit de l'instrument.

3.1 Création des textes

3.1.1 Choix des textes

La première étape, déterminante pour la suite du travail, consiste en la sélection de textes adéquats. Afin d'obtenir un instrument applicable à différentes populations, les textes retenus doivent traiter d'un sujet d'intérêt général, et ce, tout en respectant plusieurs critères de sélection. Le travail de conception d'une première version du test TVP appelle des mises en garde pour éviter que des variables additionnelles n'interfèrent avec la lecture normale. En ce qui touche la structure, les textes retenus doivent comporter un début, des éléments d'information, et une fin (voir Royer, 2004). Toutefois comme les textes sont courts, on ne peut pas y retrouver un schéma narratif adéquat. Ainsi, pour le contenu, il importe d'éviter certains pièges.

- Éviter les textes avec une quantité trop importante de données, c'est-à-dire des nombres, des tableaux, des figures, etc. Ce serait le cas si un texte sur le Titanic comportait de nombreuses données sur le tonnage, le tirant d'eau, les dimensions, le nombre de passagers, etc. Dans ce cas, non seulement l'attention du lecteur pourrait s'éloigner du scénario principal pour se tourner sur de tels détails, mais par

conséquent, le test TVP évaluerait davantage la mémoire des données chiffrées que la compréhension du contenu même.

- Éviter les histoires «connues» (p. ex., le Petit Chaperon rouge) qui, pour les uns, seraient trop faciles – les bonnes réponses pouvant être retracées sans une lecture du texte – et qui, pour les autres, pourraient être véhiculées dans leur culture sous une autre version ou pire, aucunement, ce qui fausserait les données du test.
- Éviter les histoires de nature fantaisiste ou relevant du fantastique, au contenu anthropomorphiste (avec des personnages hors du commun tels des sorciers et des animaux qui parlent). Cette précaution est nécessaire parce que de telles histoires peuvent mener à diverses interprétations.
- Éviter des sujets tels que la violence, le sexe, la religion, la politique, la guerre et autres, propices à susciter la controverse ou, pire, des réactions indésirables chez les participants.
- Éviter l'humour, en raison du rôle prépondérant et distrayant joué par un tel facteur affectif. Toutefois, l'humour bien dosé peut à l'occasion être toléré.

Comme des textes d'une longueur de 12 phrases ne sont pas fréquents, il nous a fallu sélectionner, puis modifier et adapter le contenu de textes existants. Dans le cas présent, une fois appliqués ces critères de sélection au sein d'une banque de quelque 100 textes tirés de revues de type «grand public» (p. ex., *Sélection du Reader's Digest*) et de divers sites Web[2], huit textes ont été retenus (voir tableau 1). Chaque texte a ensuite été modifié pour totaliser 12 phrases.

[2] Plusieurs des textes compilés se retrouvaient à plus d'un endroit, et les modifications importantes de forme et de contenu apportées aux textes les rendent parfois méconnaissables par rapport aux versions originales, d'où ici l'absence de mention des sources précises.

Tableau 1[3]

Huit premiers textes conservés accompagnés de leur score de lisibilité respectif

Titre	Nombre de phrases à l'origine	Score FK	Interprétation*	Score SCI	Interprétation
**A puzzling Parrot*	11	45	Difficile (12e année)	45	Plutôt difficile
**A special volunteer*	15	54	Plutôt difficile (9e année)	48	Plutôt difficile
Honesty without frontiers	12	54	Plutôt difficile (9e année)	19	Facile
Origin of the potato	9	33	Difficile (15e année)	57	Difficile
**The first frog without lungs*	9	23	Très difficile (18e année)	77	Très difficile
**The person that caused the Titanic to sink*	8	47	Difficile (12e année)	34	Moyen
The water is fine!	15	75	Plutôt facile (6e année)	X	(non calculable)
They tie the knot at McDonalds	9	53	Plutôt difficile (11e année)	76	Très difficile

* L'interprétation, basée sur une estimation de l'âge pour lequel le test est adéquat lorsqu'administré en langue maternelle, est fournie à titre indicatif et sert d'intermédiaire pour comparer les scores FK aux scores SCI. Dans ce cas, par exemple, les qualificatifs associés aux scores sont tirés de TxReadability (ND) de l'Université du Texas.

3.1.2 Échelles de lisibilité

Afin que l'instrument élaboré soit applicable à un vaste éventail de niveaux de compétence en lecture, les textes *a priori* ciblés devaient varier quant à leur degré de lisibilité. Pour des raisons évoquées plus tôt, il importait que ceux jugés comme étant les plus faciles soient à la portée de lecteurs dont le niveau de compétence en anglais serait peu élevé – sans toutefois en être aux balbutiements, rappelons-le. Il est normal que certains éléments lexicaux, discursifs et même graphiques d'un texte soient inconnus pour la personne dont c'est une langue seconde, tout en considérant que de telles carences peuvent également poindre chez des lecteurs en langue maternelle, notamment chez les enfants.

[3] Deux textes avec un même score FK peuvent se voir attribuer un niveau académique différent puisque la formule Flesch pour calculer le niveau académique est différente, et ce, quoiqu'elle prenne en compte les mêmes deux variables.

