Fixer un seuil de suffisance pour une épreuve de maîtrise : apports et limites de la méthode d'Angoff

Setting a passing score for a mastery test: benefits and limits of the Angoff method

Daniel Bain
Groupe Edumétrie, Société suisse pour la recherche en éducation (SSRE)[1]
daniel.bain@bluewin.ch

Résumé
La recherche dont nous rendons compte se situe dans le cadre de l'édumétrie, de la mesure des apprentissages scolaires. Lorsqu'on recourt à des tests critériés (*criterion referenced tests*), un des problèmes qui se posent de façon cruciale est l'élaboration d'un barème, et plus particulièrement la fixation d'un *seuil de suffisance* (de réussite ou de passage) sur l'échelle de l'épreuve. Les pratiques observées laissent soupçonner une bonne part d'arbitraire dans la détermination de seuils ou de standards, y compris pour des contrôles à enjeux élevés. Pour diminuer cet arbitraire, la *méthode d'Angoff modifiée* demande à un panel d'experts d'estimer item par item la probabilité de réussite d'apprenants « minimalement compétents » (juste suffisants). La procédure se déroule en deux ou trois étapes (*rounds*) entre lesquelles les experts reçoivent différentes informations et s'efforcent, lors de la discussion, de diminuer leurs divergences d'estimation.
Pour tester l'intérêt, les problèmes et les limites de cette méthode, nous l'avons appliquée à un examen de grammaire passé par une centaine de futurs instituteurs à la fin de leur formation universitaire. Le panel des experts était constitué de dix formateurs universitaires en didactique de la grammaire. Conformément à la méthode, le seuil final a été calculé à partir de la moyenne des estimations des dix experts en fin de procédure. Dans une discussion conclusive, à la lumière de notre expérience, nous faisons une analyse critique de la méthode à l'intention des chercheurs qui souhaiteraient l'appliquer.

Mots-clés
Tests à référence critériée – tests à enjeux élevés – fixation de standards – détermination du seuil de réussite – méthode d'Angoff modifiée – marge d'erreur.

Summary
The research reported here belongs to the domain of edumetrics, the measurement of learning outcomes**.** Whenever we use criterion-referenced tests, one of the crucial problems is to set various standards of achievement and, more particularly, to determine a *passing score* or *cut score* on the test scale. The observed practices suggest a significant part of arbitrary decisions when it comes to set thresholds or standards, including for high-stakes assessments**.** To reduce arbitrary decisions, the *modified Angoff method* asks a panel of experts to estimate, item by item, the probability of success of a

[1] Nous exprimons toute notre reconnaissance aux membres de ce groupe, qui ont suivi cette recherche dès sa conception. Ils nous ont fait bénéficier de leurs précieux conseils tout au long de son déroulement et de la rédaction du présent texte. Nos remerciements vont donc à Weimar Agudelo, Marion Dutrevis, Dagmar Hexel, Gianreto Pini, Emiel Reith, Anne Soussi et Laura Weiss.

"minimally competent examinee" (just sufficient). The procedure is conducted in two or three *rounds*. The experts receive different types of information between each round and then, through discussion, endeavour to reduce rating discrepancies.
In order to test the interest, the problems and limits of this method, we have applied it to a grammar exam given to roughly a hundred future primary school teachers at the end of their academic studies**.** The expert panel was composed of ten university teachers in the field of grammar didactics. According to the method, the final passing score was set using the mean of the estimates by all ten experts at the end of the procedure. In a conclusive discussion – in light of our experience – we develop a critical analysis of the method, aimed at researchers intending to use it.

Keywords
Criterion-referenced tests – high-stakes testing – standard setting – passing score – modified Angoff method – generalizability theory – margin of error.

Pour citer cet article : Bain, D. (2017). Fixer un seuil de suffisance pour une épreuve de maîtrise : apports et limites de la méthode d'Angoff. *Evaluer. Journal international de Recherche en Education et Formation, 3*(3), 69-95.

1. Introduction

1.1 Tests de maîtrise vs tests de niveau

L'examen que nous utiliserons comme fil rouge pour l'illustration de la méthode d'Angoff se présentait comme un *test de maîtrise* de type sommatif. Précisons donc ce que nous entendons par test de maîtrise en reprenant la définition de Cardinet et Tourneur (1985, p. 252) avant de décrire les modalités et le contenu de l'examen lui-même.

« Un test de maîtrise possède les propriétés suivantes :

1. Au contraire des tests classiques, la performance d'un étudiant n'est pas située par rapport à une performance moyenne d'un groupe qui sert de référence ; elle est comparée à un seuil absolu de réussite dans un univers de tâches.
2. L'univers doit être suffisamment bien défini pour qu'on puisse en extraire un échantillon aléatoire de tâches ou d'items, et surtout donner une définition précise à la performance qui est observée. Le modèle statistique que nous utilisons [dans les analyses de généralisabilité] (l'analyse de variance) ne suppose pas l'homogénéité à l'intérieur de l'univers, mais seulement l'échantillonnage aléatoire des items.
3. L'intérêt de l'examinateur étant de savoir si le score univers de l'étudiant (le score que ce dernier obtiendrait si, au lieu d'aborder un échantillon d'items, il abordait tous les items de l'univers d'items) est situé au-dessus ou en dessous du critère de maîtrise, la performance de l'étudiant n'a d'intérêt que dans la mesure où elle permet d'estimer le score univers (ou score vrai). »

Dans le présent texte, nous considérerons plus particulièrement les épreuves estimant l'état des connaissances ou des compétences des apprenants à la fin d'une des étapes de la formation et se présentant comme des tests de maîtrise. Nous aurons en tête plus spécifiquement les *contrôles sommatifs à enjeux élevés* passés à une étape cruciale de cette formation où une décision importante doit être prise qui engage la suite de la carrière scolaire ou professionnelle de l'apprenant ou une modification majeure du curriculum.

Dans la conception que nous nous donnons du test de maîtrise, celui-ci devrait alors avoir en principe les caractéristiques suivantes :

- en tant qu'épreuve sommative et certificative, porter strictement sur les acquis d'apprentissage (*learning outcomes*)[2] qui ont fait l'objet, sous une forme ou sous une autre, d'un contrat pédagogique entre l'enseignant et ses élèves ; concerner exclusivement les connaissances ou compétences qui ont été enseignées ou entraînées, à l'exclusion par exemple de questions dites d'aptitudes ou destinées à distinguer - voire sélectionner- les meilleurs candidats ;
- comporter une échelle d'*évaluation critériée*[3], souvent en pourcents de réussite.

Nous distinguerons deux types de contrôles de maîtrise selon qu'ils se situent

- *sur le plan individuel*, à la fin d'une étape de la formation des apprenants, caractérisant les résultats de chacun d'entre eux, généralement en prévision de l'accès à l'étape suivante, parfois comme condition à cet accès ; on peut citer à titre d'exemples les examens de fin d'études, secondaires ou universitaires, ou, pour la Suisse romande, les épreuves cantonales dites de référence recensant périodiquement les connaissances et compétences des élèves à des étapes clés de la scolarité obligatoire ;
- *sur le plan collectif ou institutionnel*, les enquêtes visant à établir un bilan des acquis d'un ensemble d'apprenants à l'intention des autorités scolaires ou politiques à des fins de « rendre compte » (*accountability*) et pour contribuer au monitorage du système scolaire. On peut citer à ce sujet *les épreuves romandes communes à l'Espace romand de la formation* (EpRoCom ; Marc & Wirthner, 2012 et 2013)[4] ainsi que les contrôles prévus par le concordat *HarmoS* pour vérifier les objectifs ou standards nationaux de formation pour la scolarité obligatoire (CDIP, 2007). On retrouve des épreuves de mêmes types, à certaines variantes près, en Belgique, France ou au Canada (cf. Yerly, 2014, chap. 4).

Si l'enseignement a été donné et reçu dans de bonnes conditions, le degré de réussite attendu de chaque question d'un test de maîtrise devrait être relativement homogène, dans l'idéal pas trop éloigné de 100%, et la distribution des scores de type courbe en J. Le standard de performance défini prioritairement est un *seuil de suffisance* (dit *de maîtrise, de réussite ou de passage,* selon le cas ou le contexte) pour distinguer dichotomiquement échecs et réussites. Les différents échelons de l'échelle considérée sont généralement définis subsidiairement (souvent en notes) à partir de ce point de repère.

Il est important pour la suite de notre propos de distinguer *tests de maîtrise* et *tests de niveau* (Cardinet, 1972). Souvent de type critérié eux aussi, ces derniers visent à distinguer différents degrés de performances, généralement en vue d'une orientation dans une filière de formation (cours à niveau, section), ou dans le cas où tel standard (autre que la simple suffisance) est exigé pour l'accès à une formation. On peut citer, parmi d'autres à titre d'exemples, les tests situant le niveau de compétence linguistique (de A1 à C2) des apprenants par rapport au *Cadre européen de référence pour l'enseignement des langues* (COE, 2017) ou, sur le plan collectif, les enquêtes PISA (Programme international pour le suivi des acquis des élèves)[5], avec ses 6

[2] Sur ce concept, cf. notamment D'Hoop E., Lemenu D., Malhomme, Chr. Coupremanne, M. (2012).
[3] vs normatives comme dans les tests psychologiques (échelles en centiles, rangs sur cent, stanines).
[4] Cf. l'article 15, al.1 de la Convention scolaire romande (CIIP, 2007) :
CIIP : Conférence intercantonale de l'instruction publique de la Suisse romande et du Tessin.
[5] Cf. http://www.oecd.org/pisa/ Noter que ces tests ne se réfèrent pas spécifiquement aux plans d'études nationaux.

niveaux de compétences. Sur le plan docimologique, ces *tests de niveau* (*de compétence, de performance*) se caractérisent, par construction, par toute une gamme de questions de difficultés différentes, ajustées aux différentes orientations visées et aux exigences correspondantes en termes de connaissances et de compétences. La distribution des résultats attendue devrait en principe prendre la forme d'une courbe de Gauss faiblement acuminée et ces contrôles impliquent la fixation de différents seuils ou standards correspondant aux catégories de performance ou aux orientations envisagées. Comme Kane (1994), nous distinguerons donc *score de passage* (*passing score*), pour nous *seuil de suffisance*, et *standard de performance* (*performance standard*). Au niveau de leur conception même, on ne peut donc considérer le test de maîtrise comme un cas particulier du test de niveau ne comportant que deux niveaux (suffisant – insuffisant) ; le degré de dispersion des scores distingue clairement les deux types de contrôle.

Dans ce qui suit, nous ne traiterons pas spécifiquement des tests de niveau ; nous y ferons cependant allusion, à l'occasion, à titre de contraste avec les tests de maîtrise.

1.2 De l'arbitraire des barèmes d'épreuves critériées : un problème d'arbitrage

> *« An absolute [criterion based] standard determines the pass/fail outcome by how well a candidate performs and he/she is usually judged against an arbitrarily set external standard. Hence it is independent of the performance of the group. »*
>
> George, S., Haque, M. S. & Oyebode, F. (2006).

Les barèmes des épreuves qui nous intéressent – et les standards qui y sont fixés – sont en principe le résultat d'un *arbitrage*, d'une décision prise par un ou plusieurs *arbitres* sur la base d'informations et de critères divers. Ce sont alors les modalités de cet arbitrage qui importent pour porter un jugement éventuel sur l'adéquation du barème à la décision à prendre et sur son degré d'arbitraire, au sens courant et connoté négativement de ce terme. Cet arbitraire s'exerce à différentes étapes du processus complexe d'évaluation des apprentissages, mais nous nous limiterons à deux phases, cruciales pour l'établissement des barèmes en nous centrant ultérieurement sur la seconde.

