Estimer la difficulté des questions en compréhension de l'écrit en français Vérification empirique d'un modèle théorique

Estimating the difficulty of reading comprehension questions in French. Empirical verification of a theoretical model

Verónica Sánchez Abchi – veronica.sanchez@irdp.ch
Murielle Roth – murielle.roth@irdp.ch
Alina Matei –alina.matei@irdp.ch
Institut de recherche et de documentation pédagogique (IRDP), Neuchâtel, Suisse

Pour citer cet article : Sánchez Abchi, V., Roth, M., & Matei; A. (2022). Estimer la difficulté des questions en compréhension de l'écrit en français. Vérification empirique d'un modèle théorique. *Évaluer. Journal international de recherche en éducation et formation, 8*(1), 29-46. https://doi.org/10.48782/e-jiref-8-1-29

Résumé

Estimer la difficulté des questions pour évaluer la compréhension de l'écrit des élèves s'avère indispensable dans le processus de conception de matériels d'évaluation. Le but de cet article est de décrire un outil d'estimation de la difficulté des questions d'évaluation en compréhension de l'écrit et, dans une perspective pratique, d'examiner sa valeur prédictive. Pour ce faire, le niveau de difficulté a priori de 77 questions, basées sur quatre textes, a été estimé avec cet outil théorique. Puis, les questions ont été testées auprès de 700 élèves, afin de vérifier si leurs taux de réussite empiriques confirmaient les niveaux de difficulté a priori obtenus avec l'outil théorique. Les résultats indiquent un niveau de corroboration élevé, montrant ainsi la fiabilité de l'outil pour évaluer la difficulté a priori des questions. Vu nos résultats, l'outil d'estimation pourrait être une aide précieuse pour les enseignant·es en articulation avec d'autres critères davantage en lien avec les pratiques de classe. En effet, l'outil donne un panorama de difficultés potentielles que l'enseignant·e doit ensuite pouvoir réinterpréter et adapter en fonction du contexte.

Mots-clés

Évaluation de la compréhension de l'écrit, difficulté des questions, niveau de réussite, évaluation en français.

Abstract

Estimating the difficulty of questions to assess students' reading comprehension has proven to be indispensable in the process of designing assessment materials. This article aims to describe a tool for estimating the difficulty of reading comprehension assessment questions and, from a practical perspective, to examine its predictive value. To do so, the a priori difficulty level of 77 questions, based on four texts, was estimated with this theoretical tool. Then, the questions were tested with 700 students, in order to verify if their empirical success rates corroborated the a priori difficulty levels obtained with the theoretical tool. The results indicate a high level of agreement, showing the reliability of the tool in assessing the a priori difficulty of the questions. In view of our results, the estimation tool proves to be a valuable aid for teachers, in conjunction with other criteria more related to classroom practices. The tool gives an overview of potential difficulties in terms of questions that the teacher must then be able to reinterpret and adapt according to the context.

Keywords

Reading comprehension assessment, difficulty of questions, level of achievement, French assessment.

1. Introduction

Cet article vise à explorer l'ampleur et l'utilité des différents facteurs permettant d'estimer la difficulté a priori de questions pour évaluer la compréhension écrite en français langue de scolarisation. Les facteurs pris en considération ont donné lieu à la conception d'un outil d'estimation de la difficulté (Sánchez Abchi, De Pietro & Roth, 2016), qui permet d'identifier et caractériser les dimensions ayant potentiellement une influence sur la réussite de la tâche d'évaluation. Dans cette étude, nous examinons plus spécifiquement la valeur prédictive de cet outil, c'est-à-dire que nous vérifions si l'estimation de la difficulté a priori des tâches utilisées pour l'évaluation permet effectivement d'anticiper le niveau de réussite des élèves dans des situations d'évaluation concrètes.

Il s'avère en effet que le niveau de difficulté des tâches – pas seulement le fait de savoir si une activité est facile ou difficile, mais aussi celui de pouvoir saisir les « degrés de difficulté » d'une tâche – a une importance fondamentale dans le domaine de l'évaluation. En effet, si la difficulté des tâches ne peut pas être estimée de manière fiable, c'est toute l'évaluation dans son ensemble qui devient problématique, car leur niveau de difficulté a des implications sur les opérations et les processus mobilisés et, par conséquent, sur le construit lui-même de l'évaluation, c'est-à-dire ce qu'on est effectivement en train d'évaluer. Dans cette perspective, le niveau de difficulté d'une tâche est donc en lien étroit avec la validité de celle-ci (Akib et al., 2020; Chapelle, 2017; De Pietro & Roth, 2017; Roth et al., 2019).

Plusieurs facteurs peuvent influencer l'estimation de la difficulté des questions dans le contexte de l'évaluation. Tout d'abord, ceux liés à l'élève : ses capacités langagières, ses connaissances préalables, son niveau d'attention, sa mémoire, son développement cognitif. Ensuite, certains facteurs qui relèvent du contexte, comme la communauté d'appartenance, le contexte socioculturel, l'horizon de lecture (voir également OECD, 2019) peuvent également avoir une incidence sur la compréhension du texte et sur la manière d'interpréter les questions. Finalement, les matériels d'évaluation (texte et questions associées) en soi se révèlent cruciaux. Certaines approches visent à mesurer la difficulté en se concentrant spécifiquement sur les caractéristiques des tâches, tandis que d'autres analysent l'impact de l'interaction entre les différents facteurs impliqués (voir Timpe, 2013, pour une synthèse).

Dans cet article, l'accent est mis sur les matériels d'évaluation et, de manière plus spécifique, sur les caractéristiques des tâches – composées d'un texte et de questions – mesurant la compréhension de l'écrit (CE) en français langue de scolarisation, pour des élèves de 8e (âgés de 11 à 12 ans). Pour réaliser cette exploration, nous avons utilisé les tâches évaluatives en compréhension de l'écrit du projet EpRoCom[1].

La difficulté des questions peut généralement être estimée soit *a priori*, par un groupe d'experts ou en fonction d'un modèle théorique, soit *a posteriori*, si le niveau de difficulté est déterminé après la passation d'un test auprès des élèves. Cette dernière méthode est couteuse en termes de temps et de ressources. En revanche, l'utilisation d'un modèle théorique peut

[1] Ce projet est mis en œuvre par l'Institut de Recherche et de Documentation pédagogique (IRDP) sur mandat de l'Assemblée plénière de la Conférence intercantonale de l'instruction publique de la Suisse romande et du Tessin (CIIP). Il vise à mettre à disposition des cantons romands (pour leurs propres épreuves cantonales) et de leurs enseignant·es, des activités évaluatives via une banque de données appelée banque romande d'items.

s'avérer plus pratique, si son pouvoir pour estimer le niveau de difficulté des questions est important.