Pour répondre à cette exigence de lisibilité variée, les textes désormais de 12 phrases ont été subséquemment soumis à deux échelles de lisibilité de nature différente. La principale échelle utilisée a été celle de Flesch-Kincaid (dorénavant FK) (Kincaid, Fishburne, Rogers et Chissom, 1975), incluse dans le logiciel Word. Reconnue comme la plus fiable parmi celles facilement disponibles (Schinka et Borum, 1993), l'échelle FK est basée sur la longueur des phrases – évaluée selon le nombre de mots cumulés – et la longueur des mots – selon leur nombre de syllabes, puisque la longueur des mots est fortement liée à leur fréquence. D'ailleurs, il s'agit là d'un phénomène présent dans de nombreuses langues et connu depuis longtemps (Miller, Newman et Friedman, 1958; Siburg, Eeg-Olofsson et van de Weijer, 2004; Strauss, Grzybek et Altmann, 2007). De toute évidence, la longueur des mots ne se mesure pas en nombre de lettres puisqu'on ne pourrait mesurer que les langues pourvues d'un système d'écriture alphabétique. Le nombre de syllabes est une mesure plus stable d'une langue à l'autre, non soumise aux disparités oral-écrit, et constitue une unité psychologique réelle sur la base de laquelle les mots sont acquis (Peters, 1983). Des études ont d'ailleurs démontré que pour chaque syllabe supplémentaire le taux de rappel d'un mot nouveau diminue de moitié (Campaña Rubio et Ecke, 2001; Pichette, 2002). Cela dit, la stabilité du nombre de syllabes n'implique pas que cette unité de mesure soit universellement fiable, compte tenu de la grande diversité des langues humaines.

Il importe de préciser ici que les scores utilisés ne servent pas sous la forme de mesures absolues, mais plutôt en qualité d'indication de la difficulté relative des textes les uns par rapport aux autres. Ainsi, même si chaque échelle de lisibilité fournit un score qui se veut précis pour chaque texte à laquelle on l'applique, la précision de chaque score importe moins que l'ampleur de l'écart entre ces scores. De cette façon, les textes retenus peuvent couvrir un vaste éventail de difficultés. Ainsi, même si ces échelles assignent un score de lisibilité en fonction de la langue maternelle, les scores en retour reposent sur des caractéristiques intrinsèques aux textes. Tout nous permet donc d'avancer que la difficulté relative des textes sera la même en L2 qu'en L1, c'est-à-dire qu'un texte identifié comme plus difficile qu'un autre en L1 le sera de même en L2. En d'autres mots, un texte qui comporte des phrases longues et des mots rares se voudra généralement plus difficile qu'un autre avec des phrases courtes et des mots courants, que la langue du texte soit une langue première ou seconde

pour le lecteur. Il est difficile de concevoir une langue maternelle qui ferait en sorte qu'un apprenant de l'anglais aurait plus de facilité à lire un texte truffé de mots rares et de syntaxe complexe qu'un texte en phrases simples avec des mots courants.

Dans les écrits sur la lisibilité, il est parfois recommandé de considérer également des variables de nature phrastique et discursive (Wagenaar, Schreuder et Wijlhuizen, 1987). À cet égard, les quatre textes ont également été soumis à l'Indice de complexité de phrases (Sentence Complexity Index, dorénavant SCI) inclus dans le logiciel WordPerfect (voir Wampler et Williams, 1991). Cette échelle, basée sur la densité des propositions subordonnées, tient compte de la complexité des phrases. Le tableau 1 montre les indices obtenus aux deux échelles, et ce, pour les huit textes. Pour maximiser les probabilités que les textes soient significativement différents les uns des autres en termes de difficulté, notre objectif était d'atteindre des niveaux équidistants à chacune des échelles de lisibilité. Par ailleurs, pour limiter le nombre de modifications majeures à apporter, les textes retenus furent ceux qui étaient déjà les plus près des niveaux visés. De plus, il a paru sensé de rejeter les textes pour lesquels les deux échelles se contredisaient quant au niveau de difficulté attribué. Enfin, divers obstacles linguistiques ont aussi influencé le choix final, dont la présence de discours direct ou de dialogues. À la lecture du tableau, on constate que les quatre textes conservés sont précédés d'un astérisque (*). Les deux échelles considérées (FK et SCI) sont inversées: un score élevé pour FK signale un texte facile alors que pour SCI, il annonce un texte difficile. Pour l'interprétation du score FK, de source américaine, le niveau fourni entre parenthèses renvoie audit système scolaire: primaire, 1re à 7e, secondaire, 8e à 12e. Dans ces cas, le chiffre indique le nombre d'années de scolarité nécessaires en référence avec le système scolaire américain pour bien comprendre le texte dans un contexte de L1. Les nombres supérieurs à 12 correspondent au niveau universitaire.

Une fois les quatre textes choisis, s'en est suivie une série de modifications tant pour qu'ils totalisent 12 phrases que pour les rapprocher des niveaux de lisibilité ciblés. Pour que les textes soient à peu près équidistants en ce qui touche leur degré de difficulté, nous visions des scores approximatifs de 25, 40, 60 et 75 sur les échelles de lisibilité retenues.