La première phase correspond au choix des questions. Dans le cas des tests de maîtrise tels que nous les entendons, le corpus des items – le 100% de réussite de l'échelle critériée – devrait correspondre à l'ensemble des connaissances et compétences dont on peut légitimement attendre l'acquisition si l'enseignement a été donné et reçu dans de bonnes conditions. L'arbitraire de ce choix est modéré si certaines conditions sont remplies :

- s'il prend comme référentiel un plan d'études, mais celui-ci est rarement conçu à cet effet : généralement, il est insuffisamment précis pour permettre une opérationnalisation sous forme d'items (Marc &Wirtner, 2012) ;
- s'il est fondé sur les propositions de plus d'un évaluateur, encore faut-il une procédure d'arbitrage pour gérer les divergences entre experts ;
- si les épreuves ont fait l'objet d'essais préalables, mais les échantillons d'élèves utilisés sont souvent restreints et pas nécessairement représentatifs de la population visée.

De ce point de vue, un des problèmes qui se posent couramment pour les tests de maîtrise, comme le relevait déjà Cardinet en 1972, tient au fait que les concepteurs de ce type de contrôle se limitent rarement aux objectifs fondamentaux du plan d'études, légitimement exigibles des apprenants. Ils ajoutent souvent des questions plus difficiles, parfois de transfert

ou d'aptitude, à l'intention des « bons élèves » ; en outre, ils craignent parfois qu'un test réduit aux fondamentaux ne donne une image insatisfaisante de leur enseignement.

Cette tendance à confondre tests de maîtrise et tests de niveau fait problème lors de l'autre phase qui nous intéresse plus particulièrement ici : celle lors de laquelle sont fixés des standards sur l'échelle du test, et notamment le seuil de suffisance ou de passage. Une enquête sur la construction des barèmes critériés et des normes adoptées se heurte d'emblée au fait que les instances qui en sont chargées ne livrent pas volontiers (litote !) des détails sur les procédures et critères adoptés *de facto* habituellement ; ceux-ci relèvent parfois de véritables recettes de cuisine, qu'on ne souhaite pas dévoiler dans leurs détails. Signalons des pratiques fréquentes, observées par nous pendant plus de trois décennies en Suisse romande : elle consiste à considérer la distribution finale des résultats et à ajuster un barème jugé acceptable par les intéressés : élèves, parents, collègues, voire autorités scolaires. Par ailleurs, la fixation du seuil de suffisance aux trois quarts ou aux deux tiers des points passe parfois pour une « bonne pratique », défendable face à des tiers.

Une détermination *a priori* des notes n'est éventuellement possible que si l'on dispose pour l'épreuve des résultats d'un essai préalable, valide et fiable, ce qui est surtout le cas pour les épreuves à enjeux élevés. Mais dans ce cas, qui nous intéresse particulièrement, se pose alors le problème des modalités de l'arbitrage. Nous allons donc présenter dans ce qui suit, au fil d'un exemple, une méthode, celle d'Angoff dite modifiée, qui propose une procédure convoquant un panel d'experts évaluateurs et organisant leurs échanges en vue de la fixation d'un seuil de suffisance ou de passage. Auparavant, nous décrivons le test de maîtrise sur lequel nous avons expérimenté la méthode d'Angoff.

1.3 Présentation de l'examen de grammaire, exemple de test de maîtrise[6]

Pour illustrer notre propos, nous avons choisi un examen de grammaire que nous avions eu l'occasion antérieurement d'analyser du point de vue de ses caractéristiques docimologiques, notamment au moyen du modèle de la généralisabilité (Bain, 2010). Cette épreuve a été passée en fin de formation par les étudiants se préparant au brevet d'enseignement primaire ; elle intervenait de façon importante dans leur certification finale. On doit donc la considérer comme un contrôle à enjeu élevé, ce qui justifie qu'on s'intéresse de plus près à la fixation d'un *seuil de suffisance*, en l'occurrence un *seuil de passage*, valide et fiable.

Les six parties de l'examen couvraient chaque fois l'ensemble des domaines grammaticaux enseignés ou révisés : 1 : *Les sortes de phrases* (phrases de base, transformées simples et complexes, non standards) ; 2. *Les trois fonctions majeures de la phrase* (sujet, prédicat ou groupe verbal, complément de phrase) ; 3. *Les fonctions dans le groupe verbal* (cpl. de verbe ; cpl. du verbe de type être, attribut). 4. *Les fonctions des groupes prépositionnels* (cpl. de phrase, cpl. du nom, cpl. de l'adjectif, cpl. du verbe/de type être, attribut, modificateur du verbe) ; 5. *Les fonctions des groupes nominaux* (sujet, cpl. de phrase, cpl. du nom, cpl. du verbe/de type être, attribut) ; 6. *Les fonctions des groupes adverbiaux* (modificateurs du verbe, de l'adjectif et de l'adverbe ; cpl. du verbe, de phrase, du verbe de type être).

Les étudiants devaient généralement repérer dans une phrase une catégorie ou une fonction grammaticales, par exemple en soulignant le groupe de mots correspondant, et l'identifier en

[6] Nous sommes particulièrement reconnaissant au professeur Jean-Paul Bronckart d'avoir mis cet examen à notre disposition ainsi que toutes informations à son sujet, nécessaires pour faciliter le déroulement de la procédure d'Angoff.

donnant son nom. Dans le mode de cotation que nous avons adoptée pour cette recherche, le test comportait 56 questions corrigées justes ou fausses (1 ou 0).

Notre étude antérieure de l'examen (Bain, 2010) attestait à la fois sa validité et sa fiabilité au sens actuel que la docimologie donne à ces termes : « Validity is a unitary concept. It is the degree to which all of the accumulated evidence supports the intended interpretation of test scores for the intended purposes. » (AERA, APA, & NCME, 1999, p. 11)[7]. Cette définition recouvre de fait, comme autant de conditions corrélées, les principales validités traditionnelles : de contenu, de construit, écologique, prédictive ou de conséquences, auxquelles s'ajoute une condition de fiabilité du dispositif d'évaluation.

En ce qui concerne la validité de cet examen, on notera d'abord que les analyses grammaticales à réaliser dans l'épreuve sont très proches de celles que l'étudiant – futur enseignant – aura à faire en situation de classe dans ses cours de français. De plus, le choix des domaines à traiter ainsi que le type de questions grammaticales se réfèrent étroitement aux objectifs du Plan d'études romand (PER, CIIP, 2010-2016) que suivront les élèves du futur enseignant et qu'il devra lui-même respecter. Enfin, les candidats ont eu l'occasion de se familiariser avec ce genre de questionnement pendant le cours et de consulter des exercices analogues dans le document soutenant la formation (Bronckart, 2004).

Pour contrôler la fiabilité de cet examen, partie intégrante de sa validité, nous avions eu recours au modèle de la généralisabilité et au logiciel *EduG* (Cardinet, Johnson & Pini, 2010). L'enseignant ayant choisi le seuil de suffisance de 75%, nous avions calculé un coefficient critérié *Phi(lambda)*, qui s'est révélé très élevé : 0.97, avec une erreur type absolue de 3% (intervalle de confiance : 6%). Contrôle préalable nécessaire si l'on suit la recommandation évidente de Çetin & Gelbal (2013, p. 2170): « Tests which are low reliable should not be used in standard setting process. »

Enfin, la distribution des scores obtenus par les étudiants à cet examen correspondait à ce que l'on attend d'un test de maîtrise, soit une courbe en J attestant que la quasi-totalité d'entre eux a réussi les trois-quarts des points.

2. La méthode d'Angoff et son application à un examen de grammaire

Dans ce chapitre, nous exposerons en détail la méthodologie appliquée à notre recherche lors des diverses étapes de notre travail, c'est-à-dire tout au long des différentes phases (*rounds*) de la procédure d'Angoff : de sa préparation à la fixation du seuil de suffisance. Nous commencerons par justifier notre choix de cette méthode.

2.1 Fixation d'un seuil de suffisance : choix de la méthode d'Angoff

« There can be no single method for determining cut scores for all tests or for all purposes, nor can there be any single set of procedures for establishing their defensibility » (AERA/APA/NCME, 1999, p. 53).

L'ouvrage de Cizek & Bunch (2007) *Standard setting* présente et commente une douzaine de méthodes envisageables pour fixer des standards ; le lecteur s'y reportera pour avoir une vision détaillée des possibilités dans ce domaine de la docimologie. Nous avons choisi la

[7] Comme le rappellent Cizek & Bunch (2007, p. 17), en citant Messick (1989) puis Cronbach & Mehl (1955), « strictement parlant, on ne peut pas dire que des tests ou des scores de tests sont valides ou non valides. […] On ne valide pas un test mais seulement un principe guidant des inférences » (notre traduction).

méthode d'Angoff – en l'occurrence la *méthode d'Angoff modifiée* (*Modified Angoff Method*) illustrée plus loin – pour des raisons souvent évoquées dans la littérature (o.c., chapitre 2). Sous différentes variantes, elle est utilisée depuis plus de quarante ans dans des contextes divers, en particulier dans ceux qui sont les plus exigeants en ce qui concerne la pertinence et la fiabilité des seuils de passage comme les formations en médecine. Elle passe pour relativement facile à appliquer, même par des novices disposant d'un minimum d'entraînement (Wheaton & Parry, 2012) et offrirait le meilleur équilibre entre adéquation technique et praticabilité (Berk, 1986, p. 147, cité par Cizek & Bunch, 2007, p. 82). Elle satisfait aux exigences légales fixées par certains pays (notamment les USA) pour les épreuves à enjeux élevés, en particulier lorsqu'il s'agit d'obtenir un *permis d'exercice* dans une profession sensible sur le plan social (métiers de la santé, de l'éducation ou de la police, par exemple). Notons de plus que, appliquée généralement à des questions à choix multiples, la méthode peut être également utilisée dans une autre variante, dite *Méthode d'Angoff étendue*, où les experts estiment le nombre de points obtenus par des élèves *borderlines* à des questions à réponse construite (o. c, p. 87).

2.2 Le choix des experts évaluateurs

Le choix des évaluateurs a évidemment un impact potentiellement important sur le résultat de la procédure : « In standard setting techniques involving panels of judges, the attributes of judges may affect the cut-scores » (Shulruf, Wilkinson, Weller, Jones & Poole, 2016, p. 1; cf. aussi Busch & Jaeger, 1990). Il est donc nécessaire de préciser les conditions de leur recrutement. Compte tenu du type de contrôle (examen universitaire de grammaire), nous avons sollicité des collègues enseignants universitaires, experts à la fois en grammaire et en didactique de cette branche au niveau de formation considéré. Nous nous sommes adressé pour cela à la commission GRAFE'MAIRE, membre du Groupe d'analyse du français enseigné (GRAFE) de la Faculté de psychologie et des sciences de l'éducation (FAPSE) de l'Université de Genève. Nous avons pu ainsi recruter dix experts[8], nombre souvent préconisé pour une application fiable de la procédure (Wheaton & Parry, 2012, p. 3).

Le hasard de ce recrutement nous a permis de scinder ce groupe en deux sous-groupes en fonction de leur origine institutionnelle pour vérifier si la familiarité des évaluateurs avec le type d'étudiants évalués jouait un rôle dans leurs estimations. Le premier groupe a été constitué de collègues enseignant à l'Université de Genève (où l'examen a été passé), le second d'experts provenant d'autres institutions universitaires : de Fribourg, de Vaud ou de Grenoble.

Nous avons écarté de cette sélection l'auteur de l'examen : il connaissait les résultats effectivement obtenus par ses étudiants à cette épreuve en ce qui concerne tant le seuil fixé à la note suffisante que le taux de réussite aux diverses questions ; il aurait donc pu influencer les estimations de ses collègues.

