L'évaluation par les pairs dans le contexte de la formation

Virginie Galtier

université
PARIS-SACLAY

CentraleSupélec

L'évaluation par les pairs dans le contexte de la formation

Virginie Galtier

2016

Objectifs du document

L'évaluation par les pairs se pratique dans plusieurs contextes ; dans ce document je me concentre sur celui de la formation. L'objectif est d'une part de préparer le cahier des charges détaillé du projet CAFPA [1] et d'autre part de réunir quelques éléments pour présenter la pratique aux collègues.

Pour la rédaction je m'appuie sur l'expérience d'évaluation par les pairs que je mène depuis plusieurs années dans un cours électif (1A (L3) et 2A (M1), une vingtaine d'élèves à chaque fois) et sur celle acquise sur Coursera (en tant que formatrice mais également en tant qu'apprenante). Le développement des MOOC a mis en lumière ce type d'évaluation et des paramètres diffèrent par rapport au contexte des classes traditionnelles (notamment l'échelle, les liens inter-apprenants et les possibilités d'échange, le synchronisme, la culture commune ou non...), mais le socle de l'évaluation par les pairs reste identique. Ce document s'appuie aussi sur une étude bibliographique limitée. Il a vocation à être complété par un autre document réalisant une analyse comparative des outils disponibles, et par une étude sur les fonctions de fiabilité des notes attribuées.

1. Projet de création d'une plate-forme d'évaluation par les pairs capable de supporter plusieurs cas d'usage. Soutenu par l'Université de Paris-Saclay en 2016.

Table des matières

Chapitre 1

Description du processus

En première définition, on peut décrire l'évaluation par les pairs de la manière suivante :

- L'enseignant fournit aux étudiants des consignes pour produire un travail ("énoncé du devoir").
- Chaque étudiant réalise alors le travail demandé.
- Chaque production/copie est transmise à un ou plusieurs autres étudiants.
- Chaque étudiant recevant le travail d'un camarade l'évalue.
- Parfois un professeur évalue également tout ou partie des copies.
- La note finale d'une copie est une combinaison des notes qui lui ont été attribuées par les différents évaluateurs. La fonction de combinaison peut être relativement sophistiquée, de manière à minimiser les conséquences du manque de fiabilité de certaines évaluations.
- La méta-évaluation est une étape supplémentaire et facultative au cours de laquelle est évaluée la qualité des évaluations fournies par les élèves à leurs camarades ; elle peut donner lieu à une note d'évaluateur.

Cette première définition regroupe en réalité de nombreuses variantes, des cas d'usage variés. Dans la suite de ce chapitre, nous décrivons des principales options possibles.

1.1 Travail personnel ou collectif

Si dans un MOOC tout travail est généralement personnel, c'est souvent moins vrai dans le cas d'une classe "physique". On peut donc imaginer que les travaux à rendre soient des productions individuelles ou bien des productions collectives : compte-rendu de séance de travaux pratiques réalisée en binôme, projet réalisé en petit groupe etc. De même, on peut aussi imaginer que l'évaluation d'une production (qu'elle soit individuelle ou collective) se fasse de manière individuelle ou bien collégiale. Cependant, hormis [4] et `https://www.economicsnetwork.ac.uk/showcase/crockett_peer.htm`, je n'ai pas trouvé d'étude impliquant des groupes dans l'évaluation par les pairs. Quand il est question d'évaluation par les pairs et de groupe dans la littérature le plus souvent cela fait référence à l'évaluation du comportement d'un membre du groupe par rapport au travail d'équipe ("évaluation des co-équipiers").

