Analyser la validité des évaluations au sein du réseau EVADIDA (Evaluations et Didactique) de l'ADMEE : approches et exemples en mathématiques et en lecture

Nadine Grapin - Laboratoire de Didactique André Revuz - Université Paris Est Créteil

Marc Vantourout – Laboratoire Education, Discours, Apprentissages – Université de Paris

Brigitte Grugeon-Allys - Laboratoire de Didactique André Revuz - Université Paris Est Créteil

Le réseau EVADIDA a été créé au sein de l'ADMEE-Europe en 2013-2014 afin de rapprocher didacticiens et spécialistes de l'évaluation pour que les contenus évalués soient effectivement pris en compte dans les travaux consacrés à l'évaluation. Les différentes didactiques disciplinaires, mais aussi la didactique professionnelle, peuvent être impliquées.

Avant 2015, quelques travaux en didactique des mathématiques traitaient explicitement d'évaluation, alors que l'intérêt des didacticiens du français pour ce domaine s'est manifesté dès la fin des années 1970. Ces dernières années, les mathématiques très largement prises en compte puis le français et les sciences, sont les trois « disciplines » les plus travaillées dans le cadre EVADIDA.

Une synthèse des travaux menés au sein du réseau EVADIDA

Nous résumons ici des éléments d'une synthèse réalisée à partir de l'analyse d'un corpus de vingt et un articles publiés au sein des deux revues de l'ADMEE[1], principalement entre 2015 et fin 2019 (Vantourout, à paraître). Ces articles, qui intègrent des considérations didactiques et sont en lien avec des apprentissages scolaires, portent sur des évaluations en « classe interne » (des évaluations conçues par un enseignant pour organiser et certifier les apprentissages en classe), en « classe externe » (des évaluations non conçues par un enseignant et ne répondant pas à une demande institutionnelle) et à « grande échelle » (toutes les formes d'évaluations institutionnelles standardisées). Plus de la moitié de ces articles concernent des tests à grande échelle.

Deux visées complémentaires émergent de cette synthèse :

D'une part, *outiller et comprendre*. Il s'agit à la fois d'outiller enseignants et formateurs afin de faciliter leur travail d'évaluation et de formation, mais aussi de comprendre les pratiques évaluatives (conception des épreuves, recueil d'informations sur ce qui est évalué, interprétation des résultats, décisions prises). Des démarches et méthodes sont proposées pour mener des analyses qualitatives des épreuves. La contribution des didactiques y est essentielle pour disposer de diagnostics qui informent qualitativement sur les connaissances et compétences des évalués ou des évaluateurs sur un domaine disciplinaire donné. Les articles s'organisent autour d'un important travail empirique qui repose sur des apports des didactiques. Ceux-ci se manifestent lors de l'identification de l'objet d'évaluation et l'étude des jugements évaluatifs et de leurs processus.

D'autre part, *la notion de validité* qui figure dans plus de la moitié des articles. Dans le domaine de la psychométrie où elle a été initialement définie, la validité est définie comme un critère garantissant que le test mesure ce qu'il est censé mesurer. Différentes preuves de validité définies *a posteriori* de la passation d'un test ont été listées, en lien par exemple avec la couverture du domaine, les processus de réponse ou encore avec la structure interne du test. Chacune de ces preuves est accompagnée d'indicateurs statistiques (difficulté, discrimination par exemple) ou d'indices permettant de les

[1] Il s'agit de « Mesure et évaluation en éducation » (MEE) et de « Évaluer, Journal international de recherche en éducation et formation » (e-JIREF).

garantir ou non. Pour nous, aborder la question de la validité dans un cadre didactique permet non seulement d'intégrer des aspects épistémologique et psychologique dans l'analyse *a priori* du contenu et des processus de réponses mais aussi de la concevoir comme interrogeant *a posteriori* la pertinence et l'utilité du contrôle évaluatif pour la régulation de l'enseignement et de l'apprentissage dans le cadre d'une séquence didactique.

Les relations entre les approches didactiques et psychométriques constituent une question centrale. Les psychométriciens s'appuient plutôt sur des analyses *a posteriori*, basées sur les scores de réussite, lors de la démarche de validation psychométrique d'un instrument de mesure, alors que les didacticiens insistent sur la dynamique entre études *a priori* / *a posteriori*. Si la sélection des items reste le principal point d'achoppement, la complémentarité de ces deux approches commence cependant à faire ses preuves dans la conception d'évaluations diagnostiques et de bilans.