Pour rendre un texte plus difficile sur la base de l'échelle FK, on a, par exemple, augmenté le nombre moyen de syllabes par mot, de même que le nombre moyen de mots par phrase. Pour ce faire, les principales techniques consistent à remplacer des mots du texte source par des synonymes de longueur supérieure, par des périphrases, ou à leur associer des qualificatifs de plusieurs syllabes. Pour rendre un texte plus facile, l'opération inverse s'impose: maximiser le nombre de mots brefs et de phrases courtes. Le tableau 2 fournit quelques exemples de modifications pour atteindre les scores ciblés sur les échelles, de façon à obtenir des textes de difficulté équidistante.

Tableau 2[4]

Exemples de reformulations pour modifier le degré de lisibilité

Formulation d'origine	Reformulation	Échelle	Impact
1. *It was not possible to open…*	*It was impossible to open…*	FK	-
2. *In a matter of seconds*	*In just a few seconds*	FK	+
3. *A sailor had the keys to a closet…*	*A sailor, David Blair, had the keys to a closet…*	SCI	-
4. *Until now, other parrots have failed in their attempt…*	*Other parrots have failed in their attempt…*	SCI	+

Ainsi, dans l'exemple 1, un mot court a été enlevé (*not*) et un mot de trois syllabes (*possible*) remplacé par un autre de quatre syllabes (*impossible*). Cela a eu pour effet de diminuer le score de lisibilité et d'augmenter la moyenne de syllabes par mot, pour accroître la difficulté du texte. De même, l'exemple 2 illustre la transformation d'un passage pour le rendre plus difficile; avec la modification effectuée, le nombre de mots reste maintenu alors que décroit de sept à six le nombre de syllabes. Du fait que le nombre moyen de syllabes par mot diminue, le score de lisibilité croît. Si ces transformations semblent mineures pour d'aucuns, il importe de se rappeler que chaque texte en a subi plusieurs dizaines.

Pour rendre un texte plus difficile selon l'échelle SCI, des propositions subordonnées ont été ajoutées, notamment en prenant deux phrases du texte originel pour n'en former qu'une seule, et ce, par l'ajout de virgules ou d'appositions. À cet effet, l'exemple 3 du tableau 2

[4] Pour l'impact, + signifie une plus grande lisibilité, - indique l'inverse.

montre un cas d'ajout d'une apposition afin de réduire le score de lisibilité; l'exemple 4 illustre un retrait de syntagme pour obtenir l'effet contraire. Le tableau 3 ci-dessous présente les quatre textes retenus de même que les scores de lisibilité atteints, une fois leur version de 12 phrases modifiée aux fins de lisibilité.

Tableau 3

Scores de lisibilité finaux des quatre textes formant le test TVP

Titre	Score FK	Interprétation	Score SCI	Interprétation
A special volunteer	76	Facile (6e année)	27	Facile
The person that caused the Titanic to sink	62	Moyen (7e année)	40	Moyen
A puzzling parrot	48	Difficile (10e année)	42	Moyen
The first frog without lungs	27	Très difficile (16e année)	75	Très difficile

À cette étape de création de textes qui portait sur la longueur des textes et sur leur lisibilité, succède l'élaboration des 16 énoncés réponses qui accompagneront chacun des quatre textes.

3.2 Création des énoncés

3.2.1 Ratio des énoncés

Dans le but de créer le meilleur instrument possible, il nous a paru justifié de modifier le ratio traditionnel de quatre énoncés par catégorie. Dans l'instrument élaboré, chaque texte est accompagné de 16 énoncés, mais ce, dans les proportions suivantes: cinq paraphrases, cinq changements de sens, deux phrases intactes et quatre leurres. Cette décision fait écho à une recommandation formulée par le créateur de la technique (Royer, 2004). Selon lui, un tel ratio offrirait une fiabilité supérieure. En effet, les paraphrases et les changements de sens constitueraient des éléments plus discriminants en compréhension que ne le sont les phrases laissées intactes et les leurres, alors plus facilement identifiables.

3.2.2 Mise sur pied des énoncés

Les énoncés qu'il convient de créer en premier lieu sont ceux ayant trait aux changements de sens. Ces énoncés, de nature fausse et au nombre de cinq, sont relativement difficiles à mettre sur pied en raison du faible nombre de phrases du texte, dont le sens est modifiable au degré voulu par le seul remplacement d'un mot ou deux. Une fois les phrases potentielles repérées, les modifications doivent être effectuées à l'aide de mots dont la fréquence est similaire à celles de ceux déjà présents dans le texte. La phrase proposée ne doit pas provoquer un effet inattendu. Par conséquent, le changement ne doit s'avérer ni trop subtil, ni trop évident. En guise d'exemple du caractère difficile de cette tâche, prenons la phrase suivante, tirée du texte sur le Titanic: *Blair forgot to put the keys back one night.* Un changement de sens proposé par un des chercheurs était: *Blair forgot to put the keys back one morning.* Cet énoncé a été *a priori* écarté, car la modification a été jugée trop mineure. La solution a finalement consisté à modifier deux mots: *Blair forgot to put the binoculars back one morning*.