Mon expérience dans le cours électif de développement d'applications mo-

biles : les élèves sont répartis en binômes. Chaque binôme imagine le cahier des charges d'une application. Un enseignant vérifie que ces cahiers des charges présentent un intérêt et un niveau de difficulté acceptable. Chaque binôme reçoit ensuite le cahier des charges établi par un autre groupe et développe l'application spécifiée. L'application développée est ensuite évaluée par 3 binômes : celui ayant rédigé le cahier des charges et 2 autres groupes (et également un enseignant, qui met la note finale). Ces évaluations du travail sont ensuite analysées par un enseignant (qui attribue la note d'évaluateur). De 2013 à 2015 les évaluations été réalisées de manière individuelle, en 2016 les évaluations sont devenues collectives (comme décrit, par binôme). En effet, l'évaluation individuelle demande plus de travail au moment de la ventilation des copies (2 évaluateurs sont faciles à affecter, ce sont les auteurs du cahier des charges ; pour les 2 autres avis je voulais qu'ils soient issus d'élèves n'ayant pas travaillé sur le même projet pour minimiser les risques d'avoir des retours très similaires) ; d'autre part, il y avait en moyenne 8 applications, évaluées chacune par 4 élèves soit 32 évaluations à relire, avec l'évaluation par binômes on tombe à 8 applications évaluées par 3 binômes soit 24 évaluations pour la même diversité de feedback pour le binôme évalué. Disposer d'un outil adapté à ce cas d'usage facilitera la ventilation et si la méta-évaluation peut être (même partiellement) automatisée, on pourrait revenir à une évaluation individuelle. Cependant l'évaluation réalisée en groupe présente des avantages pour l'évaluateur (voir section 2.2) et je pense la maintenir.

Pour le dépôt sur une plate-forme de devoirs ou d'évaluations réalisés en groupe, il faudra être attentif à ce qu'un membre du groupe ne puisse pas se désolidariser du rendu ("j'étais d'accord pour qu'on réponde X, et à la dernière minute sans me consulter mes co-équipiers ont décidé de mettre Y").

1.2 Grille d'évaluation

De nombreuses évaluations s'accompagnent de grilles critériées. La plupart du temps seul l'enseignant à connaissance et utilise cette grille (par exemple pour moduler l'importance des questions, ou rester constant dans sa notation et limiter les variations dues au niveau de fatigue, à l'ordre de correction, à l'effet pygmalion...). Parfois cette grille est révélée à l'élève, souvent avec la mention "barème fourni à titre indicatif", comme moyen d'informer l'élève des questions les plus difficiles, ou plus importantes. Dans le cas de l'évaluation par les pairs cette grille revêt une importance capitale puisqu'elle va permettre d'éviter des écarts trop importants entre les évaluateurs.

1.2.1 Contenu d'une grille critériée

Dans sa définition habituelle, une grille critériée est constituée, pour chaque question, de :

- une dimension : sur quoi porte l'évaluation. Par exemple : la présentation d'un dessert.
- une pondération : relativement à l'ensemble du devoir, quel est le poids de cette question. Par exemple la présentation du dessert compte pour 10% de la note attribuée à un repas.

- une échelle d'appréciation (niveau de performance) avec des descripteurs. Par exemple, une présentation est jugée "pauvre" si le dessert comporte moins de 2 couleurs ou s'il est servi dans une assiette de moins de 10 cm de diamètre, la présentation est "satisfaisante" si les bords du gâteau sont nets et que les framboises sont alignées, et la présentation est "exceptionnelle" s'il y a des éléments de décoration en sucre filé ou que le dessert est flambé au service.

De manière générale et intuitive, plus la grille est précise et moins les notes seront subjectives [1].

[6] a conclu de ses expériences qu'un barème adoptant des structures grammaticales similaires d'un item à l'autre conduisent à des évaluations plus fiables.

Soulignons une proposition originale de [6] qui fournit en plus de la rubrique une liste de "conseils clef-en-main" (*"fortune-cookie feedback"*) que les évaluateurs peuvent adresser à l'évalué pour compléter la note par un "feedback activable" c'est-à-dire un feedback qui ne se contente pas de pointer les faiblesses dans une production mais qui propose aussi des pistes d'amélioration.

Fournir la grille d'évaluation en même temps (voire avant) l'énoncé permet aux apprenants de mieux cerner ce que l'enseignant attend d'eux, clarifie leur objectif. Dans certains cas néanmoins cela n'est pas souhaitable car cela pourrait pousser les étudiants à travailler de manière très scolaire en ne se focalisant que sur les points listés dans la grille d'évaluation. Cependant si la grille est bien conçue et suffisamment complète, cet écueil disparaît. Un compromis est alors à trouver dans la taille de la grille car si elle devient trop longue les évaluateurs vont avoir tendance à être moins appliqués.

1.2.2 Construction de la grille

L'enseignant peut choisir de ne pas fournir de grille ou guide d'évaluation et laisser les évaluateurs dresser eux-mêmes la liste des critères d'évaluation. Dans ce cas il sera certainement intéressant pour lui d'observer les critères qui ont été retenus par les élèves car ils donnent une indication de la valeur que les élèves accordent à différents aspects du travail. A priori dans cette situation on ne pourra guère se fier aux notes attribuées. La notation massive ne sera sans doute pas l'objectif recherché. Je n'ai pas trouvé d'article décrivant cette pratique.