[8] Nous leur exprimons toute notre reconnaissance ; sans leur amicale disponibilité, cette recherche n'aurait pas pu avoir lieu. Il s'agit de Ecaterina Bulea, Sandra Canelas Trevisi, Christopher Länzlinger, Anouk Darme, Jean-François de Pietro, Serge Erard, Roxane Gagnon, Martine Panchout-Dubois, Véronique Marmy Cusin, Vincent Capt.

2.3 La phase préparatoire

Cette étape est cruciale pour que les participants saisissent l'objet ainsi que les modalités de la méthode et de la procédure ; elle l'est d'autant plus si, comme dans le cas de cette recherche, les experts sont novices dans l'application de ce type d'évaluation. Comme recommandé (Cizek & Bunch, 2007, chap. 2), nous avons d'abord fourni aux participants le syllabus du cours (Bronckart, 2004) et un exemplaire d'un examen antérieur équivalent[9], comportant les six mêmes types d'exercices, avec prière de le passer eux-mêmes à la maison. Ils ont reçu par ailleurs par écrit le plan de l'ensemble de l'opération pour qu'ils puissent en situer chaque étape.

2.4 Première phase de la procédure d'Angoff

2.4.1 Présentation et discussion de la consigne

Les tableaux 1 et 2 reproduisent deux formulations de la consigne analogues à celles que proposées habituellement (Angoff, 1971, p. 515 ; Cizek & Bunch, 2007, pp. 82-83).

Tableau 1. Formulation de la consigne

C1. Pour chaque question de l'examen, estimer la probabilité, en %, par pas de 5%, qu'un étudiant minimalement compétent (borderline) devrait donner une réponse juste et complète. ~~Référez-vous pour cela à votre expérience ces dernières années~~.

Tableau 2. Seconde formulation de la consigne

C2. Imaginez, ~~en fonction de vos expériences antérieures~~, un groupe de 100 **étudiants *borderlines*** (minimalement compétents) et estimez pour chaque item de l'épreuve la proportion d'entre eux (en %) qui devraient donner une réponse juste et complète.

A l'expérience, la présentation de ces deux consignes a posé les problèmes suivants.

- La difficulté de se représenter un *étudiant minimalement compétent* (*borderline*). La solution finalement adoptée a été de renvoyer les évaluateurs au contexte institutionnel et notamment à la fonction sélective de l'examen : écarter les candidats à l'enseignement ne disposant pas des qualifications nécessaires en grammaire. Sous cette forme générale et globale, la figure du *borderline* est encore difficile à saisir. En revanche, la consigne est plus facile à comprendre quand on l'applique à une question en particulier. Par exemple, un formateur peut difficilement accepter qu'un futur instituteur ne soit pas capable d'identifier un complément de verbe ou le confonde avec un complément de phrase, fonctions qui figurent déjà dans le plan d'études aux degrés 5 et 6 de la scolarité obligatoire ; 100% des enseignants devraient donc réussir la question.
- La difficulté d'exprimer les estimations sous forme de *probabilité,* même si les experts étaient familiarisés avec cette notion du fait de leurs activités de recherche. La consigne alternative C2 (tableau 2) n'aide guère : il est tout aussi difficile de se représenter de façon réaliste un groupe de 100 étudiants minimalement compétents, d'autant plus si la cohorte à laquelle on a généralement affaire ne comporte qu'une centaine d'étudiants.

[9] Équivalence attestée par notre étude de 2010.

La discussion sur ce mode d'évaluation peut d'ailleurs déboucher dans un premier temps sur la question, docimologiquement intéressante et que nous n'avons pu esquiver : « Pourquoi, s'agissant d'un test de maîtrise (cf. la définition supra), ne pas exiger des étudiants jugés compétents 100% de réussite à chaque question et, *a fortiori*, exiger le score maximum ? » Dans un premier temps, on peut alléguer un aléa dans le choix des questions et dans leur rédaction ainsi que les erreurs aléatoires inhérentes à toute passation de test ; on admettra aussi que certains items dépassent dans leurs exigences le niveau attendu par le plan d'étude ou représentent des cas particuliers du fait du choix de leur formulation. On doit généralement aller plus loin dans l'argumentation : concéder que l'échec à quelques items ne devrait pas porter à conséquence ; qu'il pourrait être compensé par la réussite à d'autres items, conformément au *modèle compensatoire* adopté habituellement pour la notation de ce genre de test (*scoring model* : Cizek & Bunch, 2007, p. 20) ; que le futur instituteur, au cours de sa carrière, aura très probablement l'occasion d'approfondir ses connaissances sur certains points contrôlés par l'examen et qu'il ne maîtriserait pas encore.

Constatant après la première phase que cette estimation de probabilité posait toujours problème, nous avons proposé pour la seconde phase quelques points de repères sur l'échelle en % ; nous y reviendrons donc plus loin.

Le plus souvent, dans les consignes données pour la méthode d'Angoff modifiée, cette probabilité de réussite à un item s'exprime par pas de 10% (100%, 90%, 80%…). Compte tenu du type d'épreuve, assimilable à un test de maîtrise, nous avons proposé des pas de 5% ; ce degré de précision permettait de nuancer les estimations, supposées se situant majoritairement entre 80% et 100%.

Noter encore que, pour cette application de la méthode d'Angoff, nous avons modifié la consigne de correction par rapport à celle appliquée primitivement à l'examen en exigeant pour chaque question une réponse juste et complète ; nous avons renoncé à considérer comme items des éléments de réponses *de facto* non indépendants[10], et à prendre en compte des pénalisations ou des bons points liés à des groupes de réponses. Pour cette raison, nous désignerons par la suite nos unités d'observation/d'estimation par le terme de *questions* plutôt que d'items.

Enfin, nous avons renoncé à renvoyer les experts à leurs expériences antérieures, potentiellement différentes en fonction des publics d'apprenants enseignés (cf. partie de la consigne doublement barrée), mais surtout pour centrer les estimations sur les exigences à la fois du plan d'études et de l'exercice futur de la profession.

2.4.2 Récolte des données, traitement et présentation des résultats

Les participants reçoivent un fichier Word avec un tableau de 56 cases correspondant aux estimations demandées pour l'ensemble des items (cf. extrait au tableau 3). Ils remplissent individuellement, à la maison, toutes les cases du tableau. Ces données sont ensuite retranscrites par copier-coller sur un tableau Excel.

Après traitement des données (calcul des moyennes et des écarts types par question et par évaluateur), les experts reçoivent un tableau de l'ensemble des résultats, anonymisé (cf. extrait au tableau 4). Seules les lignes du tableau correspondant au destinataire du document

[10] Par exemple, si un complément n'était pas repéré/souligné dans la phrase, il ne pouvait *a fortiori* pas être correctement identifié : les deux points attribués à cette question n'étaient pas indépendants. Or, cette indépendance est supposée par le calcul d'un total et le traitement statistique de ces données.

(ici l'expert no 3) sont identifiées, par surlignement ; la dernière est prévue pour recevoir les estimations de la seconde phase du même évaluateur.

Tableau 3. Examen de grammaire 2008. Relevé des évaluations (en %) par exercice et par question (extrait)

1.01	1.02	1.03	1.04	1.05	1.06	1.07	1.08	1.09	1.10

1.11	1.12	1.13	1.14	1.15	1.16	1.17	1.18	1.19	1.20

[…]

Tableau 4. Résultats de la première phase (extrait)

Exp.	Q1.10	Q1.11	Q1.12	Q1.13	Q1.14	...	Moy. exp.	σ
E1	70	60	40	60	70	...	**66.4**	**12.8**
E2	100	50	85	100	100	…	**82.8**	**19.3**
E3	90	60	90	70	100	...	**86.6**	**12.8**
E4	60	60	50	60	80	...	**64.3**	**12.3**
E5	80	70	70	90	90	...	**80.0**	**9.3**
E6	80	60	50	70	80	...	**69.8**	**11.7**
E7	80	60	70	75	100	...	**64.3**	**14.7**
E8	100	90	40	60	100	...	**80.7**	**20.3**
E9	100	90	100	90	80	...	**82.7**	**16.0**
E10	90	80	100	70	90	...	**82.1**	**18.9**
m qu.	**85.0**	**68.0**	**69.5**	**74.5**	**89.0**	**...**	**m gén.76.0**	**8.7**
σ qu.	**13.5**	**14.0**	**23.6**	**14.2**	**11.0**	**...**		
E3 P2						**...**		

Lecture et commentaire du tableau 4

- Les lignes E1 à E10 fournissent le détail des estimations de chaque expert pour les 56 questions.
- La première ligne qui suit (en grisé : « m qu. ») donne la moyenne des dix experts pour chaque question. Elle traduit le taux de réussite moyen attendu pour les étudiants juste suffisants selon la notion évaluée. Cette exigence est nettement plus élevée (89%) pour la question 1.14, par exemple (« Les étudiants ont festoyé pendant plusieurs semaines. », à identifier comme une phrase de base) que pour la phrase transformée complexe par subordination 1.11 (« La marchande, qui rêvait, n'avait pas entendu le client. » ; 68%).

- La ligne suivante (« σ qu. »), reproduisant le vecteur des écarts types pour les différentes questions, permet de repérer celles pour lesquelles les estimations divergent le plus entre experts, soit celles où l'écart type est le plus élevé. C'est le cas de la question 1.12 (« Il l'a regardée puis il lui a souri. » ; phrase transformée complexe par coordination), pour laquelle le degré d'exigence est estimé très différemment selon les experts (de 40% à 100%). Lors d'une discussion ultérieure sur cette question, certains évaluateurs craignaient qu'une partie des *borderlines* n'identifient pas le *puis* comme un coordonnant analogue à *et*.
- L'avant-dernière colonne reproduit la moyenne de chaque *expert* pour les 56 questions. Chaque évaluateur peut ainsi situer lui-même son degré de sévérité par rapport à ses collègues, mais cette donnée ne fait pas l'objet de discussion ou de commentaire dans le groupe : il n'est évidemment pas question de stigmatiser qui que ce soit comme plus sévère ou plus indulgent que les autres.
- La moyenne générale (des questions, identique à celle des évaluateurs ; m gén. : 76%) constitue le *seuil de suffisance* recherché. Elle correspond à environ les trois quarts des points. Elle est pratiquement identique au seuil choisi par l'auteur de l'examen (75%), mais cette information, à ce stade, n'est pas fournie aux participants pour ne pas influencer la suite des opérations.
- L'écart type des 56 estimations de chaque expert, qui figure dans la dernière colonne, n'a pas été commenté. Il est surtout intéressant pour le chercheur dans la mesure où il permet de repérer des experts dont les estimations – donc la sévérité – varient plus ou moins fortement selon les questions.[11] Le cas le plus problématique serait celui d'un écart type particulièrement bas ; il signalerait le cas d'un évaluateur tendant à attribuer systématiquement la même estimation à des questions dont on saurait par ailleurs qu'elles sont de difficultés différentes.
- A ce moment de la procédure, des statistiques purement descriptives peuvent être suffisantes. L'*écart type* de 8.7% correspondant à ce seuil suffisait à justifier une seconde étape visant à réduire si possible les divergences entre évaluateurs.