La création de paraphrases constitue la deuxième étape de la conception des énoncés. Une fois encore, le choix des mots importe; de plus, les phrases rédigées doivent équivaloir à celles remplacées quant à la longueur, à la structure et au style. Le texte *A special volunteer* contient un exemple de paraphrase problématique qu'il nous a fallu résoudre. À partir de la phrase source *When we played ball in the yard, I noticed that he could not catch it*, un énoncé a été proposé: *When we played in the garden, I realized that he could not catch the ball*. Toutefois, puisque *yard* et *garden* ne sont pas tout à fait synonymes, il a été convenu qu'un participant pourrait ne pas se rappeler avoir lu quoi que ce soit au sujet d'un jardin et indiquerait, à tort, que l'énoncé ne correspond pas au contenu du texte. Pour remédier à cela, le terme *outside* a été retenu pour remplacer *in the yard*.

Vient ensuite la sélection des énoncés qui seront empruntés tels quels du texte. Comme exemples de tels énoncés se prêtant difficilement à un changement de sens ou à une paraphrase, on retrouve: *Parrots are birds found in warm regions* et *I took him to the veterinarian, who found that he had an eye illness*. Toute modification donnerait des items trop faciles, en raison d'un non-sens ou d'un manque de correspondance évident avec le texte.

Enfin, la dernière étape consiste à créer des leurres. À titre d'exemple, l'un des textes porte sur les visites qu'un chien effectue auprès d'enfants malades et le bonheur qui en résulte. Toutefois, il n'est jamais fait mention que ces visites entrainent la guérison de malades. Or, cette idée constitue un item potentiel pour construire une phrase de la catégorie «leurre», pour lequel une réponse positive serait erronée. Par contre, certains leurres proposés ont été refusés. Par exemple, dans le texte sur le Titanic, le leurre *Fred Fleet was born in Montreal* a été rejeté par accord inter-juges, non seulement parce qu'aucune ville n'était mentionnée dans le texte, mais surtout parce que les participants étaient tous au Québec et que s'il avait été question de l'implication d'un Québécois dans le naufrage du Titanic, tous s'en seraient probablement souvenus, ce qui en aurait fait un item non valide.

Une fois les 16 énoncés créés pour chaque texte, ils ont été réunis et présentés en ordre aléatoire. Un détail de présentation matérielle revêt de l'importance à cette étape: pour encourager le lecteur à respecter la consigne du non-retour, le texte et les énoncés ont été respectivement disposés au recto et au verso d'une même feuille. Une précaution a aussi été prise pour présenter en premier les énoncés liés à la première moitié du texte. Cette mesure visait à empêcher que le lecteur ne retrouve, parmi les premiers énoncés, une phrase qu'il viendrait tout juste de lire à la fin du texte, par exemple. Dans un tel cas, la bonne réponse aurait pu être imputable à l'effet de récence en mémoire de travail (Byrne et Arnold, 1981; Howard et Kahana, 1999; Paivio, 1971) plutôt que refléter la compréhension *per se* du texte lu.

3.3 Gestion du test

3.3.1 Passation

Le test pouvait être complété individuellement ou en groupe, selon la réponse aux appels visant à recruter des participants. Hormis quelques cas individuels, la plupart des tests ont été menés auprès de groupes variant entre 5 et 30 personnes. Sa mise sur page Web logée chez un fournisseur de tests en ligne comporte toutefois plusieurs avantages: bloquer le retour au texte lors de la lecture des énoncés; empêcher les doubles réponses pour un même énoncé; indiquer au participant la présence de réponses manquantes; et enfin,

compiler le temps pris par chacun pour compléter le test. Cette ressource permet également au chercheur de proposer les 16 énoncés de chaque texte en ordre aléatoire. Bien que cette dernière fonction contrevienne à une précaution exposée plus tôt, nous considérons la possibilité d'éventuellement obvier au problème en appliquant l'ordre aléatoire à des sous-groupes de huit énoncés.

3.3.2 Correction

Pour la correction: une bonne réponse vaut un point, une mauvaise réponse ou une absence de réponse commande un zéro. Le score de chacun des quatre textes est transposé en pourcentage. La version informatisée permet quant à elle la compilation automatique des scores.

3.4 Mise à l'essai du test

3.4.1 Étude pilote

But. Dans le processus de mise sur pied du test, l'étude pilote avait pour but de valider les contenus des textes et les énoncés inhérents à chacun, cette fois non par accord-interjuges, mais en administrant le test auprès d'un échantillon de population représentatif.

Participants. L'instrument a été mis à l'épreuve en début d'année scolaire auprès de 21 étudiants universitaires francophones ayant l'anglais comme L2. Leur âge variait entre 18 et 26 ans, avec une moyenne de 19,6. Ils ont été recrutés par des annonces faites dans leurs salles de classe. Ce nombre est le maximum que nous avons pu obtenir malgré une sollicitation intensive tout au long du semestre concerné. Selon les résultats d'une autoévaluation sur une échelle de compétence répartie en sept niveaux (les six niveaux du CERL), le niveau de compétence en anglais des participants variait d'intermédiaire à avancé, ce qui nous assurait qu'ils dépassaient le minimum voulu, c'est-à-dire le niveau A2.