Afin de pouvoir estimer le niveau de difficulté *a priori* des tâches d'évaluation en CE en français, nous avons, dans un premier temps, utilisé un outil théorique et une procédure de calcul de la difficulté des textes et des questions (Sánchez Abchi, De Pietro, & Roth, 2016)). Si ce dispositif s'est avéré très utile pour concevoir et mettre au point des matériels d'évaluation pour les enseignant·es, la valeur prédictive de cet outil devait encore être mise à l'épreuve. C'est ainsi que 77 questions, basées sur quatre textes, dont le niveau de difficulté a été estimé *a priori*, ont été testées auprès d'environ 700 élèves de 8e dans les différents cantons suisses romands. Les résultats de cette passation ont permis de calculer, dans un second temps, les taux de réussite des questions, donc leurs difficultés *a posteriori*.

Cette étude cherche donc à tester la valeur prédictive de l'outil d'estimation de Sánchez Abchi, De Pietro, & Roth, (2016), en mettant en lien les estimations de la difficulté *a priori* des questions avec la difficulté *a posteriori*, issue des résultats au test pilote. Dans la première partie de cet article, nous présenterons l'outil d'estimation de la difficulté utilisé dans le contexte de cette expérimentation et la procédure qui l'accompagne. Ensuite, nous montrerons les résultats au test pilote et nous examinerons si les dimensions qui nous ont permis d'estimer le niveau de difficulté *a priori* des questions ont une valeur prédictive de la difficulté *a posteriori*. Autrement dit, nous vérifierons si les dimensions sélectionnées pour estimer de manière théorique la difficulté *a priori* des questions permettent aussi de prédire le niveau de réussite des élèves à ces questions. Cette étape est importante pour la validation de l'outil d'estimation de la difficulté des questions, qui pourrait ainsi être utilisé sans avoir recours à leur passation auprès des élèves.

2. L'estimation de la difficulté

Étant donné l'importance du problème, il n'est pas surprenant que la question de la difficulté d'une tâche d'évaluation ait inspiré un grand nombre de recherches. Dans le domaine de l'évaluation des langues, l'impact du niveau de difficulté des tâches sur la réussite dans les évaluations de compréhension et production orale et écrite a fait l'objet de nombreuses études (Aeby et al., 2000; Cardinet, 1989; Eason et al., 2012; Mullis et al., 2015; OECD, 2019; Rog & Burton, 2002; Weiss & Wirthner, 1991, pour l'évaluation en langue première; Brindley & Slatyer, 2002 pour une synthèse dans le domaine des langues secondes).

Une manière de considérer le niveau de difficulté consiste à prendre en compte et à classer les tâches d'après leur niveau de réussite. Ces études se concentrent sur l'estimation de la difficulté *a posteriori*, à l'aide d'outils statistiques. C'est le cas des grandes enquêtes internationales comme PISA (OCDE, 2000; OECD, 2010, 2019) ou encore PIRLS (Mullis et al., 2015).

Mais il existe des approches différentes. En effet, certaines études mettent l'accent sur l'impact des caractéristiques des tâches. D'autres, sur l'articulation de ces caractéristiques avec les capacités des élèves (Anderson et al., 1991; Bachman et al., 1996; Norris et al., 2002). Ces modèles prennent en considération des facteurs tels que les caractéristiques linguistiques de la tâche, la familiarité de la thématique, et même les possibles facteurs liés au stress (limite de temps à disposition, nombre de participant·es – s'il s'agit d'une interaction orale –, longueur du texte). En général, ces études, qui ont été fondamentalement menées en langue seconde (L2), ont montré qu'il n'y a pas de relation systématique entre les estimations *a priori* des

facteurs de difficulté et les indicateurs de difficulté empiriques, c'est-à-dire, les résultats des élèves (Bachman, 2002 ; Brindley & Slatyer, 2002). C'est ainsi que Bachman (idem), considère que la difficulté n'est pas une caractéristique des tâches d'évaluation elles-mêmes, mais plutôt un aspect à mettre en lien avec la performance aux tests.

En langue première (L1), plusieurs études sur l'évaluation de la compréhension se sont intéressées aux caractéristiques des tâches et aux matériels dans l'évaluation de la compréhension (Mullis et al., 2015 ; OECD, 2019), sans toutefois les considérer comme un indicateur de difficulté *a priori* en soi. En effet, dans les évaluations de la compréhension écrite à grande échelle, comme PISA, les évaluations sont conçues en tenant compte des caractéristiques du matériel. Toutefois, c'est souvent l'estimation de la difficulté *a posteriori*, laquelle émerge de l'analyse des résultats, qui est prise en compte pour calculer la difficulté des tâches d'évaluation. Comme évoqué avant (Bachman, 2002), les deux aspects – la difficulté *a priori* et la performance – sont très liés. Cette perspective de calcul de la difficulté *a posteriori* s'avère très utile dans le processus de développement des tests à grande échelle lorsqu'on dispose de divers tests et de diverses tâches pouvant être testés de manière successive et ordonnés selon la difficulté démontrée par les analyses statistiques. Mais cette possibilité ne semble pas être à la portée des enseignant·es dans leurs pratiques d'évaluation en classe, étant donné la complexité de ces calculs et le temps qu'un tel dispositif demande.

Dans cet article, et dans le but de proposer des éléments de mesure à la portée des enseignant·es, nous nous intéressons aux caractéristiques des tâches. Nous analysons en particulier celles qui permettent d'estimer la difficulté des questions associées à un texte, afin de mieux mettre en évidence les sources potentielles de difficulté des ressources évaluatives utilisées par les enseignant·es.

3. L'outil théorique d'estimation

L'outil que nous allons décrire dans ce chapitre (Sánchez Abchi, De Pietro, & Roth, 2016) prend en compte deux aspects principaux : d'une part, la difficulté du texte que les élèves doivent lire, et, de l'autre, la difficulté de la question. Dans notre étude, la « question » correspond à l'« item », terme principalement utilisé dans les évaluations sommatives se basant sur le paradigme de la mesure (Laveault & Grégoire, 2014).