2.4.3 Feed-back normatif et discussion des résultats par les experts

La transmission de l'ensemble des estimations aux experts sous la forme du tableau 4 a pour but de permettre la comparaison entre les différentes réponses ; cette étape de la procédure constitue une forme de *feed-back normatif* (Cizek & Bunch, 2007, p. 53). Le terme de normatif est bien choisi : il signale selon nous un risque de glissement de références critériées (le taux de réussite des étudiants tout juste suffisants) à une perspective normative des évaluations. Celle-ci viserait avant tout à situer les exigences des évaluateurs par rapport à celles de leurs collègues. L'objectif – ou le résultat – majeur de l'opération serait alors de susciter une espèce de régression artificielle des estimations à la moyenne générale pour diminuer (à tout prix ?) les divergences. Cizek & Bunch, 2007, précisent (p. 53) : « The purpose of providing this feedback is not to suggest that they align their individual judgments with a group mean or alter their judgments based solely on relative stringency or leniency compared to other participants ». Pour éviter ce qui pour nous aurait été un biais de la recherche, mélangeant les approches critériée et normative, nous nous sommes donc efforcé de centrer la discussion qui suivait sur la consigne précisant l'objectif et les modalités de la méthode. Nous avons notamment rappelé que le critère majeur auquel se référer était les exigences qu'on peut avoir

[11] Compte tenu d'un certain effet plafond, limitant la variance des estimations dans le cas de certains experts dont la sévérité (cf. moyenne) est élevée.

quant aux connaissances des futurs instituteurs en début de carrière. Enfin, pour diminuer les divergences d'estimation, nous avons posé quelques balises le long de l'échelle d'estimation (tableau 5, nouvelle consigne), qui apportent des précisions – très relatives, à vrai dire – pour le travail d'évaluation des questions.

Tableau 5. Reprise de la consigne en vue de la 2e phase (extrait)

[…] [Dans la consigne[12]] « **devrait »** correspond à une double modalité (verbe modal et conditionnel) impliquant à la fois un *devoir* face aux futures responsabilités de l'enseignant à l'égard de ses élèves et l'expression d'une *probabilité*. Certaines connaissances grammaticales peuvent être considérées comme élémentaires, incontournables, leur non-maîtrise comme inadmissible par rapport à l'objectif du test. Il s'agit de s'assurer que les futurs enseignants certifiés maîtrisent les notions qu'ils auront à faire apprendre le plus souvent à leurs élèves ; les items correspondant à cette définition devraient être cotés 100% (ces questions devraient être résolues correctement même par des *borderlines*). Pour d'autres connaissances, donc d'autres questions ne correspondant pas à cette exigence forte (par exemple portant sur des cas moins fréquents ou particuliers, qu'il s'agisse de classes grammaticales ou de fonctions), on considérera que le groupe des étudiants *borderlines* est très probablement relativement hétérogène et qu'une partie seulement d'entre eux pourraient réussir l'item en question : pour la plupart (1 sur 20 ou 1 sur 10 → 95% ou 90%), en nette majorité (probabilité entre 65% et 85%), ou dans un cas sur deux environ (entre 40% et 60%), ou enfin nettement plus rarement (moins de 40%), voire jamais (0%).

Les échanges sur les items sélectionnés (cf. supra) ont été aussi l'occasion pour les experts de confronter les conceptions qu'ils avaient de la formation en grammaire des candidats à l'enseignement. Toutefois, le temps disponible pour cette phase a été trop court pour un approfondissement des convergences ou divergences sur plus de quelques items.

2.5 Seconde phase de la procédure

2.5.1 Seconde estimation des questions

Le tableau 4 (supra) des premiers résultats comportait une dernière ligne sur laquelle les 10 experts ont inscrit leurs nouvelles estimations. Ils avaient ainsi sous les yeux, à titre de comparaison ou de référence, leurs premières estimations et celles de leurs collègues. Ils étaient parfaitement libres de conserver ou modifier leurs estimations. Dans la perspective de la présente recherche, il ne s'agissait pas pour nous de les « forcer » à la convergence, mais d'observer si la phase précédente d'information et de discussion avait un effet – et de quelle ampleur – sur leurs estimations et si on observait des différences individuelles plus ou moins importantes.

[12] Cf. supra la consigne C1 (tableau 1) : « Pour chaque question de l'examen, estimer la probabilité, en %, par pas de 5%, qu'un étudiant minimalement compétent (borderline) **devrait** donner une réponse juste et complète. »

Tableau 6. Extrait des résultats de la seconde étape

Expert	P	Q1	Q2	Q3	Q4	Q5	Q6	…	m E	éc.t. E
E1	1	70	70	40	60	80	50	…	66.4	12.8
E1	2	80	75	100	90	85	90	…	83.5	11.1
E2	1	80	95	100	100	95	65	…	82.8	19.3
E2	2	80	95	100	100	95	65	…	82.8	19.3
E3	1	70	60	80	50	60	40	…	64.3	12.3
E3	2	80	70	80	60	80	60	…	75.7	10.6
…	…	…	…	…	…	…	…	…	…	…
Moy.	1	74.0	79.5	87.0	81.0	84.0	67.5	…	76.2	m. gén. P1
Moy.	2	83.5	87.0	97.0	90.0	91.5	82.5	…	84.3	m. gén. P2
éc.-t.	1	12.6	16.4	18.9	19.7	14.1	19.9	…	9.0	éc. type P1
éc.-t.	2	12.9	13.2	6.7	16.3	8.5	15.5	…	7.7	éc. type P2

2.5.2 Analyse descriptive et commentaire des résultats

Le traitement des données de cette phase aboutit à un nouveau tableau (tableau 6), analogue à celui établi pour la première phase, mais comportant deux lignes par expert qui permettent de confronter les données pour les deux phases (P1 et P2). Ce tableau (transmis et commenté aux experts : second feedback) peut donner lieu à de multiples analyses et commentaires ; nous nous limiterons ici aux principaux constats intéressants pour le fonctionnement de la méthode modifiée d'Angoff.

Dans un premier temps de la discussion, il peut être instructif de revenir sur les questions ayant suscité le plus de divergences pour vérifier l'impact potentiel de la discussion. A titre d'exemple, on constate que l'écart type de la question 1.12 (tableau 6) a diminué, passant de 24% à 15%. Mais, simultanément, le degré d'exigence pour cette même question a monté de près de 15% (70% → 84%) : si l'on accepte de considérer « puis » comme une coordonnant équivalent à « et », on a affaire à une phrase transformée complexe par coordination (« Il l'a regardée puis il lui a souri. »), que devrait en principe maîtriser une grande majorité des futurs instituteurs.

On observe ensuite et surtout une augmentation moyenne appréciable des exigences lors de la seconde phase : le seuil de suffisance passe ainsi de 76.2% à 84.3% (tableau 6). Relevons que ce second seuil de 84% est nettement plus élevé que ceux fixés traditionnellement et notamment que celui qui avait été décidé effectivement pour l'examen en question (75%). On constate en outre que pour chaque phase les moyennes des deux catégories d'experts, appartenant ou non à l'université de Genève, sont pratiquement équivalents (P1 : 76% vs 77% ; P2 : 84% vs 85%), compte tenu de la variabilité des estimations à l'intérieur des deux groupes (σ >13% ; cf. aussi graphique 7). La familiarité avec le type d'étudiants ayant passé l'examen ne paraît donc pas influencer les évaluations.

Par ailleurs, on constate des différences de stratégie individuelle en ce qui concerne le réajustement des estimations entre les deux phases. L'analyse de l'évolution des seuils des dix experts (graphique 7) fait ressortir deux groupes d'effectifs équivalents : les experts qui ne modifient pratiquement pas leur niveau moyen d'estimations entre les deux étapes (la différence entre P1 et P2 est inférieure ou égale 2%) et ceux qui augmentent leurs exigences.

Le respect déontologique de l'anonymat des réponses, mais surtout le manque de temps, ont empêché une discussion qui aurait clarifié les raisons de cette différence de comportements.

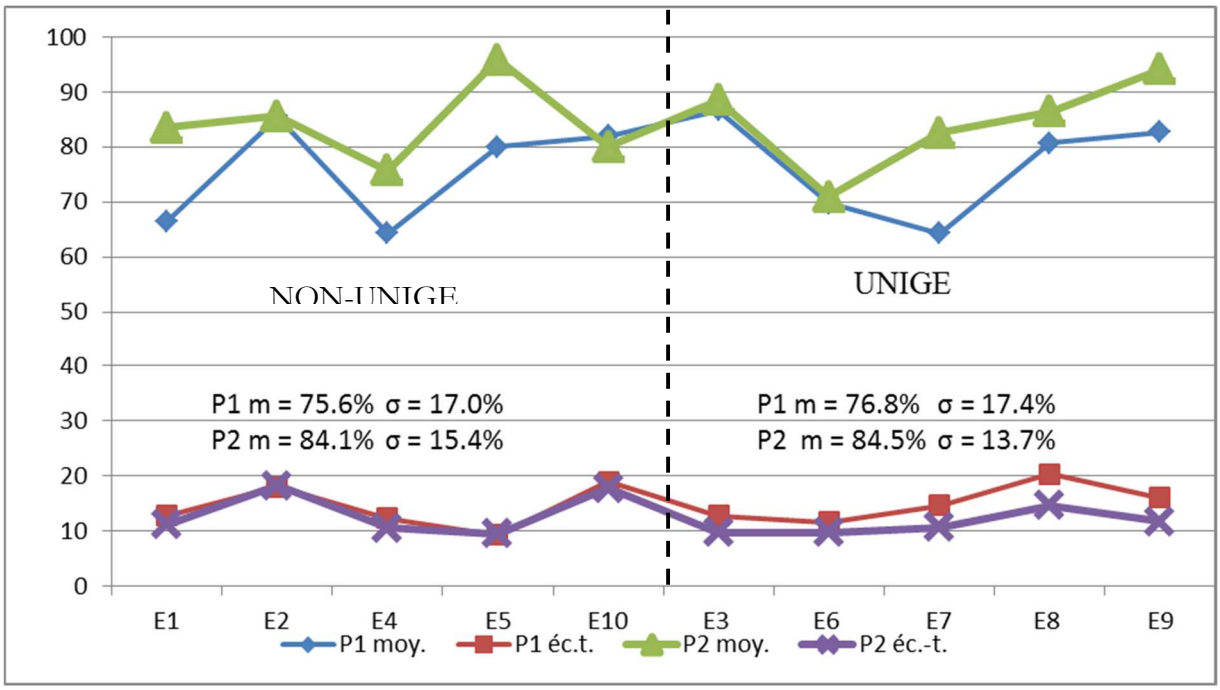

Figure 7. Moyennes générales (seuils de suffisance) selon les experts (En) et écarts types de leurs estimations pour les premières et secondes phases (P1 et P2) de la procédure d'Angoff

Les exigences ont augmenté peu ou prou en phase 2 pour la quasi-totalité des questions (55 sur 56), du fait des estimations d'une moitié des experts, rappelons-le. On peut expliquer cette évolution par le fait que certains évaluateurs ont été sensibles à notre insistance, lors de la seconde version de la consigne (tableau 5 supra), sur la responsabilité de l'institution de formation à l'égard des futurs élèves des candidats à l'enseignement. Nous ne pouvons pas présenter ici le détail de la réussite pour les 56 questions. Il serait cependant intéressant d'un point de vue didactique et docimologique ; en effet, dans les estimations des experts, le degré de difficulté des questions varie assez fortement : de 52% à 89% pour la phase 1 et de 50% à 98% pour la phase 2. Dans la perspective d'un test de maîtrise, on s'attend en principe à ce que la marge de variation inter-questions soit plus étroite, centrée sur un seuil de suffisance anticipé relativement élevé. Les questions jugées difficiles pour les élèves *borderlines* auraient mérité un examen plus approfondi des raisons des échecs anticipés par les experts. Toutefois, pratiquement, le temps a manqué pour ce type d'analyse.

Enfin, entre les deux étapes, les divergences entre évaluateurs ont apparemment un peu diminué : l'écart type des estimations passe de 9% à 8% (tableau 6). Cette différence est faible, voire pratiquement nulle : il est possible que cette diminution soit due en partie à un effet plafond : pour plusieurs questions, la moyenne des estimations avoisine 100%. Les divergences restent donc relativement importantes du fait des deux types d'évolutions signalés à l'instant : maintien ou augmentation du niveau d'exigences.