La construction de la grille d'évaluation peut prendre la forme d'un atelier dirigé par l'enseignant ([9] mentionne d'ailleurs une telle construction collaborative dans des classes de collège).

De nombreux enseignants ont déjà fait l'expérience de devoir revoir leur barème initial après la correction des premières copies car ils réalisent alors qu'une question était en fait ambiguë par exemple. Examiner un échantillon de quelques productions peut servir pour valider ou éventuellement modifier sa grille.

1. Dans ce document on parlera souvent de note "fiable" pour désigner une note proche de celle qu'aurait attribuée un professeur et qui serait considérée comme note "de référence". Voir section 1.4.1

1.2.3 Comment favoriser l'appropriation de la grille par les évaluateurs ?

Rappelons que dans le contexte qui nous intéresse, les évaluateurs ne sont pas des professionnels. Il est le plus souvent bénéfique de les accompagner dans cette tâche d'évaluation qui ne leur est pas familière.

Une manière de s'assurer qu'ils s'approprient bien la grille d'évaluation est de la construire avec eux lors d'une séance de travaux dirigés par exemple. La conception collaborative de la grille pourrait peut-être s'envisager aussi dans le cadre d'un cours en ligne mais cela me semble plus difficile à mettre en place.

La grille d'évaluation peut être accompagnée de "devoirs exemplaires" (DE). C'est une des recommandations faites par Coursera [2]. Ces devoirs ne sont pas forcément exemplaires dans le sens où ils correspondent pas tous à une production excellente, mais pour chaque critère de la grille d'évaluation on peut les utiliser comme illustration. Par exemple : pour le critère "présentation du dessert", le DE n° 1 obtient la note C car le gâteau est bien découpé mais que l'assiette ne contient que des ingrédients bruns ; le DE n° 2 obtient la note A car on découvre une fumée parfumée lorsqu'on soulève le dôme en chocolat. Ces exemples fournissent de précieux repères aux évaluateurs. Les exemples peuvent parfois être piochés directement dans les productions des "promotions" précédentes (en les anonymisant). Voir aussi la section 1.4.2.

1.3 Réalisation des évaluations

1.3.1 Attribution des copies / sélection des pairs évaluateurs

La distribution des copies peut se faire de manière aléatoire, semi-aléatoire ou bien supervisée.

La distribution semi-aléatoire peut tenir compte de contraintes telles que :

- Dans le cas d'un travail réalisé et évalué en groupes (les groupes peuvent éventuellement être remaniés entre la composition du devoir et son évaluation), un élève ne doit pas évaluer le travail de quelqu'un de son groupe de production.
- Dans le cas où on dispose pour chaque élève d'une note d'évaluateur sur un devoir précédent, deux élèves ayant été particulièrement sévères par le passé ne doivent pas évaluer le même devoir (sinon on aura plus de mal à repérer que cette production est peut-être sous-notée).

Dans le cas d'une ventilation supervisée, le professeur attribue les copies selon des critères liés à sa connaissance des étudiants (en quantité d'évaluateur et d'évalué). Par exemple :

- [9] met en garde (et c'est compréhensible) que l'amitié ou la simple camaraderie peuvent troubler la fiabilité des évaluations. Si le processus ne s'effectue pas en double-aveugle le professeur peut alors utiliser les éventuelles informations qu'il a sur les affinités entre élèves pour répartir les productions. Cela ne me semble pas applicable dans le contexte de l'enseignement supérieur (et encore moins des MOOC).
- Si l'enseignant a connaissance de compétences particulières de certains élèves, et des particularités de certaines copies, il peut appairer en conséquence. Par exemple, si un devoir est rédigé en anglais et qu'il y a un étudiant bilingue dans la promotion, on peut lui attribuer la copie.

1.3.2 Période d'évaluation

Dans tous les cas rencontrés dans la bibliographie les étudiants doivent d'abord rendre leur propre devoir avant de pouvoir évaluer leurs pairs. En outre il y a toujours une date limite de rendu des devoirs et une date limite de rendu des évaluations.