Dans les deux parties suivantes, nous nous arrêtons sur deux recherches qui intègrent une conception renouvelée de la validité et qui l'opérationnalisent. Nous présentons en conclusion d'autres axes de réflexion actuels.

Appréhender la validité d'une évaluation à partir d'une approche anthropologique et didactique : un exemple en mathématiques

La méthodologie d'analyse de la validité que nous présentons (Grapin & Grugeon-Allys, 2018) fait suite à des travaux antérieurs que nous avons menés autour de la conception de diagnostics en algèbre (Grugeon, 1997 ; Grugeon-Allys, 2016) et en arithmétique (Grapin, 2015) puis de l'analyse des évaluations CEDRE.

Afin de pouvoir étudier la validité didactique d'une évaluation, nous avons tout d'abord cherché à définir un référent épistémologique nous permettant non seulement d'étudier le référent au sens des concepteurs d'évaluations (que ce soit dans le cas d'évaluations externes à grande échelle ou d'évaluation internes conçues et mises en œuvre par l'enseignant dans sa classe) mais aussi de prendre en compte le processus de transposition didactique. Pour ce faire, nous avons choisi de nous placer dans le cadre de la théorie anthropologique du didactique (Chevallard, 1999) et de définir, sur un domaine mathématique donné, un modèle praxéologique de référence (MPR) (Bosch & Gascon, 2005). Nous faisons l'hypothèse que ce choix théorique permet l'étude de la validité didactique d'une évaluation. Les connaissances et savoirs mobilisés dans le cadre d'une activité mathématique, et qui dans notre cas, seront évalués, sont ainsi modélisés par des praxéologies, formées par des types de tâches, des techniques pour les effectuer, des discours technologiques pour justifier, expliquer, décrire les techniques et des éléments théoriques qui apportent un dernier niveau de justification aux technologies. La définition de ce modèle praxéologique de référence (MPR) pour un domaine mathématique, permettra de décrire les attentes, en termes de types de tâches et de savoirs mathématiques, relatives à une institution donnée (que ce soit par exemple le ministère, la circonscription, l'établissement, la classe) et les activités des élèves, tout en le complétant par les erreurs des élèves : nous utilisons ainsi ce MPR pour concevoir et étudier les tâches de l'évaluation (couverture du domaine, pertinence et complexité des tâches) mais aussi pour analyser les productions des élèves.

La définition d'une praxéologie de référence sur les domaines de l'arithmétique en fin d'école et de l'algèbre en fin de collège nous a ainsi conduit à modéliser ces domaines : le premier, à partir de praxéologies portant sur la résolution de problèmes arithmétiques, la numération et le calcul et le second, à partir de praxéologies de modélisation, de preuve et de calcul relatives aux expressions algébriques, aux équations.

Chacune des tâches d'une évaluation peut ainsi être décrite par :

- le type de tâche et la/les technique(s) attendue(s) en lien avec les éléments technologico-théoriques qui les sous-tendent ;
- les objets mathématiques en jeu, leur ancienneté et leur nombre ;
- les variables didactiques associées aux objets mathématiques en jeu et leurs valeurs ;
- les registres de représentation sémiotiques en jeu ;
- le nombre de types de tâches convoqués dans la résolution et le type de convocation (par la tâche ou par l'élève) définissant du côté épistémologique la complexité de la tâche.

D'autres éléments d'analyse peuvent intervenir dans cette analyse *a priori* comme la nature du format des questions, la nature du contexte dans le cas d'une situation extra-mathématiques, etc. ; ce qui est développé davantage dans une approche psycho-didactique.

A partir de l'analyse praxéologique de chacune des tâches, il est possible de déterminer pour une institution donnée leur pertinence, à partir de l'étude des techniques et des technologies, c'est-à-dire, si sa résolution mobilise les savoirs qui doivent être évalués et donc apporter de premières preuves de validité. L'analyse transversale de l'ensemble des tâches d'une évaluation permet ensuite d'étudier la couverture du domaine mathématique selon les types de tâches définis dans la praxéologie de référence. Ce deuxième type de preuve de la validité, à l'échelle globale du test, peut être complété par une étude de la complexité des tâches mais aussi des technologies sous-tendant chacune des techniques. Ces deux derniers éléments (complexité – technologies) sont particulièrement importants dans la conception d'évaluations conduisant à la construction d'échelles de score, comme le sont les évaluations CEDRE sur lesquelles nous avons travaillé.