3.1. L'estimation de la difficulté du texte

Afin de pouvoir estimer le niveau de difficulté des textes, nous avons fait le choix de retenir trois dimensions, reprises de la littérature et réinterprétées sur la base de nos analyses (François, 2009; Kandel & Moles, 1958; Lafontaine, 2003; McNamara et al., 2012) : la complexité lexicale – évaluée à l'aide d'un logiciel (Mesnager & Bres, 2008) – ; la complexité syntaxique – calculée sur la base des mesures de lisibilité (Gunning, 1952) – et, finalement, la complexité de la structure textuelle (Bogaerds-Hazenberg et al., 2021; McNamara et al., 2012) – évaluée en définissant si celle-ci est typique du genre dont relève le texte puis nuancée par la présence ou l'absence d'organisateurs textuels et leur pertinence.

3.2. L'estimation de la difficulté des questions

En ce qui concerne les questions, elles peuvent également être de difficulté variable en fonction de plusieurs dimensions. Tout d'abord « leur formulation » est susceptible d'être plus ou moins difficile en raison d'un certain nombre d'éléments, tels que le vocabulaire

utilisé ou la structure syntaxique. (Guay, 2011 ; OCDE, 2000). Ensuite, les différents processus de lecture impliqués comme ceux définis dans les enquêtes PISA (OCDE, 2000) et PIRLS (Mullis et al., 2015). Les processus de lecture, que nous opérationnalisons dans notre dispositif comme « opérations cognitives », semblent bien être un critère à retenir pour estimer le niveau de difficulté des questions posées. Plusieurs recherches ont également pointé l'importance du format des questions lorsqu'il s'agit d'évaluer (Cardinet, 1987; Fletcher, 2006).

Dans notre outil, nous avons fait le choix de retenir deux facteurs : a) le contenu de la question et b) son « enveloppe ». Chacun d'eux se compose de plusieurs dimensions décrites ci-après.

Le facteur « contenu » est composé de trois dimensions : les opérations cognitives que l'élève est censé mobiliser, l'ampleur de l'objet langagier (renvoie à l'étendue de ce que l'élève doit avoir compris pour répondre à la question : le texte en entier, un paragraphe, une phrase ou un mot) et le nombre d'activités à réaliser (c'est-à-dire le nombre de réponses à donner).

Quant au facteur « enveloppe », il est constitué de trois dimensions : la formulation de la consigne, le format de questionnement et le matériel complémentaire – les aides fournies pour réussir la question, tels que les dictionnaires, les tableaux de conjugaison, la grammaire de référence ou les moyens d'enseignement.

3.3. Les facteurs de l'outil d'estimation

Pour caractériser la tâche d'évaluation (le texte et les questions associées), notre outil retient au final trois facteurs incluant eux-mêmes plusieurs dimensions :

a) le texte,

b) le « contenu » de la question,

c) l' « enveloppe » de la question.

L'ensemble de ces dimensions définit, en quelque sorte, la manière dont la compréhension du texte est appréhendée et permet de mieux saisir les difficultés potentielles que les élèves vont rencontrer. Pour chaque dimension, nous avons attribué une de ces trois valeurs : -1 = facile, 0 = intermédiaire ou 1 = difficile. Ensuite, les moyennes de ces valeurs pour chacun des facteurs – le contenu, l'enveloppe et le texte – sont transformées en indices qualitatifs d'après la tabelle A.

Tabelle A : correspondance entre la valeur moyenne obtenue et l'indice de difficulté

Valeurs moyennes	**Indices de difficulté**
-1 à -0.61	Facile (F)
-0.6 à -0.21	Plutôt facile (PF)
-0.2 à 0.2	Intermédiaire (I)
0.21 à 0.6	Plutôt difficile (PD)
0.61 à 1	Difficile (D)

Le tableau 1 ci-dessous synthétise les éléments principaux de l'outil (les variables théoriques « difficulté *a priori* valeur texte », et pour les questions, « difficulté *a priori* valeur contenu » et « difficulté *a priori* valeur enveloppe ») ainsi que la manière de calculer leurs valeurs. La procédure détaillée de calcul pour chaque variable sera décrite dans la partie méthodologique.

Tableau 1. Les variables de l'outil d'estimation de la difficulté

Variable Texte			**Variable Contenu questions**			**Variable Enveloppe questions**		
Vocabulaire (Va)	Complexité syntaxique (Vb)	Structure texte (Vc)	Type d'opération cognitive (Vd)	Ampleur objet langagier (Ve)	Nombre activités à réaliser (Vf)	Consigne (Vg)	Format de questionnement (Vh)	Matériel à disposition (Vi)
= (Va + Vb + Vc) / 3 Valeur moyenne qui peut être convertie en indice qualitatif			= (Vd + Ve + Vf) / 3 Valeur moyenne qui peut être convertie en indice qualitatif			= (Vg + Vh + Vi) / 3 Valeur moyenne qui peut être convertie en indice qualitatif		

4. Les questions de recherche

Afin d'explorer la valeur prédictive de l'outil d'estimation de la difficulté décrit dans la section précédente, les questions de recherche qui guident notre étude sont les suivantes :

1) Est-ce que les trois variables théoriques (« difficulté *a priori* valeur texte », « difficulté *a priori* valeur contenu » d'une question et « difficulté *a priori* valeur enveloppe » d'une question) qui sont à la base de l'outil d'estimation sont liées au taux de réussite des questions calculé auprès des élèves ?

2) Dans quelle mesure les variables de l'outil d'estimation de la difficulté *a priori* permettent-elles de prédire le taux de réussite d'une question non testée auprès des élèves ?

Nous formulons l'hypothèse que le taux de réussite d'une question peut être en partie prédit par l'outil d'estimation, mais que d'autres facteurs peuvent également jouer un rôle.

5. Méthodologie

L'outil d'estimation a été utilisé dans le cadre du projet EpRoCom pour estimer la difficulté des tâches évaluatives en compréhension de l'écrit en français langue de scolarisation. Celles-ci provenaient d'épreuves cantonales de 8e que les cantons romands avaient mises à disposition. Quatre textes et 77 questions, distribués en six cahiers, ont été soumis, via un test pilote, à environ 700 élèves de Suisse romande. Chaque cahier de test était constitué de deux textes et de huit questions par texte.