Procédure. Avec une durée moyenne de 21 minutes et un éventail de durée du test qui s'étendait de 16 à 24 minutes, nous avons obtenu un temps de passation réduit en comparaison avec d'autres tests. De fait, les tests standardisés de compréhension ou

d'habileté en lecture durent habituellement plus longtemps, ce qui limite, notamment les possibilités de s'en servir auprès d'apprenants pendant les heures de cours. Par exemple, le WJ III Diagnostic Reading Battery (Woodcock, McGrew et Mather, 2001, 2007) dure entre 50 et 60 minutes; le Gates-MacGinitie (MacGinitie, MacGinitie, Maria et Dreyer, 2002) nécessite quelque 55 minutes; et le Nelson Denny (Brown, Fishco et Hanna, 1993) exige environ 45 minutes. Pour l'étude pilote, tout comme l'étude principale, le projet a reçu l'aval de deux comités universitaires d'éthique de la recherche, qui en a approuvé le design et les instruments de collecte de données.

Résultats. Le score moyen de compréhension obtenu pour les quatre textes lus s'est élevé à 87.8 % (é.t.: 4.5). Les scores s'échelonnaient de 79 % à 93 %, c'est-à-dire nettement au-dessus du 50 % imputable au hasard, sans effet de plafonnement. Des commentaires formulés par les participants peu après la passation du test ont permis de peaufiner l'instrument: suppression de coquilles, élimination d'ambiguïtés dans les items proposés, clarification de certaines consignes, etc.

3.4.2 Étude principale

Participants. À la suite de cette phase pilote, le test a été administré pendant deux semestres auprès de 171 étudiants universitaires affichant un profil comparable aux participants de l'étude pilote. Il s'agissait de 119 femmes et 52 hommes. Les participants résidaient à Montréal, Trois-Rivières et Québec. Leur âge variait de 18 à 52 ans, avec une moyenne de 20,7 ans (seuls trois participants avaient plus de 30 ans). Quelque 25 classes ont été visitées afin d'informer les étudiants sur l'étude et en leur laissant l'information nécessaire pour participer. À l'exception de six personnes anglophones, tous les participants étaient des francophones qui possédaient l'anglais comme L2 au niveau A2 et plus. Après vérification, l'exclusion de ces six participants anglophones n'aurait eu aucun impact notable sur les moyennes, les écarts types et les alphas de Cronbach de notre étude.

Nous avons fait le choix de les conserver pour les analyses. Il convient ici de préciser que la langue première tout comme la compétence en anglais L2 ne sont pas incluses comme variables dans les analyses. Tel qu'argumenté plus tôt, si un texte s'avère plus

difficile qu'un autre en vertu de ses propriétés, il le sera autant en L1 qu'en L2 et le sera autant pour un lecteur fort qu'un lecteur faible. L'objectif poursuivi ici est de comparer les items les uns aux autres et rien ne permet de croire que ce rapport entre eux serait différent en raison du statut de L1 ou de L2. En revanche, trois participants qui ont affiché des profils de réponses aberrants selon l'indice lz (Drasgow, Levine et Williams, 1985) ont été retirés des analyses. Le choix de cet indice est dû au fait qu'il est l'un des plus puissants et populaires indices statistiques pour détecter les patrons de réponses inappropriés (voir Raîche, Magis, Blais et Brochu 2012).

3.5 Résultats

3.5.1 Moyenne

La moyenne obtenue au test TVP est ici de 82.3 % (é.t.: 10.1) avec une étendue de scores de 37.3 à 96.9. Un test T de comparaison des deux moyennes indique que la moyenne au test pilote était statistiquement supérieure ($T = 2.33$; $p = .02$). Cette moyenne élevée pourrait s'expliquer par le faible nombre de participants à l'étude pilote, et le fait que la moitié d'entre eux provenaient d'un même cours auquel tendent à s'inscrire des étudiants de plus haut rendement académique.

3.5.2 Consistance interne

La consistance interne de l'instrument a été évaluée par l'alpha de Cronbach, qui est, en un mot, la moyenne de toutes les fiabilités fractionnées (*split-halves*) d'une matrice de données. Il est basé sur la proportion de réponses correctes de chaque item, ce que l'on considère refléter la difficulté de l'item dans la théorie classique des tests (Lord et Novick, 1968). Notre instrument présente un alpha de Cronbach de 0.82 (N=168), ce qui est considéré comme satisfaisant (Kline, 2000; Lowie et Seton, 2013). Aucun item, si éliminé, ne fait augmenter ou diminuer de manière marquée l'alpha de Cronbach.

3.5.3 Indépendance des items

Bien qu'aucun des 64 items n'ait été échoué ou réussi par tous les participants, le pourcentage de réussite pour les items individuels s'échelonne de 31 % à 99 %. Dans des tests de compréhension en lecture de cette nature, des pourcentages de réussite près de 100 %, tels ceux parmi les nôtres qui se situent entre 96 % et 99 %, reflètent souvent un problème d'indépendance des items. Cela indique la présence d'items qui habituellement pourraient être réussis sans même qu'un participant ait lu le texte. Par exemple, la bonne réponse ressortirait comme évidente pour un item tel *le Titanic a heurté un iceberg* ou *Le Titanic transportait de nombreux passagers lorsqu'il a sombré*, et ce, en raison du fait que de telles informations sont relativement connues de tout un chacun.