Ensuite on distingue plusieurs fonctionnements : dans notre MOOC, dès qu'un étudiant a déposé sa production il peut commencer à évaluer ses pairs si des productions sont en attente de notation (nous appelons ce fonctionnement où il n'y pas de date de début d'évaluation "évaluation à la volée"). Cela permet aux étudiants de mieux gérer leur emploi du temps, d'obtenir éventuellement un retour plus rapidement. De plus, un étudiant peut soumettre successivement plusieurs devoirs ; cela lui permet d'améliorer sa production sur la base des évaluations de ses propositions précédentes mais cela peut épuiser le "stock" d'évaluateurs (si dans une promotion de 30 élèves on exige de chacun 3 évaluations, qu'un élève soumet 10 devoirs avant ses camarades, les derniers élèves à soumettre ne trouveront plus d'évaluateurs n'ayant pas déjà fait 3 évaluations...).

Dans notre cours électif, nous adoptons un fonctionnement synchrone : les devoirs sont tous remis à la même date et les évaluations commencent alors. Ce choix a été fait pour faciliter la gestion (ventilation manuelle, partie "papier" pour le devoir...).

1.3.3 Nombre d'évaluations par production

Chaque étudiant peut se voir confier une ou plusieurs copies (et donc, dans la plupart des situations, chaque étudiant verra son travail évalué par une ou plusieurs personnes). La valeur par défaut du nombre d'évaluations par production est fixée à 3 sur la plate-forme Coursera. Pour le MOOC "développement durable" [4] a choisi de demander 4 évaluations. Dans le MOOC HCI de Stanford sur Coursera, chaque évaluateur doit faire 5 évaluations dont une d'étalonnage (voir plus loin). Ces choix ne sont pas argumentés.

Le nombre d'évaluations peut-être fixé selon plusieurs critères :

- de combien de temps dispose l'évaluateur pour rendre son évaluation, et combien de temps estime-t-on qu'une évaluation doit prendre ? [8] vérifie l'hypothèse assez intuitive que si l'évaluateur n'a pas assez de temps (ou ne prend pas assez de temps), son évaluation ne sera pas fiable (la note qu'il attribuera sera à plus d'un écart type de la note étalon, et généralement au dessus).
- combien d'évaluations semblent suffisantes pour faire des statistiques permettant d'identifier les évaluateurs trop ou trop peu exigeants et de "neutraliser" leurs effets ?
- quelle variété est attendue dans les productions et donc quel intérêt peut trouver l'évaluateur à en examiner plusieurs ? Par exemple, dans notre MOOC Android une des évaluations consiste à examiner une idée d'application que l'étudiant souhaite développer. Il y a une grande diversité dans les réponses, de l'application ludique à celle permettant de gérer la distribution de médicaments et l'évaluateur peut être intéressé et apprendre de cette diversité. Une autre évaluation consiste à vérifier que l'apprenant évalué a réussi à installer tout le logiciel nécessaire et à re-

produire un code simple. L'objectif est de faire une simple vérification avant de passer à des codes plus complexes et également de vérifier que les apprenants ont bien compris comment soumettre leurs productions composées d'une multitude de fichiers car les évaluations suivantes (qui elles jugeront plus les compétences acquises en développement Android) nécessiteront le même genre de dépôts sur la plate-forme. L'exercice ne présente donc quasiment aucun intérêt pour l'évaluateur. Dans ce cas on peut plutôt viser la borne minimum pour le nombre d'évaluations.

Il est également envisageable de fixer un nombre initial d'évaluations par production relativement faible (3 par exemple) et de demander une 4ème ou même 5ème évaluation si l'écart-type constaté dans les 3 premières notes est jugé trop important.

1.3.4 Anonymat (des copies et des évaluations)

Pour éliminer le biais lié aux amitiés/rivalités entre élèves ou à l'effet Pygmalion par exemple, on peut choisir d'anonymiser les copies. En revanche si l'exercice consiste à évaluer la contribution des co-équipiers à un travail de groupe, l'anonymat doit être levé.

Certains évaluateurs peuvent se censurer s'ils savent que leur identité sera connue des évalués et noter plus généreusement, on peut donc être tenté de ne pas révéler aux évalués l'identité des évaluateurs. D'autres évaluateurs au contraire risquent de profiter de l'anonymat pour faire des commentaires agressifs par exemple, à l'image des comportements qu'on observe parfois sur Internet... Une solution dans ce cas est de permettre aux évalués de signaler ("flag") les commentaires qui leur semblent injustifiés afin que l'enseignant les contrôle.