5.1. Procédure d'estimation de la difficulté des matériels

Trois chercheures ont analysé – de manière indépendante – les textes et les questions et leur ont attribué un niveau de difficulté *a priori*. Ensuite, une procédure d'accord interjuges a été mise en place pour déterminer le degré de concordance entre les personnes. Les trois

chercheures ont comparé leurs résultats pour s'accorder. Les cas de désaccord ont été discutés pour ajuster les analyses.

5.1.1. Difficulté des textes

Pour estimer la difficulté des textes, une valeur – -1 (= facile), 0 (= intermédiaire) ou 1 (= difficile) – a été attribuée à chacune des trois dimensions (voir annexe 1). Puis nous avons effectué une moyenne dont le résultat a été mis en relation avec la Tabelle A pour obtenir un indice de difficulté qualitatif. Celui-ci pouvait encore être nuancé par plusieurs autres critères, proposés sous forme de question à se poser (par exemple : Est-ce que le contenu est familier (ou non) pour le lecteur ?). Ces critères pouvaient ainsi apporter des bonus ou des malus[2].

5.1.2. Procédure d'estimation de la difficulté des questions

Comme pour les textes, une des trois valeurs a été octroyée pour chaque dimension. Ensuite, la moyenne pour chacun des facteurs (contenu et enveloppe) a été calculée et les valeurs ont été attribuées en relation avec la Tabelle A, présentée avant.

Pour le contenu de la question :

- L'opération cognitive mobilisée par l'élève est considérée comme facile (valeur -1) lorsqu'il s'agit d'identification ou de repérage ; intermédiaire (valeur 0) dans le cas d'une inférence simple avec mise en lien d'informations données dans le texte ; et difficile (valeur 1) quand l'inférence est complexe (par exemple l'inférence de cause) (Graesser et al., 1994) ou lors d'une interprétation.
- Le nombre d'activités que l'élève doit réaliser est considéré comme facile (valeur -1) entre 1 et 3 ; intermédiaire (valeur 0) entre 4 et 6 ; et difficile (valeur 1) dès 7.
- L'ampleur de l'objet langagier est estimée comme facile (valeur -1) dans le cas d'un mot ou d'un syntagme ; intermédiaire (valeur 0) lorsqu'il s'agit d'une phrase ou d'un fragment ; et difficile (valeur 1) quand c'est un texte ou une situation communicative/contexte.

Pour l'enveloppe de la question :

- La formulation de la consigne est considérée comme facile (valeur -1) quand le vocabulaire est accessible, la structure claire et que le nombre de réponses à donner est indiqué ; intermédiaire (valeur 0) lorsqu'un mot et/ou un élément structurel risque(nt) de rendre la consigne plus difficile à comprendre et que le nombre de réponses attendues n'est pas indiqué ; et difficile (valeur 1) lorsque le vocabulaire est complexe, que la structure n'est pas claire, que le nombre de réponses attendues n'est pas précisé et qu'éventuellement il y a des éléments de contextualisation ambigus ou artificiels.
- Le format de questionnement est évalué sur la base de la typologie de Pini et al. (2006) avec quelques adaptations. Les questions à réponse visuelle, les vrai-faux et celles à choix multiples avec une seule bonne réponse sont considérées comme faciles (valeur -1) ; les questions à choix multiples avec plusieurs bonnes réponses, les appariements,

[2] Pour davantage d'explication sur l'attribution des bonus et malus, voir : Sánchez Abchi, De Pietro& Roth (2016).

les classements, les ordinations et les questions à réponse brève sont estimées intermédiaires (valeur 0) ; et les questions à réponse textuelle et demandant une performance sont évaluées comme difficiles (valeur 1).

- Le matériel à disposition (dictionnaire, grammaire, conjugaison) est uniquement pris en compte s'il peut faciliter la résolution de la question ; dans ce cas il est considéré comme facilitateur et une valeur de -1 lui est attribuée.

5.2. Procédure d'analyses statistiques

Afin de répondre aux questions de recherche, une série d'analyses statistiques ont été effectuées. Pour ce faire, le taux de réussite des 77 questions utilisées dans le test pilote a été considéré. Ces taux ont été calculés sur la base des résultats des élèves, en prenant en compte les réponses correctes et partiellement correctes aux questions. Le taux de réussite pour les 77 questions varie entre 22.8% et 97.1%.

L'analyse des données a été effectuée en deux étapes. Dans un premier temps, nous avons testé s'il existait un lien significatif entre les trois variables théoriques de difficulté (le texte, le contenu et l'enveloppe d'une question) et le taux de réussite des questions. Dans un deuxième temps, plusieurs modèles de régression (voir Howell, 2008 pour une synthèse) ont été mis en œuvre afin de déterminer dans quelle mesure les variables théoriques de difficulté prédisent le taux de réussite aux questions. Autrement dit, ces modèles devaient nous fournir une prédiction de la difficulté d'une nouvelle question (non encore testée auprès des élèves), à partir des variables théoriques. Finalement, un seul modèle de régression a été retenu pour nos résultats.

6. Résultats

6.1. Difficulté a priori et taux de réussite

L'outil d'estimation décrit précédemment prévoit 7 valeurs possibles pour la variable « difficulté *a priori* valeur contenu » (-1, -0.66, -0.33, 0, 0.33, 0.66 et 1) qui prend en compte les 3 dimensions possibles – opération cognitive, ampleur de l'objet langagier et nombre d'activités à réaliser – et 5 valeurs possibles pour la variable « difficulté *a priori* valeur enveloppe » (-1, -0.5, 0, 0.5 et 1). Il faut signaler que la valeur de l'enveloppe a été calculée sur la base de seulement deux dimensions – consigne et format de questionnement, parce que la troisième dimension – matériel complémentaire – n'a pas été identifiée dans les questions. En effet, aucune question ne comportait de matériel supplémentaire pour faciliter la tâche, donc l'indice final est basé sur deux dimensions au lieu de trois (cf. tableau 1).