Bien qu'il soit difficile de prévoir les techniques de recoupement d'items et d'autres inférences auxquelles pourraient recourir les personnes testées, l'indépendance d'items est fréquemment vérifiée par un accord interjuges. Pour ce faire, quelques personnes doivent, par exemple, tenter indépendamment d'identifier les items faisables hors contexte, et corréler ensuite leurs jugements pour identifier le niveau de consensus (Ebel et Frisbie, 1991). Afin de sonder cette possibilité d'indépendance des items, nous avons opté pour une variante de cette méthode. Nous avons mis sur pied une version du test TVP, en tous points identiques à celle ici présentée, mais sans les textes. Cette version en ligne a été complétée par 12 personnes de profil semblable à celui des participants à l'étude pilote et à l'étude principale, c'est-à-dire des francophones universitaires, ayant l'anglais comme L2, de niveau de compétence intermédiaire ou avancé.

Tel qu'anticipé, la moyenne au test, de 56.8 % (é.t.: 6.2) avec une étendue de 48.4 à 67.2, a été nettement inférieure en l'absence des textes à lire, pour se situer près des probabilités liées au hasard. Pareil exercice fut donc concluant: les items réussis par presque tous n'ont pas eu le même succès pour ces 12 participants. À titre d'exemple, les items 6, 49 et 54, qui affichaient des scores de 98 %, 98 % et 96 % ont été réussis respectivement par seulement 7, 3 et 2 des 12 participants de la version modifiée (respectivement 58 %, 25 % et 17 %). Si les scores étaient demeurés élevés en l'absence des textes, ils auraient confirmé la thèse de l'indépendance des items. Le fait que les scores reflètent les probabilités suggère

alors que les items du test étaient simplement beaucoup plus faciles que d'autres. C'était là un motif suffisant pour décider, au final, de conserver ces items.

3.5.4 Fonctionnement différentiel des items

Deux méthodes non paramétriques ont été appliquées à nos données pour tester la présence de fonctionnement différentiel d'items (FDI) lié au sexe des participants: le test Mantel-Haenszel (Holland et Thayer, 1988) et le modèle de régression logistique (Swaminathan et Rogers, 1990). Les deux formules confirment que deux items sur 64 présentent un FDI: les items 15 (Blair forgot to put the binoculars back one morning; FDI non uniforme) et 51 (The dog would go near the children and wait to be petted; FDI uniforme). Ce faible nombre d'items suggère une bonne équité du test. L'item 51 favorise les femmes pour tous les niveaux de compétence alors que l'item 15 les favorise pour les scores moins élevés, les hommes étant favorisés pour les scores élevés. Puisque rien de particulier ne ressort dans la nature même de ces items, cela soulève des questions sur la cause de ces effets observés (pour plus de détails sur cet aspect du test, voir Pichette, Béland, Raîche et Magis, 2011).

4. Conclusion

Dans le présent article, nous avons présenté les étapes de la mise sur pied d'un premier test TVP pour mesurer la compréhension en lecture. Le but était de créer un instrument qui soit plus simple, plus court à compléter, en libre accès et gratuit pour les utilisateurs, et ce, en comparaison des tests standardisés actuels. Les premières données issues de l'administration de l'instrument auprès de notre population à l'étude, c'est-à-dire plus de 200 adultes universitaires dont une grande majorité est de langue maternelle française et possède l'anglais comme L2 suggèrent que ce test atteint ce triple objectif.

En outre, les commentaires recueillis au cours de l'étude pilote témoignent d'avantages certains: un faible coût d'administration, une économie de temps de passation, une validité apparente avérée étant donné sa convivialité tant pour l'enseignant, l'apprenant que le chercheur, le cas échéant. En effet, la brièveté des textes et leur contenu, parfois intrigant,

semblent motiver les participants. De plus, ils désirent ne pas avoir à écrire de réponses détaillées, et admettent pouvoir ainsi mieux se concentrer sur la lecture et la compréhension.

La prochaine étape consistera à évaluer plus avant les qualités psychométriques de l'instrument. Son utilisation par un nombre élevé de personnes permettra de mesurer plus finement des propriétés liées à chacun des items. Par la suite, sa validité de construit pourra être établie par l'administration conjointe auprès des mêmes personnes, du test TVP et de tests standardisés, censés mesurer les mêmes habiletés.

Références

Alderson, J.C. (1979). The cloze procedure and proficiency in English as a foreign language. *TESOL Quarterly, 13*, 219-227.

Alderson, J.C. (1980). Native and non-native performance on cloze tests. *Language Learning, 30*, 59-76.

Alderson, J.C. (1984). Reading: A reading problem or a language problem? *In* J.C. Alderson et A.H. Urquhart (éd.), *Reading in a Foreign Language* (p. 1-24). Essex: Longman.

Bachman, L.F. (1985). Performance on cloze tests with fixed-ratio and rational deletions. *TESOL Quarterly, 19*, 535-556.

Baddeley, A.D., Logie, R., Nimmo-Smith, I. et Brereton, N. (1985). Components of fluent reading. *Journal of Memory and Language, 24*, 119-131.

Bernhardt, E.B. (1991). Reading Development in a Second Language: Theoretical, Empirical and Classroom Perspectives. Norwood, NJ: Ablex.

Brantmeier, C. (2003). Does gender make a difference? Passage content and comprehension in second language reading. *Reading in a Foreign Language*, *15*(1), 1-27.

Brown, J.D. (1988). Tailored cloze: Improved with classical item analysis techniques. *Language Testing, 5*, 19-31.

Brown, J.A., Fishco, V.V. et Hanna, G. (1993). *Nelson-Denny Reading Test: Manual for Scoring and Interpretation, Forms G et H*. Rolling Meadows, IL: Riverside Publishing.