Nous montrerons durant notre présentation et à l'aide d'exemples issus des évaluations CEDRE fin d'école et fin de collège la façon dont nous avons opérationnalisé cette méthodologie pour analyser le contenu de ces évaluations ainsi que leurs résultats. Nous montrerons aussi comment cette approche didactique peut s'articuler sur une dynamique *a priori / a posteriori* avec les méthodologies statistiques qui sous-tendent les évaluations à grande échelle, en particulier en revenant sur la distinction complexité (*a priori)* et difficulté (*a posteriori* à partir des scores de réussite*)*. Nous conclurons enfin cette partie sur l'exploitation de notre méthodologie pour étudier les pratiques d'évaluation des enseignants en classe, au quotidien et pour réaliser une typologie des pratiques enseignantes lors de l'enquête PRAESCO en parallèle de l'évaluation CEDRE (3e).

Approche psycho-didactique : un exemple en lecture

La validité psycho-didactique (VPD), qui se détache de toute considération psychométrique, est une approche qualitative et renouvelée qui a pour origine un questionnement « local » sur la validité des évaluations – de chaque question ou item – qui sont utilisées en classe par les enseignants (Vantourout & Goasdoué, 2014). L'évaluation vise dans ce cadre psycho-didactique à décrire/qualifier le fonctionnement cognitif à partir de l'activité de réponse de l'évalué qui est pour une part déterminée par le couple sujet-situation. La VPD conjugue les apports des didactiques disciplinaires et de l'ergonomie cognitive : on complète l'analyse conceptuelle, impliquant une analyse *a priori*, par une analyse de la tâche, en s'interrogeant sur l'activité de réponse que la tâche et ses spécificités peuvent engendrer.

La VPD concerne à l'origine l'évaluation en classe et renvoie à l'activité de l'enseignant-évaluateur et à des préoccupations évaluatives telles que s'assurer de la validité des évaluations proposées à ses élèves. Notons que :

- ces préoccupations ne sont pas largement partagées par les enseignants (ni d'ailleurs par leur encadrement), avant qu'ils ne soient sensibilisés à cette question ;

- s'assurer de la VPD vaut aussi pour des évaluations externes, pour la qualité des diagnostics produits, mais aussi pour pouvoir si on le souhaite utiliser en classe les « résultats » des évaluations ; cela vaut indépendamment du rôle de l'évaluation (sommative, formative, …), la question de la validité étant pour nous à privilégier par rapport à celles relatives à ces différentes formes.

A l'origine, la VPD a été développée pour analyser des épreuves de mathématiques ; nous avons aussi exploité cette approche de la validité pour concevoir un dispositif d'évaluation – une ingénierie évaluative courte – de la lecture en fin d'année au CP. Pour cela, nous avons fabriqué un texte et trois épreuves, une de fluence, et deux de compréhension, rappel de récit et questionnaire avec demande de justifications.

Pour s'assurer au mieux de la validité des épreuves, condition nécessaire (mais pas suffisante) pour accéder à des jugements valides, nous prenons en compte lors de leur élaboration (et du choix du texte), des éléments théoriques et résultats relatifs au fonctionnement cognitif des élèves dans le domaine de la lecture (Vantourout & Blanc, 2017 ; Vantourout & Maury, 2017). Ces épreuves sont en adéquation avec notre conception du « savoir-lire » et font l'objet d'une analyse *a priori*. Cette analyse permet d'apprécier le niveau de complexité que pose, pour un élève de CP, l'identification des mots du texte (score de difficulté *a priori* selon deux facteurs calculés à partir de critères issus de la littérature et renvoyant à l'identification automatique et à une procédure orthographique). Elle permet, d'autre part après analyse du texte du point de vue de sa compréhension, d'anticiper les stratégies de réponse (et de justification) aux questions. Nous distinguons plusieurs types de processus en référence au(x) niveau(x) de représentation sollicité(s), base de texte, modèle de situation.

Il est intéressant de noter que nous nous appuyons ici sur des modèles globaux, valables pour tout type de texte, à la différence de modèles praxéologiques de référence, exploités en mathématiques et qui sont spécifiques du domaine évalué.