Les valeurs des variables difficulté *a priori* du contenu, de l'enveloppe et du texte ont été calculées selon la procédure décrite dans la section 5. Les 77 questions de l'échantillon ne présentent pas toutes les valeurs théoriquement possibles signalées ci-dessus. Les valeurs retrouvées dans l'échantillon des 77 questions se distribuent de la manière suivante :

- Pour la variable « difficulté *a priori* valeur contenu » : 6 valeurs (-0.66, -0.33, 0, 0.33, 0.66 et 1) ;
- Pour la variable « difficulté *a priori* valeur enveloppe » : 4 valeurs (-1, 0.5, 0 et 0.5) ;
- Pour la variable « difficulté *a priori* valeur texte » : 3 valeurs (-0.33, 0 et 0.25). Notons que 23 questions avec la valeur texte égale à 0 ont reçu un malus de 0.25 pour la

difficulté du texte utilisé, en raison de la thématique qui est peu familière pour des élèves de 8e. Ce malus a compté pour un quart dans le calcul de la moyenne.

Les valeurs pour les trois variables théoriques, ainsi que les nombres de questions par valeur possible, sont présentées dans les figures ci-dessous. On observe que, pour la variable « difficulté *a priori* valeur contenu » par exemple, 21 questions ont la valeur -0.66 et sont donc classées comme étant faciles, tandis que seulement 4 questions de notre échantillon ont la valeur 1, donc considérées comme difficiles (voir la Figure 1).

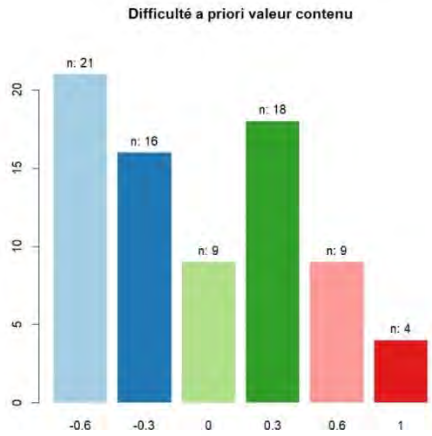

Figure 1. Valeurs de la variable « difficulté *a priori* valeur contenu » et nombre de questions par valeur possible

Pour la variable « difficulté *a priori* enveloppe », la plupart des questions, 38 au total, se voient attribuer la valeur « 0 ». Cela permet de les considérer comme étant de difficulté intermédiaire (voir la Figure 2).

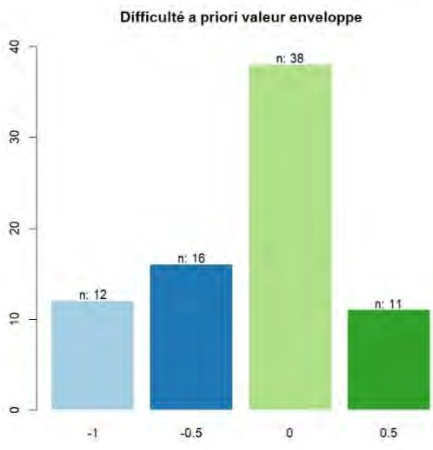

Figure 2. Valeurs de la variable « difficulté *a priori* valeur enveloppe » et nombre de questions par valeur possible

Finalement, pour ce qui concerne la difficulté des questions en lien avec le texte auquel elles sont associées, la plupart des questions (40) présentent un niveau de difficulté texte plutôt facile (Figure 3).

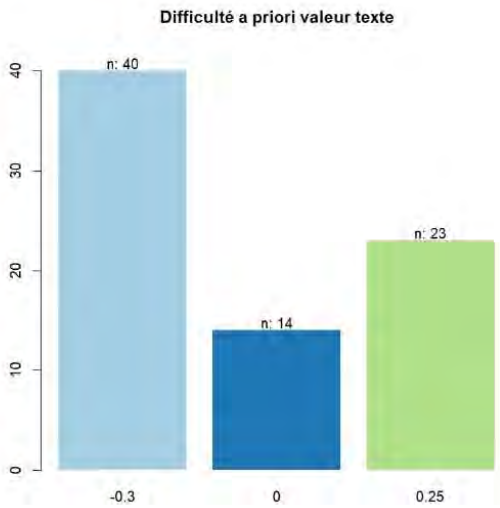

Figure 3. Valeurs de la variable « difficulté *a priori* valeur texte » et nombre de questions par valeur possible

Le nombre d'élèves ayant résolu les questions varie d'une question à l'autre, de 105 élèves à 648 élèves. Autrement dit, certaines questions ont été passées par 105 élèves, d'autres par 648 et quelques questions par un nombre d'élèves qui se situe entre 107 et 324. Pour prendre en compte cette différence d'une question à l'autre, les résultats présentés ci-dessous ont été pondérés en utilisant un poids associé à chaque question. Pour calculer celui-ci, on considère le rapport entre le nombre d'élèves qui a passé la question et la somme du nombre d'élèves pour toutes les questions.

Les corrélations (de Spearman) entre le taux de réussite aux questions et les trois variables théoriques sont significatives au seuil de 5%. Celles-ci ont les valeurs : -0.59 pour la variable « difficulté *a priori* valeur contenu », -0.39 pour la variable « difficulté *a priori* valeur enveloppe » et -0.30 pour la variable « difficulté *a priori* valeur texte ». Ces valeurs indiquent qu'une augmentation des difficultés (contenu, enveloppe, texte respectivement) est associée à une réduction du taux de réussite d'une question, comme attendu. Ces résultats montrent le lien significatif entre la variable « taux de réussite d'une question » et les trois variables théoriques, et nous permettent de répondre, ainsi, de manière positive à notre première question de recherche. Il reste encore à vérifier le pouvoir prédictif de l'outil d'estimation.

6.2. Un modèle statistique de prédiction

Afin de répondre à notre seconde question de recherche, et savoir si l'outil d'estimation de la difficulté permet de prédire le taux de réussite d'une question, plusieurs modèles de régression ont été ajustés et analysés, par rapport aux hypothèses d'application de chaque modèle. Pour rappel, le taux de réussite (réponses correctes et partiellement correctes) pour les 77 questions utilisées varient entre 22.8% et 97.1%. Comme le but est de vérifier la

pertinence de l'outil d'estimation pour la prédiction des taux de réussite, les modèles de régression utilisent comme variables indépendantes les trois variables théoriques : « la difficulté *a priori* valeur contenu », « la difficulté *a priori* valeur enveloppe » et « la difficulté *a priori* valeur texte ». Ces variables ont été prises dans les modèles comme variables catégorielles.