Les 2 premières années où j'ai utilisé l'évaluation par les pairs dans l'électif, je laissais les évaluateurs choisir de révéler ou non leur identité. Environ 1/3 des évaluateurs choisissait de rester anonyme ; j'ignore quelles étaient leurs motivations pour cela et je n'ai pas noté de différence de ton ou de note attribuée liée à ce choix.

1.4 Calcul de la note

L'évaluation par les pairs a le plus souvent pour objectif d'attribuer une note aux productions de chaque élève (en tous cas dans les MOOC). Comment peut-on calculer cette note sur la base des évaluations rendues ? Plusieurs possibilités sont envisageables. L'article [9] s'appuie sur des données récoltées dans des collèges et indique que dans des conditions favorables (barème soigné, évaluateurs accompagnés), on observe une très grande corrélation entre les notes attribuées par les élèves et celles attribuées par le professeur. L'article [8] indique au contraire que dans un MOOC (HCI de Stanford) 43% des productions se sont vues attribuer une note s'éloignant de plus de 10% de la note attribuée par un professeur, cet écart pouvant parfois aller jusqu'à 70% ! Si l'objectif prioritaire est la notation des productions il est donc critique de mettre au point une méthode permettant de limiter cet écart (en plus de mettre en place un accompagnement des évaluateurs pour tenter de limiter cet écart à la source plutôt qu'a posteriori). [8] s'appuie sur l'analyse d'une grande base d'évaluation par les pairs pour proposer quelques modèles de notes ajustées. Il sera intéressant

de creuser cette problématique pour proposer un mode de calcul de la note finale qui soit à la fois le plus adapté à fournir une note fiable, et à la fois compréhensible par tous les enseignants (et étudiants).

1.4.1 Fonctions simples

La méthode la plus commune consiste à faire la moyenne des notes, éventuellement en retirant les extrêmes, ou d'utiliser la note médiane.

Les notes peuvent être pondérées par leur écart à une note de référence. Cette référence peut être la médiane par exemple. Après avoir simulé différentes manière de pondérer les notes [6] conclut que la médiane est finalement la meilleure solution.

La note de référence peut également être la note attribuée par l'enseignant, lorsqu'il y en a une. Dans ce cas, cette note peut devenir la note finale, écrasant les évaluations par les pairs, ou devenir la nouvelle référence pondérant les notes attribuées par les pairs.

Note : [1] distingue la "*validité*" (mesure la similarité des notes attribuées par les pairs et par le professeur) et la "*fiabilité*" (dans laquelle la note attribuée par le professeur n'intervient pas).

1.4.2 Estimation et intégration d'un biais de l'évaluateur

Utilisation d'un devoir étalon

On peut demander à chaque évaluateur de corriger un devoir type déjà corrigé par un professeur et établir, en comparant la note attribuée par l'enseignant et celle de l'évaluateur, le biais de l'évaluateur (c'est-à-dire sa tendance à noter de manière trop généreuse ou trop sévère). Les autres notes fournies par l'étudiant sont alors pondérées par son biais.

Cet "étalonnage" peut se faire à l'insu de l'évaluateur (il ne sait pas qu'il corrige un devoir déjà noté par un enseignant) ou bien constituer un préalable aux "vraies" évaluations : l'outil [10] en fait un pré-requis contrôlé : avant de pouvoir évaluer le devoir d'un pair, il faut commencer par noter un *"devoir étalon"* et si les notes proposées ne sont pas assez proches de celles du professeur pour les différents critères considérés il est impossible de passer à l'étape suivante et il faut recommencer avec un autre "devoir étalon" (dans le MOOC [6] qui utilise cet outil les étudiants peuvent malgré tout passer à l'étape suivante si après le 5ème entraînement leurs notes diffèrent encore trop de celles du professeur).

[8] propose une variante de la solution précédente qui présente l'avantage de réduire la charge de travail du professeur : il s'agit choisir quelques productions parmi celles rendues qui serviront d'étalon : chaque évaluateur reçoit 5 productions à noter, dont une "étalon" (mais il l'ignore). Les productions d'étalonnage sont donc corrigées par un grand nombre d'évaluateurs. On calcule la moyenne reçue pour une production d'étalonnage donnée et ensuite on calcule le biais de chaque étudiant par rapport à cette moyenne. L'article indique que la note moyenne obtenue pour une production d'étalonnage est même plus fidèle au barème qu'une note attribuée par un professeur. A mon avis, cela révèle plutôt un défaut dans la conception de la grille d'évaluation ou bien un biais de l'évaluation par les pairs : la note du professeur devrait rester celle de

référence. L'article cite d'ailleurs ce point parmi ceux qui devront faire l'objet de plus d'étude.