2.5.3 Renoncement à une troisième étape, aux informations dites de réalité et d'impact

Compte tenu de ces divergences encore relativement importantes, n'aurait-il pas fallu passer à une troisième phase, souvent recommandée dans la littérature (Cizek & Bunch, 2007, pp. 54-56) ? Lors de cette étape, on injecte dans la procédure des informations sur les

résultats effectifs à l'examen ou à des contrôles équivalents antérieurs. L'objectif majeur est généralement d'améliorer la convergence des estimations en fournissant aux experts des références communes.

L'*information* dite *de réalité* consiste à donner aux participants des renseignements notamment sur le taux de réussite moyen (*p value*) de l'ensemble des étudiants sur un sous-ensemble d'items en se référant aux résultats d'un test précédent comparable ou aux résultats effectifs du test en cours d'évaluation. L'*information d'impact* consiste à indiquer quel serait le taux d'échecs dans le groupe testé ou dans certains sous-groupes (par exemple dans certaines filières de formation), si l'on appliquait le seuil estimé par les experts. Nous avons finalement renoncé à une telle étape faute de temps, mais surtout parce que, dans les deux types de feed-back, on tend à mélanger les références normatives et critériées, changeant ainsi le point de vue adopté au départ de notre recherche. Notre référence était, rappelons-le, la *réussite attendue ou exigible* des candidats instituteurs compte tenu de leur activité professionnelle ultérieure. Dans cette perspective, le décalage entre la réussite effective des apprenants et celle estimée par les experts ne constitue pas pour nous une remise en cause du processus de *standard setting* en soi ou de la qualification des évaluateurs mais une information intéressante à discuter sur le plan pédagogique.

2.5.4 Prise en compte de l'erreur de mesure sur le seuil : apport de l'analyse de généralisabilité[13]

Comme nous renonçons à cette troisième étape, selon la procédure habituelle d'Angoff, c'est le résultat de la seconde phase que nous prendrons en considération pour fixer le seuil final. Mais celui-ci devrait toujours prendre en compte l'erreur de mesure due à l'échantillonnage des questions et des experts, selon la recommandation des *Standards for Educational and Psychological Testing* : « Where cut scores are specified for selection or classification, the standard errors of measurement should be reported in the vicinity of each cut score. » (AERA/APA/NCME, 1999, Standard Number 2.14). Nous avons donc calculé l'*erreur type* affectant le seuil de suffisance P2 en recourant à une *analyse de généralisabilité* (plan de mesure sans facette de différenciation, proposé par Cardinet, 2014, p. 5). Cette *erreur type absolue* (il s'agit de situer le seuil sur l'échelle des scores de l'épreuve) est de 2.66%. Pour calculer l'intervalle de confiance – la marge d'erreur ou la zone d'incertitude – autour du seuil de suffisance, on multiplie cette valeur par 1.96, soit 2.66 x 1.96 = 5.21 (pour p = .05). Le « score vrai » correspondant au seuil de suffisance se situe donc entre 79% et 89% (84% ± 5%). Par ailleurs, on observe sans surprise que la principale source d'erreur absolue est due aux divergences d'évaluation entre Experts (E : 81% de la variance d'erreur absolue totale).

Comment prendre en compte la marge d'erreur pour déterminer le seuil de suffisance final ? On observe différentes solutions correspondant aux stratégies institutionnelles ; nous y revenons dans le chapitre de discussion qui suit.

[13] Faute de place, nous ne pouvons présenter ici dans le détail le recours à la *théorie de la généralisabilité* pour calculer l'erreur de mesure affectant le seuil de suffisance et pour estimer les différentes sources de cette erreur. Pour des informations détaillées à ce sujet, on se référera à la version longue du présent texte (Bain, 2018, à paraître sur le site du groupe Edumétrie : https://www.irdp.ch/institut/edumetrie-1635.html) ou à Cardinet, Johnson & Pini, 2010.

3. discussion : limites et problèmes du modèle

Nous reprenons dans ce chapitre certains objets traités ci-dessus à propos de la méthode d'Angoff pour évoquer – ou revenir sur – les limites et les problèmes rencontrés ou potentiels du recours à cette procédure de *standard setting*.

3.1 De la généralisabilité de nos conclusions

Les conclusions de notre travail sont limitées dans leur généralisabilité par le fait que nous nous situons dans le cadre d'une *étude de cas*, appartenant au paradigme de *recherche de faisabilité* (Astolfi, 1993). Nous aurions voulu nous appuyer sur d'autres études du même type pour étayer ou relativiser nos observations. Malheureusement, dans le domaine de la didactique, et plus particulièrement de la didactique du français, on ne trouve guère de travaux s'appuyant sur la méthode d'Angoff, et encore moins portant sur des épreuves de maîtrise au sens où nous les avons définies en introduction.

En consultant la littérature disponible (surtout anglophone ; cf. références bibliographiques), nous constatons en effet que, dans la plupart des cas, la méthode est appliquée à des évaluations de performances qui sont *de facto* des *épreuves de niveau*, sur lesquelles on définit *plusieurs niveaux de maîtrise*. On se contente plus rarement[14] de tester des connaissances ou des compétences fondamentales, exigibles à un certain stade de la formation. Même quand le contrôle porte sur un socle de compétences (par exemple, ceux de la DEPP[15] en France) ou des attentes fondamentales (épreuves cantonales ou romandes en Suisse), on cherche souvent à tester « jusqu'où vont les compétences des élèves », voire à les évaluer en fonction de *normes d'excellence* (Perrenoud, 1989), à des fins (pas toujours affichées) de classement des élèves ou de leurs performances. Ce qui a un impact notamment sur les conditions de passation et de fiabilité des épreuves : la forte variance due aux questions a un impact négatif sur l'erreur type absolue si l'épreuve est critériée (cf. Cardinet, 2014, à propos de l'article de Verhoeven et al., 1999, et Bain, 2018).

Par ailleurs, comme le remarquent Cisek & Bunch (2007, p. 81), il y *de facto* autant de variantes que d'utilisations de la méthode d'Angoff modifiée : « The method as described by Angoff is rarely used exactly as it was proposed. Rather, slight reconfigurations of the basic approach – each variation referred to as "modified Angoff method" – are now considerably more common, although precisely what constitutes a "modified" Angoff method is somewhat unclear ». Nous doutons donc qu'on puisse tenir un discours généralisateur sur la méthode d'Angoff. Trop de facteurs sont susceptibles de modifier son application et ses résultats : le type de connaissances, de performances ou de compétences évaluées ; le contexte et les enjeux des contrôles ; la compétence et l'expérience des experts ; l'organisation des échanges interphases ; la prise en compte ou non des résultats effectifs de l'épreuve… Pour que l'on puisse juger de l'impact de ces facteurs sur les résultats, il est malheureusement exceptionnel que des chercheurs proposent une description détaillée du contexte institutionnel et de ses contraintes ; de la façon dont ils ont appliqué la procédure d'Angoff ; des incidents ou des difficultés de parcours ; des raisons pour lesquelles certains experts ou items ont été écartés… Méta-analyse (Hurtz & Auerbach, 2003) ou simulation (Shulruf & al., 2016) s'achoppent aux mêmes types de problèmes.

[14] A l'exception notamment des examens finaux de médecine.
[15] Cf. DEPP 2014 et 2015.

La solution que nous avons alors choisie dans cette discussion est de reprendre avantages, problèmes et limites de la méthode d'Angoff telles que nous les avons expérimentés, dans le contexte décrit en introduction, tout en les confrontant à des constats semblables ou différents repérés dans les travaux consultés. Nous considérons ainsi que nos observations peuvent être pertinentes pour d'autres applications dans la mesure où les contextes institutionnels ou pratiques ne seraient pas trop différents et où nos constats recoupent des analyses faites dans d'autres recherches pas trop éloignées des nôtres.

3.2 Intérêt et avantages docimologiques de la méthode d'Angoff

3.2.1 Facilité d'application (relative)

Un premier avantage de la méthode, souvent avancé, est sa facilité d'application : « The Angoff method is easy to implement and can be perfected by novice users with only minimal training. » (Cisek & Bunch, 2007). Nous avons pu le vérifier lors de la présente recherche, mais aussi lors de deux autres opérations menées dans le cadre du groupe Edumétrie, portant sur un examen de physique pour l'admission à l'université (Bain & Weiss, 2016), ou sur des épreuves cantonales de mathématique passées en fin de scolarité obligatoire (Frey, 2016 et 2017 ; Bain, 2016). A cette dernière occasion, nous avons pu expérimenter différents formats de question (juste/faux, QCM, % de réussite de la question) et tester une autre modalité de la procédure proposée par Angoff : la *méthode Oui/Non* (*Yes/No*). Nous avons donc pu vérifier la faisabilité de la procédure, tout en étant amené à relativiser sa « facilité d'application » (cf. infra).

3.2.2 Mise en évidence et contrôle de l'arbitraire des évaluations

Mais pour nous, l'avantage majeur de la méthode est de mettre en évidence la part d'arbitraire des barèmes appliqués à bien des épreuves à enjeux élevés se présentant comme des tests de maîtrise. La méthode est particulièrement conséquente avec des contrôles critériés, pour lesquels la simple distribution des scores ne constitue pas une référence suffisante quant à l'atteinte des objectifs fixés par le plan d'études. Contraignant chaque évaluateur-arbitre à fournir séparément ses estimations pour chaque question, la procédure d'Angoff évite de pseudo-consensus à la suite de brefs échanges, influencés parfois par l'opinion prépondérante de certains participants, notamment celle des concepteurs de l'épreuve. Notre expérience montre que, dans ces conditions, peuvent se manifester des divergences non négligeables, même après la discussion sur les résultats d'une première étape évaluation question par question : les seuils de suffisance finaux en phase 2 diffèrent de 20% (76% vs 96%) d'un expert à l'autre.

Les procédures habituelles de fixation des barèmes scotomisent de telles disparités d'estimations. Nos résultats laissent supposer que ces différences sont dues principalement à un niveau global d'exigence propre à chaque évaluateur. Ce niveau de sévérité ou d'indulgence est vraisemblablement influencé par l'expérience de chaque expert, mais dans des conditions et des contextes divers, difficiles à saisir. Nous avons constaté, par exemple, que la connaissance du public estudiantin ne semble pas jouer de rôle déterminant. Tout se passe comme si l'expérience de chacun donnait lieu à l'équivalent d'une *équation personnelle* (de Landsheere, 1979, p. 112).

Comme dans d'autres recherches (Hurtz & Auerbach, 2003), nous avons constaté que cette discussion interphase avait pour effet d'augmenter le niveau du seuil (en l'occurrence de près de 10%). C'est probablement le cas quand les évaluateurs prennent conscience de l'enjeu institutionnel de l'examen. Dans notre recherche, une révision de la consigne allait dans ce sens en insistant sur le fait qu'il s'agissait d'écarter des candidats dont les compétences

seraient insuffisantes pour l'enseignement l'année suivante. Toutefois, cette tendance à la hausse des exigences ne s'observe que pour la moitié de nos experts, ce qui laisse soupçonner dans ce cas également des différences individuelles quant à la remise en cause d'estimations antérieures. Pour des raisons déontologiques, nous n'avons pas cherché à savoir s'il s'agissait d'une certaine inertie évaluative ou le maintien délibéré d'un niveau d'exigence (en l'occurrence, d'indulgence relative).

