

Communication 193.1 : Quel est l'impact de la difficulté des tâches évaluatives sur leur validité ?

Roth Murielle, Murielle.Roth@irdp.ch, Institut de recherche et de documentation pédagogique (IRDp), Suisse

Sánchez Abchi Verónica, Veronica.Sanchez@irdp.ch, Institut de recherche et de documentation pédagogique (IRDp), Suisse

De Pietro Jean-François, Jean-Francois.dePietro@irdp, Institut de recherche et de documentation pédagogique (IRDp), Suisse

Objectif de la communication

Cette contribution vise à explorer les liens entre la difficulté des tâches utilisées pour évaluer la compréhension écrite des élèves en français et leur validité. Autrement dit, nous souhaitons examiner comment et dans quelle mesure les éléments qui déterminent la difficulté d'une tâche de compréhension, ainsi que les variations que ces éléments sont susceptibles de subir, impactent la validité de l'évaluation. Dans notre perspective, didactique, c'est d'abord la question de la validité de contenu qui est en jeu : y a-t-il une adéquation entre ce que l'on cherche à évaluer et ce qu'on peut déduire de l'information recueillie à partir des résultats des élèves ? (de Pietro & Roth, 2017).

Cadre de référence

La compréhension écrite constitue un processus complexe faisant appel à des opérations cognitives diverses qui réalisent les différents aspects de ce processus. En effet, la lecture d'un texte suppose la mise en œuvre de processus de niveau inférieur, comme la reconnaissance des correspondances entre les phonèmes et les graphèmes, et de processus de niveau plus élevé, comme la réalisation d'inférences de cause ou l'interprétation du sens « caché » d'un texte (pour une synthèse voir Bianco, 2016). Si toutes les opérations sont nécessaires pour réussir à comprendre un texte, cela ne veut bien sûr pas dire pour autant que toutes les tâches d'évaluation visent les mêmes opérations et les mêmes aspects du processus. En fonction du contexte dans lequel l'évaluation est réalisée (degré scolaire ; type d'élèves ; but de l'évaluation ; manière dont l'information va être recueillie auprès des élèves), chaque épreuve définit finalement les aspects de la compétence de compréhension qu'elle va focaliser. Puis, une fois ces éléments définis, il s'agit d'élaborer des tâches qui évaluent précisément ce qu'on a décidé d'évaluer – au niveau d'exigence qu'on attend des élèves. Mais comment s'assurer que les tâches proposées aux élèves sont au niveau de difficulté souhaité par l'enseignant.e ?

Pour un meilleur contrôle, nous avons donc élaboré un outil *a priori* afin de définir les différents éléments qui contribuent à déterminer la difficulté d'une tâche en compréhension de l'écrit lorsque celle-ci est évaluée. Cet outil permet de caractériser le niveau de difficulté de la tâche selon une procédure formée de trois composantes incluant elles-mêmes plusieurs facteurs : le « contenu » de la tâche évaluative (opérations cognitives à effectuer par l'élève, ampleur de l'objet langagier et nombre d'activités que l'élève doit réaliser), son « enveloppe » (formulation de la consigne, format de questionnement et matériel complémentaire à disposition tels que dictionnaires, correcteur orthographique, etc.) et le « matériel langagier »

(texte, corpus...) auquel elle est attachée. (Sánchez Abchi, Roth, De Pietro, 2016; Roth, Sanchez Abchi & De Pietro, 2018)

La question se pose alors : comment – parmi d'autres facteurs – la difficulté des tâches utilisées pour évaluer la compréhension de textes peut avoir une incidence sur la validité de contenu ? C'est selon nous la manipulation des éléments définis comme ayant une incidence sur la difficulté des tâches qui devrait permettre de rendre visible cet impact – fort, modéré, voire nul. Nous faisons en effet l'hypothèse que certains de ces éléments pourraient avoir une incidence décisive sur la validité de contenu et *a fortiori* de l'épreuve : en fonction de la manière dont est formulée la consigne et dont sont rendues visibles les opérations cognitives en jeu, en fonction du format de questionnement utilisé, nous devons donc nous demander si on évalue toujours le même objectif.

Méthodologie

Sur la base de notre outil, nous avons analysé de manière approfondie une épreuve de compréhension de l'écrit – destinée à des élèves de 9^e secondaire (âgé.e.s de 12-13 ans) – élaborée par une enseignante en trois versions afin qu'elle puisse être passée par des élèves de différents niveaux : pré-gymnasial (A : haut niveau), moderne (B : niveau intermédiaire) et général (C : bas niveau). Nous avons également étudié les résultats de 58 élèves pour les trois niveaux.

Dans un premier temps, nous avons identifié les éléments potentiellement transformables d'après notre outil d'estimation de la difficulté. Autrement dit, nous avons observé les différents éléments de notre modélisation : les opérations censées être mobilisées par l'élève pour répondre aux différentes questions, l'ampleur de l'objet langagier, le nombre d'activité à réaliser par l'élève, la formulation de la consigne et le format de questionnement. Nous avons également estimé le niveau de difficulté du texte à comprendre – un article de journal qui traite de la thématique du réchauffement climatique – compte tenu de la difficulté du vocabulaire, du niveau de complexité syntaxique et de la complexité de sa structure.