Byrne, B. et Arnold, L. (1981). Dissociating the recency effect and immediate memory span: Evidence from beginning readers. *British Journal of Psychology, 72*(3), 371-376.

Campaña Rubio, E.B. et Ecke, P. (2001). Un estudio experimental sobre la adquisición y recuperación (parcial) de palabras en una lengua extranjera. *In* G. López Cruz et M. Morúa Leyva (éd.), *Memorias del V Encuentro Internacional de Lingüística en el Noroeste* (p. 63-84). Hermosillo, Mexico: Editorial Unison.

Chapelle, C.A. et Abraham, R.G. (1990). Cloze method: What difference does it make? *Language Testing, 7*(2), 121-146.

Clarke, M.A. (1980). The short-circuit hypothesis of ESL reading — or when language competence interferes with reading performance. *The Modern Language Journal, 64*, 203-209.

Council of Europe (2011). *Common European Framework of Reference for Language: Learning, Teaching, Assessmen*t. Council of Europe.

Crocker, L. et Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. Philadelphie, PA: Harcourt Brace Jovanovich.

Day, R.R. et Park, J.S. (2005). Developing reading comprehension questions. *Reading in a Foreign Language, 17*(1), 60-73.

De Jonge, P. et de Jong, P. F. (1996). Working memory, intelligence, and reading ability in children. *Personality and Individual Differences, 21*(6), 1007-1020.

De Landsheere, G. (1978). *Le test de closure* (2e éd.). Bruxelles: Éditions Labor.

de Serres, L. (1998). *Stratégies de lecture en langue maternelle et en langue seconde, motivation et style épistémique d'étudiants inscrits à la maîtrise*. Thèse de doctorat non publiée. Université Laval, Québec.

de Serres, L. (2003). Stratégies de lecture en français langue première et en anglais langue seconde chez des universitaires diplômés: aspects quantitatifs. *Revue canadienne de linguistique appliquée, 6*(1), 31-52.

Douglas, D. et Selinker, L. (1985). Principles for language tests within the discourse domains' Theory of interlanguage. *Language Testing, 2*, 205-226.

Drasgow, F. Levine, M.V. et Williams, E.A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*, 67-86.

Ebel, R. et Frisbie, D.A. (1991). *Essentials of Educational Measurement*. Englewood Cliffs, NJ: Prentice-Hall.

Giasson, J. (2007). *La compréhension en lecture*. Paris: De Boeck.

Gorin, J.S. (2005). Manipulating processing difficulty of reading comprehension questions: The feasibility of verbal item generation. *Journal of Educational Measurement, 42*(4), 351-373.

Grabe, W. (2009). *Reading in a Second Language: Moving from Theory to Practice*. New York: CUP.

Greene, B.B. Jr. (2001). Testing reading comprehension of theoretical discourse with cloze. *Journal of Research in Reading, 24*, 82-98.

Harrington, M. et Sawyer, M. (1992). L2 working memory capacity and L2 reading skill. *Studies in Second Language Acquisition, 14*, 25–38.

Heilenman, l. (1983). The use of cloze procedure in foreign language placement. *The Modern Language Journal, 67*, 121-126.

Holland, P.W. et Thayer, D.T. (1988). Differential item performance and the Mantel-Haenszel procedure. *In* H. Wainer et H. I. Braun (dirs), Test validity. Hillsdale, NJ: Erlbaum.

Howard, M.W. et Kahana, M. (1999). Contextual Variability and Serial Position Effects in Free Recall. *Journal of Experimental Psychology: Learning, Memory and Cognition, 25*(4), 923-941.

Hudson, T. (1982). The effects of induced schemata on the 'short circuit' in L2 reading: non-decoding factors in L2 reading performance. *Language Learning, 32,* 1-31.

In'nami, Y. et Koizumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing*, *26*(2), 219-244.

Jonz, J. (1987). Textual cohesion and second language comprehension. *Language Learning, 37*, 409-438.

Kincaid, J.P., Fishburne, R.P., Rogers, R.L. et Chissom, B.S. (1975). Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. *Research Branch Report 8*.

Kline, P. (2000). *The Handbook of Psychological Testing* (2e éd.). Londres: Routledge.

Lord, F.M. et Novick, M.R. (1968). *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley Publishing Company Reading.

Lowie, W. et Seton, B. (2013). *Essential Statistics for Applied Linguistics*. Londres: Palgrave MacMillan.

MacGinitie, W.H., MacGinitie, R.K., Maria, K. et Dreyer, L.G. (2002). *Gates-MacGinitie Reading Tests* (4e éd.). Technical Report for Forms *S* and *T*. Itasca, IL: Riverside Publishing.

McKamey, T. (2006). Getting closure on cloze: A validation study of the "rational deletion" method. *Second Language Studies, 24*(2), 114-164.

Miller, G.A., Newman, E.B. et Friedman, E.A**.** (1958). Length-frequency statistics for written English. *Information and Control, 1, 370-389*.

Miyake, A., Carpenter, P.A. et Just, M.A. (1994). A capacity approach to syntactic comprehension disorders: making normal adults perform like aphasic patients. *Cognitive Neuropsychology, 11*(6), 671-717.