Lors de la communication, les épreuves et quelques résultats seront brièvement présentés mais nous nous arrêterons sur l'usage qui est fait en formation de formateurs (master PIF) depuis 2018 de cette recherche. Nous avons observé que la prise en compte de cette approche didactique de l'évaluation faisait évoluer le rapport à l'évaluation de collègues en formation. Nous montrerons comment ils l'utilisent pour évaluer des élèves en début de CE1 et procèdent ainsi à l'analyse des productions des évalués (fluence, rappel de récit et questionnaire) ; ils « saisissent » alors les apports d'une approche didactique de l'évaluation et semblent convaincus de l'importance de s'interroger sur la validité des épreuves. Ils comprennent notamment que celle-ci, grâce aux apports didactiques, peut-être autre chose qu'un construit (comme c'est le cas en psychométrie et dans la majorité des évaluations à grande échelle).

Conclusion

L'approche didactique de l'évaluation – l'évaluation est à l'origine une question pédagogique – doit pour se développer, en particulier lorsque les enseignants sont polyvalents, parvenir à articuler une entrée disciplinaire, centrée sur les contenus évalués, et une transversalité en termes d'approches, de méthodes, de notions, c'est-à-dire articuler le didactique et le pédagogique.

Nous reviendrons d'abord sur le transfert de ces deux approches dans d'autres disciplines et sur les questions qui apparaissent, notamment en termes d'épistémologie des disciplines : si la définition d'un modèle épistémologique de référence est en cours de réalisation pour d'autres domaines mathématiques (géométrie), il semble difficile de procéder avec une même approche pour analyser la validité d'une évaluation en lecture/compréhension (Blanc, 2017). Fagnant et ses collègues (Fagnant,

2008 ; Auquiere & coll., 2018) ont par ailleurs montré par des travaux menés au sein du réseau que l'évaluation de connaissances mobilisées en résolution de problèmes, surtout dans les petites classes, est complexe puisqu'elle mobilise à la fois des connaissances en mathématiques mais aussi potentiellement en lecture/compréhension. Leurs résultats nous permettront de montrer, à titre d'exemple, l'intérêt d'analyser conjointement validité épistémo et psycho didactique.

Concevoir des évaluations valides pour la classe est une question partagée au sein du réseau, tout comme celle de l'étude des pratiques évaluatives en classe, en particulier à travers le lien entre évaluation et moments de l'étude (Grugeon-Allys, 2019). Nous reprendrons ces questions à travers celle de la formation et de son articulation avec des recherches finalisées impliquant des ingénieries évaluatives. Nous illustrerons notre propos avec des exemples de formation que nous avons mises en œuvre et en présentant le projet que nous menons actuellement autour de la validité de tâches d'évaluation en mathématiques et en français au cycle 2 en expliquant comment il répond à une double préoccupation : apporter des éléments scientifiques permettant de mieux définir et cerner la notion de validité didactique des évaluations mais aussi outiller les enseignants et formateurs pour leur pratique au quotidien.

Bibliographie partielle

Grapin, N., Grugeon-Allys, B. (2018) Approches psychométrique et didactique de la validité d'une évaluation externe en mathématiques : quelles complémentarités ? Quelles divergences ? *Mesure et évaluation en éducation.* 41-2, 37-66.

Grugeon-Allys, B. (2019). Evaluer en mathématiques : des pistes pour un enrichissement réciproque de la recherche en didactique des mathématiques et en évaluation. In S. Coppe, E. Roditi &al. (éds) *Géométrie et Evaluation,* La Pensée Sauvage – Editions.

Grugeon-Allys, B. (2016) Modéliser le profil diagnostique des élèves dans un domaine mathématique et l'exploiter pour gérer l'hétérogénéité des apprentissages en classe : une approche didactique multidimensionnelle. *Evaluer. Journal international de Recherche en Education et Formation,* 2.2, 63-88 http://e-jref.org/index.php?id=137.

Vantourout, M. (à paraître). Évaluations et didactiques, validité(s) didactique(s) : quelles synergies ? Dans, Younes, N., Grémion, C. & Sylvestre, M. (éds), *Évaluation, sources de synergies ?*

Vantourout, M. & Blanc, C. (2017). Une approche psycho-didactique pour évaluer la lecture au cours préparatoire ? *Repères*, 56-1, 215-231.

Vantourout, M. & Maury, S. (2017). Évaluation de la lecture au CP : mise en oeuvre d'une approche multiple. *Éducation et didactique*, 11-1, 45-62.

Vantourout, M., & Goasdoué, R. (2014). Approches et validité psycho-didactiques des évaluations. *Education et Formation*, e-302, 139-156.