Dans un deuxième temps, nous avons comparé ces éléments avec ceux qui ont été effectivement modifiés par l'enseignante dans les trois versions de l'épreuve. Cette comparaison devait donc, d'une part, nous permettre de saisir dans quelle mesure notre modélisation rejoint les représentations que la conceptrice de l'épreuve se fait de la difficulté de son épreuve et des éléments qu'elle-même prend en considération pour la faire varier et, d'autre part, en comparant les résultats des élèves, d'entrevoir l'impact de ces éléments sur leurs performances.

Les résultats

Plusieurs éléments que l'enseignante fait varier rejoignent ceux de notre outil :

- le type d'opération cognitive à mobiliser : lorsqu'il y a une différence entre les trois niveaux il est moins difficile pour les niveaux B et C que pour le niveau A; par exemple, pour le niveau A, l'élève doit réaliser une inférence pour répondre à la question alors que pour les niveaux B et C, il lui suffit de repérer l'information qui est donnée dans le texte ; de plus, pour le niveau C, le format de questionnement est simplifié (QCM) par rapport au niveau B (question ouverte) ;
- le nombre d'activité à réaliser : dans les questions qui diffèrent d'un niveau à l'autre, l'élève a souvent qu'une seule activité à réaliser pour les niveaux B et C alors qu'il en a au moins deux pour le niveau A ;

- le format de questionnement : l'épreuve de niveau A comporte 19 questions ouvertes, celle de niveau B 16 questions ouvertes et deux QCM et, pour le niveau C, elle comprend 10 questions ouvertes et quatre QCM.
- la formulation de la consigne : l'enseignante pose des questions moins précises, parfois ambiguës, pour le niveau A (par rapport aux autres niveaux), qui pourraient dès lors donner lieu à plusieurs réponses possibles alors même qu'une seule réponse est attendue de la part des élèves. Les résultats des élèves ont à ce propos montré qu'une grande majorité d'entre eux n'ont pas donné tout à fait la « bonne réponse » et ont été pénalisés – ce qui laisse bien entendre que ce format contribue à la difficulté de l'item. En outre, on constate que les questions sont plus nombreuses pour les élèves de niveau A (19) que pour les élèves de niveau C (14) avec le même temps à disposition. Toutefois, la longueur de l'épreuve (le nombre de questions) constitue un élément que nous n'avons pas pris en compte puisque notre outil a été créé dans une logique de tâche et non d'épreuve comme c'est le cas ici.

Un début de discussion

Après avoir abordé la question de la difficulté des tâches utilisées ici pour évaluer la compréhension écrite des élèves, il s'agit maintenant de s'interroger sur l'impact que celle-ci peut avoir sur la validité même de l'épreuve. Ce lien est loin d'être évident. Mais il nous semble tout de même, par exemple, qu'un certain nombre de questions adressées aux élèves de niveau A posent un problème de validité didactique : nous avons parfois le sentiment que – dans les questions ouvertes – c'est peut-être moins la compréhension écrite de l'élève qui est évaluée que sa capacité à deviner ce qui est attendu... En effet, plusieurs réponses nous paraissent dans certains cas possibles alors qu'une seule est considérée comme acceptable. Il y aurait donc inadéquation entre la demande qui émerge de la consigne et ce qui est visé, compromettant ainsi la validité de l'épreuve, du moins des questions concernées.

Bien que nous puissions observer quelques faits de ce type, nous ne pouvons évidemment pas affirmer que les éléments qui déterminent la difficulté d'une tâche de compréhension ont un impact direct sur la validité de l'épreuve : difficulté et validité ne sont pas directement liées. Toutefois, notre outil d'estimation de la difficulté nous a permis de rendre visible un problème de validité qui est lié au construit de la compréhension : il nous est apparu en effet que la profondeur de compréhension du texte telle qu'elle est évaluée ici n'est pas la même dans les trois versions de l'épreuve. Plus les questions de l'épreuve sont faciles – avec beaucoup de repérage facile à corriger – plus la validité est forte. Il y a donc bien adéquation entre la demande et la réponse attendue mais... on reste à un niveau de compréhension de surface. Au contraire, plus les questions sont difficiles – avec davantage de questions ouvertes, qui demandent de faire des inférences, d'interpréter le texte, et qui sont donc plus compliquées à corriger – plus la validité devient problématique : la marge d'interprétation concernant ce qui est défini et accepté comme bonne réponse s'accroît en même temps qu'on vise à un niveau de compréhension plus profond. Il nous semble qu'il y a là une vraie question pour celles et ceux qui travaillent à l'évaluation de la compréhension...

Mots-clés

Difficulté des tâches, validité de l'évaluation, compréhension de l'écrit

Références bibliographiques

- Bianco, M. (2016). Du langage oral à la compréhension de l'écrit. Grenoble : Presses universitaires de Grenoble.
- 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*, 3(3), 31-50.
- Roth, M., Sanchez Abchi, V. & De Pietro, J.-F. (2018). L'évaluation de la compréhension de l'écrit en français : comment estimer la difficulté des items et des textes proposés aux élèves ? In M. Demeuse, M. Milmeister & Ch. Weis (éds), *L'évaluation en éducation et en formation face aux transformations des sociétés contemporaines : actes du 30ème colloque de l'ADMEE-Europe*, 10-12 janvier 2018, Luxembourg (pp. 468-469). Luxembourg : Campus Belval à Esch-surAlzette.
- Sánchez Abchi, V., De Pietro, J.-F. & Roth, M. (2016). *Evaluer en français : comment prendre en compte la difficulté des items et des textes*. Neuchâtel : Institut de recherche et de documentation pédagogique (IRDP).