3.2.3 Intérêt didactique d'une analyse question par question

Sur le plan didactique, l'intérêt de la procédure est aussi d'obliger les participants à se pencher sur le détail de l'évaluation, question par question : sur son contenu et sa formulation. On évite ainsi qu'une évaluation fondée uniquement sur une statistique globale se réfère implicitement à des normes traditionnelles et arbitraires du suffisant telles que l'obtention des deux tiers ou des trois quarts des points. La sélection des items où les divergences sont les plus importantes incite les experts à chercher les raisons des différences d'estimations, notamment dans l'importance diverse donnée à certaines notions, tant dans l'enseignement que dans les contrôles. Ainsi, l'évaluation ne sert pas seulement au contrôle des apprentissages (*assessment **of** learning*) mais aussi à favoriser ces apprentissages (*assessement **for** learning*)[16]. Un rapport des experts aux instances qui les ont mandatés pourrait contenir des recommandations en ce qui concerne les prochains contrôles et éventuellement un aménagement de l'enseignement (plan d'études ou méthodes).

La méthode est également utilisable pour des *questions à réponses construites*, fréquentes dans certaines épreuves scolaires actuelles, et une *méthode d'Angoff* dite *étendue* a été développée à cet effet (Hambleton & Plake, 1995). Elle peut s'appliquer également à des *formats mixtes* (Cisek & Bunch, 2007, p. 82). Nous avons eu l'occasion de tester cette dernière approche, sans problème particulier, à une épreuve de mathématiques passée en fin de scolarité obligatoire (Bain, 2016 ; Frey, 2016).

3.3 Problèmes et limites de la méthode d'Angoff

3.3.1 Adaptation de l'épreuve à une analyse par la méthode d'Angoff

Comme le relèvent Cisek & Bunch (2007, p. 6), il est hautement recommandable d'envisager l'application de la méthode dès la conception et l'élaboration de l'épreuve : « Standard setting is best considered early enough to align with the identified purpose of the test; to align with the selected test item or task formats ». Dans notre recherche sur l'examen de grammaire, nous avons appliqué la méthode après passation de l'épreuve, ce qui nous a obligé à modifier les critères de correction pour éviter certains cas de dépendance statistique entre items tels qu'ils avaient été conçus au départ.

Nous avons rencontré des problèmes analogues dans la recherche portant sur les épreuves de mathématiques passées en fin de scolarité obligatoire (Bain, 2016 ; Frey, 2016), notamment du fait que des points supplémentaires étaient accordés ou des pénalisations infligées à des groupes d'items. Dans le cas de questions complexes ou construites, l'« itemisation » des réponses *a posteriori* conduit à bricoler des solutions pouvant introduire certains biais. Ceux-ci sont surtout gênants lors du traitement statistique des résultats, les modèles utilisés supposant la non-dépendance entre items.

[16] Cf. Sur le thème « Assessment for learning », cf. Allal & Laveault, 2016.

3.3.2 Recrutement et sélection des évaluateurs

Dans la littérature consultée et dans les expériences auxquelles nous avons participé, on relève trois possibilités dans des épreuves analogues à notre examen : faire appel à

a). des experts de la discipline, généralement des spécialistes de niveau universitaire, externes à l'institution, donc non impliqués dans la formation dispensée, souvent désignés par les autorités scolaires ou politiques à des fins précisément d'expertise ;

b). des enseignants ou praticiens de la discipline formant le collège des formateurs à l'intérieur de l'institution et souvent associés aux travaux de didactique dans la branche considérée ;

c). des étudiants ayant récemment terminé avec succès la formation évaluée.

Cette dernière solution, étonnante au premier abord, a été testée par Verhoeven & al. (1999), dont la recherche portait précisément sur « la fiabilité et la crédibilité d'une procédure d'Angoff de fixation de standard pour un test de progrès recourant à des étudiants récemment diplômés » (notre traduction du titre de l'article). Ce test était appliqué quatre fois par an tout au long des études de médecine pour aider les étudiants à situer leur progression dans les objectifs de formation. Les huit juges étaient docteurs en médecine diplômés de l'université de Maastricht depuis environ cinq mois. Sans commenter ici plus avant cette recherche, notons simplement que cette solution était envisageable dans la mesure où l'objectif de l'évaluation était formatif et centré sur la réussite académique des études de médecine et non, plus directement, sur la capacité à exercer la profession. Les autres contrôles sanctionnant la fin de la formation médicale et utilisant la méthode d'Angoff réunissent généralement des experts de la discipline.

Autre expérience avec des étudiants réalisée dans le cadre de notre groupe Edumétrie : dans son cours de didactique, Laura Weiss a demandé à 11 futurs enseignants de physique dans l'enseignement secondaire d'estimer le seuil de suffisance pour un examen d'entrée à l'université (Bain & Weiss, 2016). La compétence des évaluateurs, dans ce cas, tenait à leur expérience relativement récente des études universitaires scientifiques et de leurs exigences. L'expérience montre une assez bonne homogénéité des estimations du seuil de suffisance (moyenne de 62% des points) ; la principale source d'erreur type absolue (près de 60% de la variance totale d'erreur absolue) était due à des estimations portant sur des questions de difficultés très différentes.

Choisir des experts universitaires, externes à la formation évaluée, ou des enseignants de la discipline ? L'expérience de Cisek & Bunch (2007, p. 22) « montre, d'une part, que des conseillers indépendants, externes apportent des avis très valables : ils offrent habituellement une vision des choses, une expérience, des idées, etc., des contributions qui ne seraient pas disponibles sans eux et qui en général améliorent la qualité des procédures de fixation de standards et la légitimité des résultats. Mais d'autre part, de tels experts-conseillers ont souvent des points de vue et des objectifs qui peuvent ne pas être partagés par les instances responsables finalement des standards » (notre traduction). Dit autrement, et selon notre propre expérience, il est facile pour les responsables finaux de l'évaluation d'écarter tout ou partie des conclusions des experts externes en leur déniant, explicitement ou implicitement, une (bonne) connaissance du terrain.

La participation de praticiens dans la procédure d'Angoff apporte l'avantage d'une expérience de l'enseignement évalué, voire du suivi de certains élèves, mais parfois sans une connaissance suffisante du plan d'études ou sans une distance adéquate par rapport à leur

expérience actuelle de l'enseignement (cf. notre expérience en mathématiques, Bain 2016 et Frey, 2016).

Le choix du panel d'évaluateurs dépendra donc du cadre institutionnel et du genre de contrôle visés par l'évaluation, selon qu'il s'agit d'une évaluation formative d'une école (en général par les enseignants eux-mêmes) ou d'une expertise du fonctionnement du système de formation (la plupart du temps par des experts externes). C'est pourquoi une autre solution, préconisée par Capey et Hay (2013), nous laisse sceptique : elle consisterait à sélectionner différents types d'évaluateurs, experts et praticiens / généralistes, sous prétexte d'élargir les points de vue. Au contraire, dans bien des cas, il s'agit, comme nous venons de le dire, de préciser le point de vue adopté pour la fixation de standards et de recruter les évaluateurs en conséquence. L'hétérogénéité du groupe risque de déboucher sur des estimations très divergentes en phase 1, que la discussion ne pourra guère réduire si les références des uns et des autres sont très différentes par rapport à l'objectif de l'expertise.

3.3.3 Détermination du nombre d'experts à recruter

Dans la quasi-totalité des travaux sur la méthode d'Angoff, les auteurs indiquent ou suggèrent un nombre minimum de juges (5, 8, 10, >10…) ou une fourchette…, sorte de règle empirique (*rule of thumb*), souvent sans indiquer la source ou la justification de cette information, voire de cette prescription. Les *Standards for Educational and Psychological Testing* (AERA / APA / NCME, 1999, p. 54) recommandent d'engager dans la procédure « un groupe de juges suffisamment nombreux et représentatifs pour fournir une garantie raisonnable que les résultats ne varient pas considérablement si la procédure était répétée » (notre traduction). Ce qui renvoie implicitement à une analyse de statistique inférentielle et, pour nous, à une analyse au moyen du modèle de la généralisabilité (cf. Bain, 2018).

Sur ce point également, il est donc bien difficile de généraliser à partir de recherches très diverses quant à leur objet ou leurs modalités. Disons simplement qu'il serait imprudent de se lancer pour la première fois dans une procédure d'Angoff avec moins de 10 experts si l'on vise à minimiser l'erreur type absolue sur le seuil recherché. Il faut être conscient en effet que trois autres facteurs augmentent statistiquement l'erreur type absolue sur le standard recherché : un petit nombre de questions, une forte dispersion de leurs niveaux de difficulté (donc de leurs estimations) ou la tendance des évaluateurs à modifier leur niveau d'exigence selon la question (d'où parfois une importante interaction Évaluateurs x Questions).

En avançant ci-dessus ce nombre d'une dizaine d'experts, nous soulignons un autre problème potentiel de la méthode : trouver suffisamment d'évaluateurs ayant l'expertise visée et disponibles tout au long de la procédure, ce que constate aussi Klein (2008, p. 107) « The number of available judges is limited ».

3.3.4 Difficulté de compréhension et d'application de la consigne

Nous l'avons expérimentée en début de procédure. Il n'est pas évident de se représenter la limite séparant les « étudiants juste suffisants » et les autres, et à chiffrer en % la probabilité de leur réussite. Pour cette estimation, se représenter un groupe de 100 borderlines ne se révèle pas vraiment facilitateur. C'est certainement le cas dans d'autres recherches ; nous en donnons le témoignage, peut-être ironique, de cette consigne de Dever (2015, p. 1) :

« Pour mieux comprendre le concept du candidat minimalement compétent, il s'avère souvent utile d'observer ses collègues de travail[17]; quelques-uns sont des « vedettes », dont le

[17] En l'occurrence, il s'agit de médecins.

rendement est à un niveau bien au-dessus de la majorité, tandis que celui de certains collègues est plutôt mauvais, sans compter que certains ne devraient peut-être même pas avoir le droit d'exercice de la profession. Quelque part entre ces deux extrêmes se trouve le groupe dont le rendement constitue le niveau de compétence minimal. Le candidat limite appartient au groupe qui se qualifie tout juste pour l'agrément ou l'obtention d'un permis ».

Une origine probable de la difficulté d'estimation demandée découle du fait que l'évaluateur peut avoir de la peine à concilier plusieurs références ou critères : le *degré d'attentes ou d'exigences* à l'égard du futur instituteur, le *degré de certitude* de l'évaluation proposée, avec en arrière-fond une référence aux compétences grammaticales d'une population d'apprenants familière[18]. On peut donc se demander, avec plusieurs chercheurs cités par Wyse & Reckase (2012, p. 6), si « la tâche d'émettre des jugements de probabilités selon la méthode d'Angoff n'est pas excessivement complexe » (notre traduction). Nos observations lors de la procédure montrent que la tâche est effectivement complexe, mais que les problèmes ne sont pas insurmontables.

Il est certainement difficile d'éviter que les experts se réfèrent à la population d'étudiants à laquelle ils ont affaire régulièrement. Nous avons essayé de prévenir ce biais en écartant explicitement cette référence de la consigne pour la seconde phase. Il est toutefois assez artificiel de situer ses estimations dans l'absolu des exigences du plan d'études, par ailleurs peu explicites sur les performances attendues (Marc & Wirthner, 2013, p. 5), et la référence au niveau de compétence exigé par l'exercice de la profession est sujette à des interprétations diverses. Pour bien des évaluateurs il est probablement nécessaire de fixer, comme nous l'avons fait, quelques balises le long de l'échelle en %, se référant implicitement à un degré de difficulté de la question, toujours pour les borderlines.

Relevons encore un problème observé assez couramment dans nos travaux sur l'évaluation, et récemment dans l'application de la méthode d'Angoff à des épreuves de mathématiques au Cycle d'orientation genevois en fin de scolarité obligatoire (Bain, 2016 ; Frey, 2016 ; Frey, 2017) : la tendance des enseignants à être relativement optimistes – ou très exigeants, c'est selon – quant à la réussite de leurs élèves… ou de leur enseignement. Cela les incite à fixer *a priori*, avant tout feed-back sur les résultats de l'épreuve, un seuil relativement élevé.