Utilisation de l'historique de notation

Si le module de formation comporte plusieurs évaluations par les pairs successives, on peut également s'appuyer sur les notes attribuées précédemment par un étudiant pour estimer son biais puisque l'analyse réalisée dans [8] montre une cohérence temporelle dans ce biais [2].

Mais le biais initial doit être malgré tout réévalué car [6] indique de plus un évaluateur réalise d'évaluations, plus elles deviennent fiables (au sens "proche de la note du professeur"). Il faut souligner que dans ce MOOC les évaluateurs reçoivent après chaque évaluation une indication de comment ils se situent par rapport à la médiane des notes attribuées par les autres évaluateurs (au dessus, au même niveau, en dessus).

Utilisation du niveau de compétence de l'évaluateur

Bien que cela interroge sur la légitimité de l'évaluation par les pairs, il est intuitif de penser que le jugement d'une personne compétente sera plus fiable que celui d'une personne qui l'est moins. Ainsi, on peut affecter d'un poids plus important les notes attribuées par les étudiants qui obtiennent les meilleurs résultats. Cette idée est un des modèles proposés par [8]. Cet article montre que le biais des évaluateurs est moins important lorsqu'ils notent des productions du même niveau que les leurs : les très bons étudiants ont tendance à se montrer trop sévères avec les moins bonnes productions tandis que les moins bon étudiants ont tendance à surnoter les très bonnes productions. [9] rapporte une analyse différente : les élèves attribuent généralement de moins bonnes notes que le professeurs aux très bonnes productions.

Utilisation de la nationalité

[6] rapporte que les étudiants attribuent des notes en moyenne supérieures de 3% aux étudiants de même nationalité qu'eux ; l'évaluation étant réalisée en double aveugle, cette surnotation n'est pas volontaire. On peut l'attribuer au fait que les évaluateurs comprennent mieux le travail des étudiants partageant la même culture qu'eux, ou à une compréhension identique des consignes.

1.5 Contrôle par l'enseignant

Le professeur peut contrôler certains devoirs et évaluations de manière systématique (l'enseignant examine toutes les évaluations produites), de manière aléatoire, ou de manière dirigée. Parmi les indices pouvant orienter l'enseignant vers certaines copies à re-corriger, on peut par exemple imaginer :

2. L'article note qu'une cohérence temporelle se retrouve aussi, statistiquement, dans les notes d'une personne mais que cela ne peut éthiquement pas être utilisé pour ajuster la note d'une production future car cela reviendrait à nier les capacités de l'étudiant à progresser et son droit à un "nouveau départ" pour chaque devoir.

- Les moins bons étudiants ont tendance à être moins fiables dans leurs évaluations, ce sont les copies évaluées par ces étudiants que le professeur examinera en priorité.
- Lorsqu'une auto-évaluation complète l'évaluation par les pairs, l'enseignant peut choisir les devoirs pour lesquels l'écart entre la note auto-attribuée s'éloigne beaucoup de la médiane de celles attribuées par les pairs.
- Lorsque les évalués ont la possibilité de signaler que leur évaluation leur semble injuste, l'enseignant de penchera sur ces copies.
- Les copies obtenant des notes très différentes d'un évaluateur à un autre peuvent attirer l'attention de l'enseignant.

1.6 Méta-évaluation, note d'évaluateur

Chaque évaluateur peut recevoir une note d'évaluation, reflétant sa capacité à estimer avec justesse la qualité des travaux soumis à son évaluation. Cette note peut être obtenue en mesurant de combien l'évaluateur s'écarte des notes attribuées par ses pairs sur un devoir.

La méta-évaluation peut être plus délicate à faire pour les champs de commentaires libres. [8] propose une méthode basée sur la longueur du commentaire et une analyse à partir d'une liste de mots permettant de distinguer les commentaires positifs des critiques négatives. Cette technique peut peut-être être étendue ? Une autre solution est d'utiliser l'éventuel feedback fourni par l'évalué signalant qu'il s'estime mal jugé.