Les modèles de régression ont été comparés par rapport aux erreurs de prédiction en utilisant la méthode de la validation croisée. Trois questions ont été écartées, jugées comme atypiques du point de vie statistique. Le modèle de régression retenu est un modèle linéaire ayant comme variable dépendante la proportion de réponses correctes et partiellement correctes pour chaque question (donc le taux de réussite divisé par 100), transformée à l'aide des fonctions arc sinus et racine carrée[3]. La valeur du coefficient de détermination R^2 pour ce modèle est de 0.54[4].

Les coefficients estimés sont significatifs dans le modèle (au seuil de 5%), à l'exception de ceux associés à la valeur -0.5 de la variable « difficulté *a priori* enveloppe » et à la valeur 0 de la variable « difficulté *a priori* texte ». Nous les conservons dans le modèle, car ces deux coefficients ont des valeurs assez petites et le but du modèle est la prédiction du taux de réussite d'une nouvelle question.

La significativité de la plupart des coefficients estimés du modèle ainsi que la valeur du coefficient de détermination R^2 (qui est assez élevée pour ce type de données), nous indiquent que le modèle est assez fiable pour prédire le taux de réussite d'une question, non testée auprès des élèves. Par exemple, pour une nouvelle question avec la difficulté du contenu -0.33, la difficulté de l'enveloppe -0.5 et la difficulté du texte 0, le modèle fournit la prédiction égale à 0.82 et donc un taux de réussite de 82%.

Les prédictions du taux de réussite obtenues à l'aide de ce modèle de régression se situent dans l'intervalle 40% à 99% pour les données utilisées. La valeur 99% correspond à une question ayant la « difficulté *a priori* valeur contenu » égale à -0.6, la « difficulté *a priori* valeur enveloppe » égale à -1 et la « difficulté *a priori* valeur texte » égale à -0.3 (les plus petites valeurs sur l'échantillon qui correspondent à la question la plus facile) ; la valeur 40% correspond à une question ayant les valeurs 1, 0.5 et 0.25 pour les trois variables respectivement, c'est-à-dire, contenu, enveloppe et texte (les plus grandes valeurs de l'échantillon qui correspondent à la question la plus difficile).

Pour rappel, plus ces valeurs sont grandes, plus la question est difficile et plus le taux de réussite est bas. Le modèle de régression nous empêche de prédire des taux de réussite de moins de 40%. La limite du modèle de régression est une conséquence de l'échantillon utilisé. Comme déjà signalé, l'échantillon contient des questions qui n'ont pas toutes les valeurs possibles des variables théoriques. La valeur 40% diminuera probablement en présence d'une « difficulté *a priori* valeur enveloppe » égale à 1, par exemple. Toutefois, aucune question de notre échantillon n'a cette valeur. Une extrapolation des résultats au-delà des limites des

[3] La transformation utilisée (arc sinus et racine carrée) est employée en statistique pour des proportions et assure que les prédictions du modèle de régression sont dans l'intervalle [0,1] (ou [0%,100%] si on utilise les taux de réussite).

[4] Pour éviter une fausse augmentation de cette valeur par l'utilisation des poids, nous donnons ici la valeur de R^2 pour un modèle sans pondération.

valeurs utilisées pour les variables du modèle théorique n'est pas possible du point de vue statistique.

Malgré cette limite, la bonne qualité du modèle de régression nous permet de répondre positivement à notre seconde question de recherche et notre hypothèse est également confirmée. En effet, le taux de réussite peut être en partie prédit par l'outil d'estimation. Toutefois, d'autres facteurs peuvent également jouer un rôle, ce qui fera l'objet de notre discussion.

7. Discussion et conclusion

Dans cet article, nous nous sommes intéressées à l'outil d'estimation de la difficulté *a priori*, élaboré par Sánchez Abchi, De Pietro et Roth (2015) et plus particulièrement au modèle théorique qui est à la base de sa construction. En effet, le but de l'article était de vérifier sa fiabilité et son pouvoir prédictif. Plus précisément, la question était de savoir si cet outil, qui donne une estimation *a priori* de la difficulté, est capable de fournir une estimation fiable de la difficulté d'une question. Autrement dit, si, par exemple, une question qui est estimée difficile par l'outil sera également moins bien réussie par les élèves.

Les analyses statistiques réalisées ont confirmé qu'une augmentation de la difficulté au niveau du contenu et de l'enveloppe d'une question ainsi que du texte utilisé, engendrait une diminution de son taux de réussite, comme attendu. L'outil d'estimation *a priori* permet également de prédire, en partie, le taux de réussite des questions. Ainsi, le modèle de régression construit sur la base des variables de l'outil d'estimation explique une grande partie de la variance des taux de réussite des questions (54%).

Toutefois, d'autres facteurs pourraient jouer également un rôle dans la réussite des questions et donc influencer la difficulté des matériels évaluatifs de manière importante. Parmi ces facteurs, on trouve, dans la littérature, la familiarité des élèves avec les questions posées (Bachman, 2002; Timpe, 2013). En effet, comme le souligne Bachman, la difficulté d'une question est en lien avec deux composantes : le processus cognitif et la familiarité cognitive, qui sont tous les deux susceptibles de varier d'une personne testée à l'autre. Il faut également considérer l'incidence possible des pratiques enseignantes sur la familiarité avec la tâche évaluée (Denton et al., 2015; Kobayashi, 2009; Roth & Sánchez Abchi, 2020, 2020), et donc, avec la perception de la difficulté de la tâche par les élèves. Effectivement, les pratiques de l'enseignant·e vont amener les élèves à s'habituer à certaines formes de questions et à une manière de questionner le texte, qui peuvent moduler la perception de la difficulté (Bachman, 2002; Brindley & Slatyer, 2002). Ainsi, la prédiction de la difficulté donnée par l'outil doit être nuancée par les habitudes évaluatives des élèves. Par exemple, d'après l'outil, le format de questionnement à réponse textuelle est considéré comme difficile, car il exige la production d'une réponse d'une certaine longueur. Mais, si c'est le format privilégié de l'enseignant·e lorsqu'il ou elle pose des questions à ses élèves, les élèves seront habitués à cette pratique, ce qui peut avoir un impact sur la perception de la difficulté, et par conséquent conduire à une nuance du calcul. Les indices concernant les caractéristiques des tâches pourraient encore être nuancés par les pratiques sociales de la classe (Mottier Lopez, 2008).