Par ailleurs, nous n'avons pas relevé de difficulté majeure dans le maniement d'une échelle critériée en %. En observant l'ensemble des estimations aux deux phases, on peut se demander s'il est utile de proposer une précision par pas de 5%. Elle est utilisée pour une minorité des estimations (16%) en phase 1 comme en phase 2. Mais dans ces évaluations, comme dans d'autres, interviennent des différences personnelles marquées, opposant les experts qui n'utilisent pratiquement pas – ou très peu – les intervalles de 5% et ceux, minoritaires, qui y recourent pour plus de 10 questions sur 56, avec une forte corrélation entre les deux phases. Il nous semble finalement préférable d'autoriser cette latitude d'estimation pour éviter des blocages chez certains évaluateurs pour lesquels une échelle en déciles apparaîtrait trop sommaire.

Au départ, nous avons hésité à utiliser la *méthode d'Angoff Oui/Non*, apparemment plus simple d'emploi : il s'agit d'estimer si un candidat juste suffisant va réussir ou non la question (codage 1/0). Nous y avons renoncé pour deux raisons : cette estimation dichotomique nous paraissait d'abord trop grossière par rapport à la probabilité de réussir ou échouer telle

[18] « A criticism of Angoff has been that the validity of resulting cut scores may be threatened due to panelists' difficulty in performing the cognitively complex task of estimating probabilities and to inconsistency between panelists' item ratings and actual student performance data. » (Peterson, Schulz, Engelhard, 2011, p. 4).

question formulée de telle façon. Cette variante se révèle par ailleurs particulièrement inadéquate pour un test de maîtrise comme l'examen de grammaire, pour laquelle on postule des estimations situées entre 50% et 100% (et *de facto* plutôt entre 75% et 100%). Cisek & Bunch (2007, p. 94) en font rapidement la démonstration et l'illustration :

« Un biais potentiel se produit parce que la méthode [Oui/Non] est fondée sur un jugement implicite exigeant de décider si la probabilité d'une réponse correcte en référence au score de passage est plus grande que .5 [50%]. Pour illustrer ceci, supposons qu'un test soit composé d'items identiques qui aient tous une probabilité de réponse en référence au score de passage de .7 [70%]. Un évaluateur rigoureux dans ses estimations assignerait la valeur de 1 à chaque item, et le standard de performance résultant serait un score parfait [100% de réussite], ce qui n'est clairement pas l'intention de l'évaluateur ni une attente réaliste fondée sur la difficulté du test » (notre traduction).

On conçoit l'inadéquation de cette variante dans le cas d'un test de maîtrise comme l'examen de grammaire pour lequel on pouvait attendre *a priori* (cf. le seuil fixé par le professeur) à une moyenne de 75% de réussite.

3.3.5 Temps nécessaire pour parcourir les étapes de la procédure

C'est un des problèmes le plus souvent signalés dans la littérature consultée : « A drawback of the Angoff method is the time involved of the use of experts in panels, and therefore the costs. This was confirmed by the comments the judges made about the time consuming procedure. The judges needed about 2 hours for this procedure. » (Klein, 2008, p. 10). Il faut en effet du temps pour exposer et entraîner la méthode ; pour analyser et coter chaque question ; pour prendre connaissance du détail des résultats ; pour discuter les points de divergence entre les deux phases et décider du seuil final ; pour rédiger et discuter le rapport aux commanditaires de l'évaluation, éventuellement le commenter aux instances concernées. La durée (totale) de deux heures avancée par Klein nous semble sous-évaluée, en particulier si l'on a affaire à des experts novices ou s'ils ils fonctionnent ensemble pour la première fois. Ce travail ne figurant généralement pas dans le cahier des charges des experts, l'opération correspond également à un coût non négligeable, que les instances administratives s'efforcent actuellement de réduire.

3.3.6 Efficacité relative et problèmes de la discussion interphase

Cette étape d'échanges entre évaluateurs, avant la première étape et entre la première et la seconde (éventuellement avant la troisième), est jugée cruciale pour la qualité et la convergence des estimations. Clauser & al. (2009) en font l'objet d'une étude spécifique, qui conclut (p. 2) que la discussion des divergences entre évaluateurs a diminué la variance associée aux effets Juges et Juges x Items, indice d'un accord accru entre les juges.

Dans notre expérience, cette diminution est de faible ampleur, de même que dans la recherche sur les épreuves de mathématiques (Bain, 2016 ; Frey, 2016 et 2017), et cela risque d'être assez souvent le cas quand on ne fait pas intervenir dans les évaluations des feed-back d'impact ou de réalité. En effet, faute de temps et compte tenu du nombre de questions, il est probablement rare qu'on puisse approfondir les cas de divergences, et plusieurs d'entre elles risquent de subsister. C'est l'expérience que nous avons faite ; elle laisse par ailleurs supposer des différences individuelles quant à la capacité ou à la volonté de modifier ses estimations : la moitié de nos 10 experts n'ont guère changé globalement leurs exigences d'une phase à l'autre. D'autant que, plus généralement, ces divergences peuvent tenir à des positions personnelles tant épistémologiques (statut de la branche), que pédagogiques

(investissement dans la formation), ou docimologiques (pertinence du mode ou de la forme de l'évaluation).

Mentionnons par ailleurs un biais possible lors de cette discussion, susceptible d'influer sur les estimations de la seconde phase : le leadership exercé *de facto* par tel ou tel participant. Pour paraphraser un des commandements de la *Ferme des animaux* (Orwell, 1945, p. 89), théoriquement « tous les [experts] sont égaux, mais certains sont plus égaux que d'autres » : « Caution must be exercised in interpreting the reliability coefficient since it might be influenced by one judge who dominates others » (Mortaz & Jalili, 2014, p. 5).

3.3.7 Renonciation à une phase 3 de feed-back de réalité et d'impact ?

Compte tenu de ces divergences encore relativement importantes en fin de phase 2, n'aurait-il pas fallu passer à une troisième phase ? Rappelons que lors de cette étape (Cisek & Bunch, pp. 55-56 et 84), on injecte dans la procédure des informations sur les résultats effectifs à l'examen ou à des contrôles équivalents antérieurs (cf. supra § 2.5.3). L'objectif est d'améliorer la convergence des estimations en fournissant aux experts des références communes.

Nous avons finalement renoncé à une telle étape faute de temps, mais surtout parce que, dans les deux types de feed-back, on tend à mélanger les références normatives et critériées, changeant ainsi le point de vue adopté au départ de notre recherche. Dans la pratique en revanche, en dehors des cas d'expertise particuliers, nous constatons que le recours à un *feed-back de réalité ou d'impact* s'impose souvent. Il est notamment la conséquence de la tendance des enseignants (ainsi que des experts-enseignants) à être relativement optimistes ou exigeants dans leurs estimations. Nous l'avons remarqué notamment dans l'application de la méthode d'Angoff à des épreuves de mathématiques au Cycle d'orientation genevois en fin de scolarité obligatoire (Bain, 2016 ; Frey 2016 et 2017) : pour certaines filières, le seuil fixé *a priori* était proche de la moyenne des scores, d'où théoriquement un taux d'échecs de près de 50%, que l'institution n'aurait pu se permettre pour des raisons évidentes. La solution adoptée finalement est généralement un compromis : « Pour fixer le seuil [final], le niveau de compétence attendu des élèves (seuil Angoff) et les rendements effectifs des élèves ont été pris en considération » (Frey, 2017, p. 5). Et encore faut-il que le taux final d'échecs soit défendable face aux instances extérieures (notamment politiques) susceptibles de critiquer l'institution de formation (Cizek & Bunch, 2007, p. 2).

3.3.8 Prise en compte de l'erreur de mesure sur le seuil

Tout se passe comme si cette notion d'erreur de mesure sur un score ou un seuil n'existait pas dans la culture évaluative sous nos latitudes. Elle est rarement mentionnée dans les comptes rendus relatifs aux épreuves de référence et aux examens, et probablement pas prise en considération pour le calcul des barèmes. On peut notamment l'expliquer par le fait qu'il n'est socialement pas de bonne stratégie de parler d'erreur dans le cas d'épreuves qu'on doit, de plus en plus souvent, défendre contre des critiques extérieures ou des recours de plus en plus fréquents des évalués. Dans la pratique, cela a pour conséquence d'ignorer *de facto* l'existence d'une telle erreur ou de faire comme si cette erreur n'existait pas ; il n'y a donc pas lieu d'en évaluer l'importance ni de décider qui faire profiter de l'intervalle d'incertitude calculé : l'apprenant ou l'institution. Généralement, les responsables de l'évaluation adoptent en effet ce que Cisek & Bunch (2007, p. 29) appellent la solution « par défaut » : « In fact, the equal weighting of false positive and false negative classification errors is effectively the "default" weighting that is adopted when the issue of relative costs is not deliberated. »

En dehors de cette « solution par défaut », deux cas de figure se présentent : selon le contexte institutionnel et social, ainsi que selon les conséquences de la décision finale, on peut décider de faire bénéficier l'évalué – ou au contraire l'institution – de la marge d'erreur en la soustrayant ou en l'additionnant au seuil de réussite. Il s'agit d'écarter ce que Cizek & Bunch (2007, p. 25) appellent des *décisions de classement* respectivement *fausses positives et fausses négatives*.

Dans le cadre de la formation, de l'orientation ou de la sélection scolaires, pour éviter l'erreur consistant à considérer comme incompétent un candidat effectivement compétent, on décide généralement d'abaisser le seuil de passage en soustrayant la marge d'erreur afin de « donner leur chance » aux intéressés. Pour l'examen de grammaire pris comme exemple, dans cette perspective, on fixerait donc le seuil à 79% (84% - 5%). Un *feedback d'impact*, se référant aux résultats effectifs à l'examen, nous indiquerait que le taux d'échecs serait ainsi d'environ 5% de l'effectif des candidats enseignants (avec le seuil de 75% adopté par le professeur responsable du cours, on comptait 1% d'échecs)

Dans le cas d'un examen final de médecine, en revanche, on peut être amené à additionner au contraire la marge d'erreur (84% + 5% = 89%) pour éliminer des candidats médecins susceptibles de mettre en danger leurs premiers patients : « In cases where the consequences or costs of false positive decisions are […] serious […], those participating in a standard-setting procedure might recommend a very high standard to preclude a large proportion of false positive decisions. » (Cizek & Bunch, 2007, p. 27).

Une autre façon d'utiliser l'intervalle de confiance est de le considérer comme une *zone d'incertitude* et, pour lever cette incertitude, de recourir à une des stratégies de décision séquentielle préconisées par Cronbach & Gleser (1969) en faisant intervenir une autre information pour les étudiants concernés. Il s'agit souvent d'une nouvelle passation de l'examen[19]. Dans notre exemple de l'examen de grammaire, ce pourrait être uniquement pour ceux se situant entre 79% et 83% de réussite.

3.4 En guise de conclusion

Ce texte a été rédigé en hommage à Jean Cardinet, chercheur, collègue et ami avec lequel j'ai échangé pendant plusieurs décennies sur les problèmes posés par l'évaluation dans le domaine des sciences de l'éducation. C'est donc à lui que je donne d'abord la parole dans ces quelques mots conclusifs en reproduisant le dernier paragraphe de son texte de 2014 qui faisait la critique de l'article de Verhoeven & al. (1999).