Nuttall, C. (1996). *Teaching Reading Skills in a Foreign Language*. Oxford: Heinemann.

Oller, J.W. (1973). Cloze tests of second language proficiency and what they measure. *Language Learning, 23*, 105-118.

Oller, J.W. et Conrad, C.A. (1971). The cloze technique and ESL proficiency. *Language Learning, 21*, 183-195.

Paivio, A. (1971). *Imagery and Verbal Processes*. New York: Holt, Rinehart and Winston.

Peters, A.M. (1983). *The Units of Language Acquisition*, Cambridge, MA: CUP.

Pichette, F. (2002). *Skill Transfer and Reading in the Second Language Classroom*. Second Language Research Forum, Toronto, October.

Pichette, F. (2005). Time spent on reading and reading comprehension in second language learning. *Canadian Modern Language Review, 62*(2), 243-262.

Pichette, F., Béland, S., Raîche, G. et Magis, D. (2011). Évaluation d'un test de lecture en anglais par deux méthodes de détection du fonctionnement différentiel d'items. *Revue des sciences de l'éducation, 37*(3), 543–568.

Pichette, F., Segalowitz, N. et Connors, K. (2003). Impact of maintaining L1 reading skills on L2 reading skill development in adults: Evidence from speakers of Serbo-Croatian learning French. *The Modern Language Journal*, *87*(3), 391-403.

Pulido, D. (2007). The relationship between text comprehension and second language incidental vocabulary acquisition: A matter of topic familiarity? *Language Learning, 57*, 155-199.

Pulido, D. (2009). Vocabulary processing and acquisition through reading: evidence for the rich getting richer. *In* Z. Han et N. J. Anderson (éd.), *Second Language Reading Research and Instruction: Crossing the Boundaries*. Ann Arbor, MI: University of Michigan Press.

Raîche, G., Magis, D., Blais, J.-G. et Brochu, P. (2012). Taking atypical response pattern into account: a multidimensional measurement model from item response theory. *In* M. Simon, K. Ercikan et M. Rousseau (éd.): *Improving Large-Scale Education Assessment*. New York, NY: Taylor and Francis.

Rankin, E.F. et Thomas, S. (1980). Contextual constraints and the construct validity of the cloze procedure. *In* M.L. Kamil et A.J. Moe (éd.), *Perspectives on Reading: Research and Instruction* (p. 47-55). Washington, DC: International Reading Conference.

Royer, J.M. (2004). *Uses for the Sentence Verification Technique for Measuring Language Comprehension*. Amherst, MA: Reading success lab.

Royer, J.M. et Carlo, M.S. (1991). Assessing the language acquisition progress of Limited-English-Proficient Students: Problems and a new alternative. *Applied Measurement in Education, 4*, 85-113.

Royer, J.M., Carlo, M.S., Carlisle, J.F. et Furman, G.A. (1991). A new procedure for assessing progress in transitional bilingual education programs. *Bilingual Review, 16*, 3-14.

Royer, J.M., Hastings, C.N. et Hook, C. (1979). A sentence verification technique for measuring reading comprehension. *Journal of Reading Behavior, 11*, 355-363.

Schinka, J.A. et Borum, R. (1993). Readability of adult psychopathology inventories. *Psychological Assessment, 5*(3), 384–386.

Siburg, B., Eeg-Olofsson, M. et van de Weijer, J. (2004). Word length, sentence length and frequency – Zipf revisited. *Studia Linguistica, 58*(1), 37-52.

Stanovich, K.E. (1980). Toward an interactive-compensatory model of individual differences in the development of reading fluency. *Reading Research Quarterly 16*, 32-71.

Strauss, U., Grzybek, P. et Altmann, G. (2007). Word length and word frequency. *In* P. Grzybeck (dir.), *Contributions to the Science of Text and Language Word Length Studies and Related Issues* (p. 277-294). Dordrecht, Netherlands: Springer.

Swaminathan, H. et Rogers, H.J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 361-370.

Tagliante, C. (2005). *L'évaluation et le Cadre européen commun*. Paris: CLE international.

Taylor, W. (1953). Cloze procedure: A new tool for measuring readability. *Journalism Quarterly, 30,* 415-433.

TxReadability (ND), *TxREADABILITY: A Multilanguage readability tool*. University of Texas in Austin: The Accessibility Institute. Accessed 1, octobre 2011 at www.utexas.edu/disability/ai/resource/readability/manual/forcast-versus-flesch-English.html.

Wagenaar, W.A., Schreuder, R. et Wijlhuizen, G.J. (1987). Readability of instructional text, written for the general public. *Applied Cognitive Psychology, 1*(3), 155-167.

Wampler, B.E. et Williams, M.P. (1991). *Grammatik Windows* {Computer program}. San Francisco, CA: Reference Software.

Watanabe, Y. et Koyama, D. (2008). A meta-analysis of second language cloze testing research. *Second Language Studies, 26*(2), 103-133.

Wolf, D.F. (1993). A comparison of assessment tasks used to measure FL reading comprehension. *Modern Language Journal, 77*, 473-489.

Woodcock, R.W., McGrew, K. et Mather, N. (2001, 2007). *Woodcock-Johnson Tests of Cognitive Abilities and Tests of Achievement* (3e éd.). Rolling Meadows, IL: Riverside Publishing.