L'impact des caractéristiques de l'élève n'est pas non plus à négliger dans l'estimation de la difficulté. Assurément le bagage cognitif de l'élève, sa motivation, ses connaissances préalables et son niveau d'attention ont un impact sur sa compréhension (Graesser et al.,

2011; McLaughlin & Overturf, 2012) et pourraient amener à nuancer la difficulté du matériel d'évaluation.

En outre, il est important de garder à l'esprit que les trois composantes analysées – contenu, enveloppe et texte – sont, à leur tour, constituées de dimensions spécifiques. Ainsi, la valeur du contenu est basée sur les valeurs des opérations cognitives, de l'ampleur de l'objet langagier et du nombre d'activités. L'enveloppe, quant à elle, s'appuie sur le format de la question et la consigne. Dans la forme actuelle de l'outil d'estimation, les dimensions ont le même poids dans la définition des variables théoriques. La question du poids des différentes dimensions qui composent chacune des variables théoriques reste à interroger de même que les éventuels liens entre elles. Des analyses plus approfondies, en particulier concernant les opérations cognitives et le format de questionnement, font l'objet d'une étude complémentaire (Roth & Mathei, en préparation).

Quant à la composante « texte », l'indice de difficulté exprimé sous forme de score permet une manipulation facile. Toutefois, comme pour les autres variables qui concernent les questions, celui-ci est une moyenne qui n'est pas forcément facile à interpréter. Une piste pourrait être de prendre en compte des variables plus « qualitatives », telles que la complexité linguistique ou la familiarité de contenu, sous forme d'accord interjuges.

Au-delà de la constitution des trois composantes de la difficulté, la question de l'interaction entre les caractéristiques du texte et les questions de compréhension de l'écrit se pose également. En effet, il est difficile d'isoler les effets d'une seule des variables (Brindley & Slatyer, 2002), étant donné qu'elles sont très liées. Ainsi, il a été observé, par exemple, que les questions littérales sont mieux réussies que les questions d'interprétation dans les textes qui racontent, alors que les questions d'interprétation ont été mieux réussies que les questions littérales pour les textes qui transmettent des savoirs (Eason et al., 2012).

Par ailleurs, les limites de notre étude ne nous permettent pas de généraliser les résultats. En effet, nous avons été tributaires des données fournies par le projet EpRoCom pour tester l'outil d'estimation. Le matériel n'a pas été conçu de manière expérimentale (comme dans PISA) avec un nombre similaire de questions pour chaque modalité d'un paramètre (ex. 40% de questions avec une consigne facile, 30% avec un niveau intermédiaire et 30% difficiles), mais nous nous sommes appuyées sur des épreuves existantes dont les propriétés et l'objectif étaient différents. Pour mieux contrôler l'impact de la difficulté sur le taux de réussite, il aurait fallu avoir un plan expérimental avec toutes les catégories de difficulté représentées ; et pouvoir également proposer des variantes de difficulté au niveau des textes pour observer le lien entre un texte et ses questions.

Malgré ces observations, les résultats de cette étude nous permettent de répondre positivement à nos questions de recherche et confirment la valeur du dispositif. En effet, l'outil d'estimation pourrait être une aide précieuse pour les enseignant·es en articulation avec d'autres critères davantage en lien avec les pratiques de classe. Le dispositif donne ainsi un panorama de difficultés potentielles au niveau des questions, que l'enseignant·e doit ensuite pouvoir réinterpréter à la lumière des connaissances qu'il·elle a de ses élèves. Autrement dit, l'outil devrait donner à l'enseignant·e de précieuses informations sur la difficulté du matériel évaluatif (textes et questions) qu'il·elle pense utiliser et ainsi lui permettre d'agir (en le modifiant ou l'ajustant) pour que celui-ci soit adapté à ses intentions évaluatives.

8. Références bibliographiques

Aeby, S., De Pietro, J.-F., & Wirthner, M. (2000). Français 2000. Dossier préparatoire. L'enseignement du français en Suisse romande : Un état des lieux et des questions. *Institut de recherche et de documentation pédagogique*.

Akib, E., Syatriana, E., & Ebrahimi, S. S. (2020). Principle for the Evaluation of Reading Assessment Tools. *Journal of Critical Reviews*, *7*(15), 5.

Anderson, N. J., Bachman, L., Perkins, K., & Cohen, A. (1991). An exploratory study into the construct validity of a reading comprehension test : Triangulation of data sources. *Language Testing*, *8*(1), 41-66. https://doi.org/10.1177/026553229100800104

Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing*, *19*(4), 453-476. https://doi.org/10.1191/0265532202lt240oa

Bachman, L. F., Davidson, F., & Milanovic, M. (1996). The use of test method characteristics in the content analysis and design of EFL proficiency tests. *Language Testing*, *13*(2), 125-150. https://doi.org/10.1177/026553229601300201

Bogaerds-Hazenberg, S. T. M., Evers-Vermeul, J., & Bergh, H. (2021). A Meta-Analysis on the Effects of Text Structure Instruction on Reading Comprehension in the Upper Elementary Grades. *Reading Research Quarterly*, *56*(3), 435-462. https://doi.org/10.1002/rrq.311

Brindley, G., & Slatyer, H. (2002). Exploring task difficulty in ESL listening assessment. *Language Testing*, *19*(4), 369-394. https://doi.org/10.1191/0265532202lt236oa

Cardinet, J. (1987). *L'objectivité de l'évaluation*. Institut romand de recherches et de documentation pédagogiques.

Cardinet, J. (1989). Évaluer sans juger. *Revue française de pédagogie*, *88*, 41-52. https://doi.org/10.3406/rfp.1989.1412

Chapelle, C. A. (2017). Conceptions of validity. In G. Fulcher & F. Davidson (Éds.), *The Routledge Handbook of Language Testing* (p. 21-33). Routledge.

De Pietro, J.-F., & Roth, M. (2017). A propos de la validité « didactique » d'une évaluation. *Evaluer. Journal international de Recherche en Education et Formation (e-JIREF)*, *3*(3), 31-50.