« A plus long terme, il faut remettre en cause toute la procédure d'examen, car l'étude statistique que nous venons de faire montre bien que, pour des raisons pratiques de temps et de coût, l'incertitude ne pourra jamais être suffisamment réduite, même pour l'estimation apparemment simple du niveau minimum requis en utilisant la méthode d'Angoff. Le modèle statistique n'est pas pour autant à rejeter, car il a au moins l'intérêt de révéler quelles sont ses limites. C'est plutôt l'ambition d'une évaluation bilan qu'il faudra sans doute abandonner au profit de procédures plus élaborées d'évaluation formative. »

J'ai pris connaissance de ce texte quelques semaines seulement avant le décès de Jean Cardinet. Je n'ai donc pas pu réagir à cette conclusion et avoir avec lui un de ces échanges fructueux et passionnants, notamment quand nos avis divergeaient quelque peu. Je lui aurais

[19] Citant Millman (1989), Cizek & Bunch (2007, p. 26) soulignent le danger d'une telle stratégie si l'institution autorise de multiples passations de l'examen : « The examinee has a better than 50/100 chance of capitalizing on random error and passing the test in only five attempts! ».

dit mon accord de principe avec l'idée d'affecter l'essentiel de nos efforts à développer l'évaluation formative, à laquelle j'ai consacré quelques-uns de mes travaux. En revanche, je ne suis pas sûr qu'on puisse jamais abandonner évaluations bilans et examens. Un contrôle sommatif des résultats de la formation s'impose, en particulier pour la scolarité obligatoire ; il est en cours de façon systématique dans de nombreux pays. Il est méthodologiquement et didactiquement hautement recommandable que des experts examinent les résultats de ces enquêtes avec une méthode du type d'Angoff pour estimer s'ils correspondent au niveau de performances attendu par les formateurs, mais aussi par les instances politiques auxquelles ils doivent des comptes. En ce qui concerne les examens et les divers tests à enjeux élevés, dont dépend souvent la carrière scolaire ou professionnelle des apprenants, il s'agit de réduire un arbitraire dont on n'a pas toujours conscience. Les méthodes de *standard setting* du type de celle d'Angoff pourraient apporter une contribution importante à un tel objectif.

4. Références

AERA/APA/NCME (1999). *Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association, American Psychological Association National Council on Measurement in Education.

Allal, L., & Laveault, D. (2016). *Assessment for Learning: Meeting the Challenge of Implementation.* Cham: Springer.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (pp. 508–600). Washington, DC: American Council on Education.

Astolfi, J.-P. (1993). Trois paradigmes pour les recherches en didactique. *Revue française de pédagogie*, *volume 103*, 5-18.

Bain, D. (2010). Pour évaluer les qualités docimologiques des tests de maîtrise : l'intérêt de recourir à la généralisabilité. *Mesure et Évaluation en Éducation, 33*(2), 35-63.

Bain, D. (2016). *Math. 11e Tronc commun. Résumé des analyses de généralisabilité***.** Genève : Groupe Edumétrie, Société suisse de recherche en éducation (SSRE) ; rapport de recherche.

Bain, D. (2018). *Fixer un seuil de réussite pour un test de maîtrise : intérêt et limites de la méthode d'Angoff et de de la généralisabilité.* Genève : Groupe Edumétrie, téléchargeable sur https://www.irdp.ch/institut/edumetrie-1635.html

Bain, D., & Weiss, L. (2016). *Épreuve passerelle de physique : commentaires des résultats.* Genève : Groupe Edumétrie, Société suisse de recherche en éducation (SSRE) ; rapport de recherche.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion- referenced tests. *Review of Educational Research, 56,* 137–172.

Busch, J., & Jaeger, R. (1990). Influence of type of judge, normative information, and discussion on standards recommended for the National Teacher Examinations. *J. Educ. Meas.*, *27*(2), 145–163.

Bronckart, J.-P. (2004). *Syllabus de grammaire 1.* Genève: Faculté de psychologie et des sciences de l'éducation, Université de Genève.

Capey, St., & Hay, Fr. C. (2013). Setting the standard in assessments. In O. M. R Westwood., A. Griffin, & Fr. C. Hay (Eds). *How to Assess Students and Trainees in Medicine and Health.* New Jersey: Wiley-Blackwell, 94-113.

Cardinet, J. (1972). *Adaptation des tests aux finalités de l'évaluation.* Neuchâtel : Institut romand de recherches et de documentation pédagogiques (R72-9).

Cardinet, J., & Tourneur, Y, (1985). *Assurer la mesure.* Berne : Peter Lang.

Cardinet, J. (2014). Discussion de l'article de Verhoeven, Van der Steeg, Scherpbier, Muijtjiens, Verwijnen & van der Vleuten, *Medical Education, 33*, 832-837. Communication personnelle, s.l.n.d.[20]

Cardinet, J., Johnson, S., & Pini, G. (2010). *Applying Generalizability Theory using EduG*. New York: Routledge/Taylor & Francis (Quantitative Methodology Series).

CDIP (2007). *HarmoS : Objectifs nationaux de formation*. Conférence intercantonale de l'instruction publique de la Suisse romande et du Tessin. Accès : http://www.ciip.ch/documents/showFile.asp?ID=2910.

Çetin, S., & Gelbal, S. (2013). A Comparison of Bookmark and Angoff Standard Setting Methods, *Educational Sciences: Theory & Practice, 13*(4). Educational Consultancy and Research Center (2169-2175).

CIIP (2007), *Convention scolaire romande*, Espace romand de la formation Conférence intercantonale de l'instruction publique de la Suisse romande et du Tessin. Accès : http://www.ciip.ch/FileDownload/Get/80

CIIP (2010-2016). *Plan d'études romand*. Conférence intercantonale de l'instruction publique de la Suisse romande et du Tessin. Accès : https://www.plandetudes.ch/ consulté le 12.08.2016.

Cizek, G. J., & Bunch, M. B.(2007). Standard setting: a guide to establishing and evaluating performance standards on tests. London: Sage Publications.

Clauser, B., Harik, P., Margolis, M., McManus, I., Mollon, J., Chis, L., & Williams, S. (2009). An empirical examination of the impact of group discussion and examinee performance information on judgments made in the Angoff standard-setting procedure. *Appl. Meas. Educ.*, *22*(1):1–21.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological Tests and Personnel decisions*. Urbana: University of Illinois Press.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

COE (2017). *Cadre européen de référence pour l'enseignement des langues*. Strasbourg : Conseil de l'Europe, Unité des Politiques linguistiques, Strasbourg, Accès : https://rm.coe.int/16802fc3a8.

De Landsheere, G. (1979). *Dictionnaire de l'évaluation et de la recherché en éducation*. Paris: Presses universitaires de France.

D'Hoop, E., Lemenu, D., Malhomme, Chr., & Coupremanne, M. (2012). *Articulation entre référentiels, pratiques d'enseignement, dispositifs de formation et pratiques d'évaluation.* Texte final destiné aux Actes du 24e colloque de L'ADMÉÉ-Europe : L'évaluation des compétences en milieu scolaire et en milieu professionnel.

DEPP (2014). *Rapport technique - CEDRE Mathématiques École 2014*. Paris : Direction de l'évaluation, de la prospective et de la performance, Ministère de l'éducation nationale, de l'enseignement supérieur et de la recherche.

DEPP (2015) : *Note d'information no 19*, mai 2015. Paris : Direction de l'évaluation, de la prospective et de la performance, Ministère de l'éducation nationale, de l'enseignement supérieur et de la recherche.

Dever, E. (2015). *La fixation des normes selon la méthode Angoff*. Téléchargé sur le site de l'ACTRM : http://www.camrt.ca/fr/wp-content/uploads/sites/3/2015/05/La-m%C3%A9thode-Angoff.pdf, le 20.12.17.

Frey, J. (2016). *Étude exploratoire sur l'utilisation de la méthode d'Angoff pour déterminer a priori les seuils de suffisance des EVACOM de mathématiques*. Genève : Direction générale de l'enseignement obligatoire.

Frey, J. (2017). *Utilisation de la méthode d'Angoff pour déterminer a priori les seuils de suffisance des EVACOM de mathématiques 2017*. Genève : Direction générale de l'enseignement obligatoire.

George, S, Haque, S, & Oyebode, F. (2006). Standard setting: comparison of two methods. *BMC Med Educ. 6*(1):46.

[20] Ce texte nous a été envoyé en mai 2015 par Jean Cardinet, à la suite d'un échange sur le texte de Verhoven & al., 1999. Il ne contenait aucune information sur sa date ni sur un éventuel lieu d'édition (s.l.n.d). Nous l'avons daté de 2014 en fonction de la date du fichier reçu.

Hambleton, R. K., & Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education, 8*, 41–56.

Hurtz, G, & Auerbach M. A. (2003). A meta-analysis of the effects of modifications to the Angoff method on cutoff scores and judgment consensus. *Educ. Psychol. Meas., 63*(4): 584–601

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, *64*(3), 425.

Klein, M. E. (2008). *The use of the objective structured clinical examination (OSCE) in dental education*. Thesis. Department of Periodontology of the Academic Centre for Dentistry Amsterdam (ACTA), 152 p., Downloaded from UvA-DARE, the institutional repository of the University of Amsterdam (UvA). http://hdl.handle.net/11245/2.55005.

Marc, V., & Wirthner, M. (2012). *Épreuves romandes communes : de l'analyse des épreuves cantonales à un modèle d'évaluation adapté au PER – Rapport final du projet EpRoCom*. Neuchâtel : Institut de recherche et de documentation pédagogique.

Marc, V., & Wirthner, M. (2013). *Développement d'un modèle d'évaluation adapté au PER. Rapport scientifique du projet d'épreuves romandes communes*. Neuchâtel : Institut de recherche et de documentation pédagogique.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–104). New York: Macmillan.

Mortaz, H. S., & Jalili, M. (2014). Standard setting in medical education: fundamental concepts and emerging challenges. *Medical Journal of the Islamic Republic of Iran, 28*(34), published online 2014 May 19.

Orwell, G. (1989). *Animal Farm : A Fairy Story*. London, Penguin, coll. « Fiction », 1989 (1re éd. 1945).

Perrenoud, Ph. (1984). *La fabrication de l'excellence scolaire: du curriculum aux pratiques d'évaluation*. Genève : Droz.

Peterson, C., Schulz, E.,M., & Engelhard, G. (2011). Reliability and validity of bookmark-based methods for standard setting: comparisons to Angoff-based methods in the National Assessment of Educational Progress. *Educ. Meas., 30*(2):3–14.

Shulruf, B., Wilkinson, T., Weller, J., Jones, Ph. & Poole, Ph. (2016). Insights into the Angoff method: results from a simulation study. *BMC Med Educ.*, *16*: 134. Published online 2016 May 4.

Site Edumétrie (2017). Hébergé par l'Institut de recherche et de documentation pédagogique Neuchâtel, à l'adresse : https://www.irdp.ch/institut/edumetrie-1635.html.

Verhoeven, B, Van der Steeg, A, Scherpbier, A, Muijtjens, A, Verwijnen, G, & Van Der Vleuten, C. (1999). Reliability and credibility of an Angoff standard setting procedure in progress testing using recent graduates as judges. *Med. Educ.*, *33*(11), 832–837.

Wheaton, A., & Parry, J. (2012). *Using the Angoff Method to Set Cut Score*s. New Orleans Questionmark, 2012, March 20-23. Users conference

Wyse, A., & Reckase, M. (2012). Examining rounding rules in Angoff-type standard-setting methods. *Educ. Psychol. Meas. 72*(2), 224–244.

Yerly, G. (2014). *Les effets de l'évaluation externe des acquis des élèves sur les pratiques des enseignants. Analyse du regard des enseignants du primaire*. Université de Fribourg. Thèse de doctorat sous la direction de Gurtner, Jean-Luc.