Denton, C. A., Enos, M., York, M. J., Francis, D. J., Barnes, M. A., Kulesz, P. A., Fletcher, J. M., & Carter, S. (2015). Text-Processing Differences in Adolescent Adequate and Poor Comprehenders Reading Accessible and Challenging Narrative and Informational Text. *Reading Research Quarterly*, *50*(4), 393-416. https://doi.org/10.1002/rrq.105

Eason, S. H., Goldberg, L. F., Young, K. M., Geist, M. C., & Cutting, L. E. (2012). Reader–text interactions : How differential text and question types influence cognitive skills needed for reading comprehension. *Journal of Educational Psychology*, *104*(3), 515-528. https://doi.org/10.1037/a0027182

Fletcher, J. M. (2006). Measuring reading comprehension. *Scientific Studies of Reading*, *10*(3), 323-330. https://doi.org/10.1207/s1532799xssr1003_7

François, T. (2009). Modèles statistiques pour l'estimation automatique de la difficulté de textes de FLE. RECITAL, Senlis, France.

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix providing multilevel analyses of text characteristics. *Educational researcher*, *40*(5), 223-234.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, *101*(3), 371-395.

Guay, A. (2011). Analyse de la formulation des consignes des épreuves adaptatives (EA) de la 3e année primaire (5e année HarmoS) en français et en mathématiques. Institut de recherche et de documentation pédagogique.

Gunning, R. (1952). *The technique of clear writing*. McGraw-Hill.

Howel, D. (2008). Méthodes statistiques en sciences humaines. De Boeck.

Kandel, L., & Moles, A. (1958). Application de l'indice de Flesch à la langue française. *Cahiers d'études de radio-télévision*, *19*, 253-274.

Kobayashi, K. (2009). The influence of topic knowledge, external strategy use, and college experience on students' comprehension of controversial texts. *Learning and Individual Differences*, *19*(1), 130-134. https://doi.org/10.1016/j.lindif.2008.06.001

Lafontaine, D. (2003). *Comment faciliter, développer et évaluer la compréhension des textes aux différentes étapes de la scolarité primaire ?* http://www.cndp.fr/bienlire/01-actualite/document/lafontaine.pdf

Laveault, D., & Grégoire, J. (2014). *Introduction aux théories de tests en psychologie et en sciences de l'éducation*. De Boeck Supérieur.

McLaughlin, M. & Overturf, B. (2012). The Hunger Games and the Common Core : Determining the Complexity of Contemporary Texts. *30*(3), 8-9.

McNamara, D. S., Graesser, A. C., & Louwerse, M. M. (2012). Sources of text difficulty : Across genres and grades. In J. P. Sabatini & E. Albro (Éds.), *Measuring up : Advances in how we assess reading ability* (pp. 89-116). Rowman & Littlefield Education.

Mesnager, J., & Bres, S. (2008). *Évaluer la difficulté des textes : Cycle 2-3, 6e-5e*.

Mottier Lopez, L. (2008). *Apprentissage situé. La microculture de classe en mathématiques*. Peter Lang.

Mullis, I. V. S., Martin, M. O., & Sainsbury, M. (2015). PIRLS 2016 Reading Framework. In I. V. S. Mullis & M. O. Martin (Éds.), *PIRLS 2016 assessment framework* (2nd éd., p. 11-30). Chestnut Hill, MA: TIMSS & PIRLS International Study Center.

Norris, J. M., Brown, J. D., Hudson, T. D., & Bonk, W. (2002). Examinee abilities and task difficulty in task-based second language performance assessment. *Language Testing*, *19*(4), 395-418. https://doi.org/10.1191/0265532202lt237oa

OCDE. (2000). *Mesurer les connaissances et les compétences des élèves : Lecture, mathématiques et science : l'évaluation de PISA 2000*. OECD. https://doi.org/10.1787/9789264281561-fr

OECD. (2010). *PISA 2009 Assessment Framework : Key Competencies in Reading, Mathematics and Science*. OECD. https://doi.org/10.1787/9789264062658-en

OECD. (2019). *PISA 2018 Assessment and Analytical Framework*. OECD Publishing.

Pini, G., Reith, E., Weiss, L., & Bugniet, F. (2006). *Guide méthodologique pour l'évaluation et la mesure en éducation*. Genève : Développement et innovation pédagogique au cycle d'orientation (DIPCO).

Rog, L. J., & Burton, W. (2002). Matching Texts and Readers : Leveling Early Reading Materials for Assessment and Instruction. *The Reading Teacher*, *55*(4), 348-356.

Roth, M., & Matei, A. (en préparation). Estimer la difficulté des questions posées aux élèves âgés de 10 à 12 ans pour évaluer leur compréhension de l'écrit : Quel lien avec le taux de réussite ?

Roth, M., & Sánchez Abchi, V. (2020). Évaluation la difficulté d'une épreuve de compréhension de l'écrit : Expérimentation d'un outil sur une pratique enseignante. 3. https://doi.org/10.48325/RLEEE.003.02

Sánchez Abchi Verónica, de Pietro Jean-François, & Roth Murielle. (2016). *Evaluer en français Comment prendre en compte la difficulté des items et des textes*. Institut de recherche et de documentation pédagogique. Neuchâtel.

Timpe, V. (2013). « The difficulty with difficulty : » The issue to determining task difficulty in TBLA. *Journal of Linguistics and Language Teaching*, *4*(1), 13-28.

Weiss, J., & Wirthner, M. (1991). Evaluation de la compréhension de l'écrit correspondant au niveau de fin de deuxième année primaire. In J. Weiss & M. Wirthner (Éds.), *Enseignement du français. Premiers regards sur une rénovation* (pp. 265-296). Delval.

Annexe

Annexe 1 : synthèse des valeurs attribuées à chaque dimension de difficulté du texte

DIMENSIONS **VALEURS**	**Complexité lexicale** Pourcentage de mots peu fréquents (Mesnager et Bres, 2008)	**Complexité syntaxique** Valeur de l'indice de lisibilité de Gunning	**Complexité de la structure du texte** Structure canonique (SC) (correspond-elle à ce qu'on attend du genre de texte concerné ?) et emploi d'organisateurs textuels
Facile (-1)	Jusqu'à 5%	6-8	- SC, sans organisateurs - Peu éloigné de la SC avec des organisateurs qui structurent davantage
Intermédiaire (0)	Entre 6 et 9%	9-11	- SC sans organisateurs, mais avec la nécessité d'en ajouter - Peu éloigné de la SC, avec des organisateurs qui perturbent - Sans SC avec des organisateurs qui structurent
Difficile (1)	Dès 10%	12-20	- SC avec des organisateurs qui perturbent la compréhension - Peu éloigné de la SC, mais sans organisateurs - Structure non canonique, sans organisateurs