1.7 Autres considérations

1.7.1 Transparence

Comment présenter à un élève sa note obtenue par une manipulation statistique des évaluations reçues et éventuellement des évaluations données ? A-t-il besoin/envie de connaître le mécanisme, est-il capable de le comprendre ?

1.7.2 Mécanismes incitatifs

Pour encourager les évaluateurs à remplir sérieusement leur mission je n'ai pas trouvé beaucoup de mécanismes incitatifs. Il s'agit majoritairement de retirer des points à un évaluateur dont la note s'éloigne trop de la moyenne des notes attribuées pour les mêmes productions (ou attribuer un bonus dans le cas contraire).

[6] mentionne qu'une étudiante a réalisé beaucoup d'évaluations et s'est montrée très active dans les forums, ce qui lui a permis d'acquérir l'estime de ses pairs. La reconnaissance et le sentiment d'utilité peuvent être des leviers incitant certains étudiants à faire sérieusement les évaluations qui leur sont confiées.

Si le professeur corrige toutes les copies et que l'évaluation par les pairs n'a pas pour objectif de noter une production, [3] estime que les étudiants peuvent se sentir dévalorisés et se démotiver.

Présenter clairement les bénéfices d'apprentissage qu'un évaluateur peut retirer de cet exercice peut également faciliter l'adhésion.

1.7.3 Aspects légaux

Est-il légal de conditionner l'obtention d'un diplôme à une note partiellement issue d'évaluations par les pairs ? Aux Etats-Unis, où la question a été portée devant les tribunaux, la Cour Suprême a répondu par l'affirmative en 2001.

1.7.4 Auto-évaluation

Dans [6], la note finale prend en compte l'auto-évaluation : si la médiane des notes attribuées par les pairs et la note d'auto-évaluation présentent moins de 5% d'écart, la plus haute des deux est utilisée comme note finale (sinon c'est la médiane des pairs qui est utilisée).

[9] cite des études concluant que les filles ont tendance à se sous-noter lors d'auto-évaluations (mais d'autres études arrivent à la conclusion que la fausse-modestie – pour éviter d'être perçu comme fier ou orgueilleux par les pairs – est également répandue chez les filles et les garçons). En vieillissant les élèves perdraient la tendance qu'auraient les plus jeunes à se sur-noter. L'âge et le genre pourraient donc être des facteurs de biais dans l'auto-évaluation mais je n'ai trouvé aucune étude qui étende cette conclusion à l'évaluation des pairs.

Chapitre 2

Les bénéfices attendus

[9] discute en profondeur la pertinence des choix d'outils statistiques utilisés dans diverses études sur l'intérêt de l'évaluation par les pairs. Parmi ces articles nombreux sont ceux rédigés par des auteurs n'ayant que des compétences limitées en statistiques. Cela contribue certainement à expliquer pourquoi les conclusions des différentes études ne convergent pas toujours : pour chaque proposition exhibant des chiffres en sa faveur dans un article on trouve des conclusions différentes, ou en tous cas des chiffres bien différents, dans un autre article... Une autre raison des écarts rapportés tient certainement de la variété des contextes : une évaluation par les pairs en présentiel dans des classes de collèges présente des caractéristiques différentes d'une évaluation sur un MOOC de Stanford. Il est donc important de faire des expérimentations au niveau de l'Université de Paris-Saclay pour déterminer la ou les formules qui seront le plus adaptées à notre contexte.

2.1 Pour le formateur

2.1.1 Traiter de gros volumes

Partant du constat que les tests les plus riches sont aussi ceux qui prennent le plus de temps à corriger, les professeurs préfèrent parfois recourir à des outils moins optimaux mais plus faciles à noter. L'évaluation par les pairs permet un passage à l'échelle : plus il y a d'apprenants, plus il y a d'évaluateurs. Cela permet :

- de travailler avec de grands groupes
- de faire des évaluations plus fréquentes
- de fournir un retour plus rapidement
- de fournir un feedback plus riche

2.1.2 Évaluer la compétence à critiquer

Avec le développement de "l'approche par compétences", l'évaluation par les pairs pourrait fournir au professeur la possibilité d'évaluer, justement, une compétence des apprenants rarement testée sans cette modalité : leur capacité à évaluer une solution (et pas seulement à en formuler une). C'est un outil de

test de l'esprit critique des étudiants, ainsi que de leur capacité à formuler un feedback de manière constructive.

2.1.3 Estimer l'engagement / la motivation

[8] montre qu'analyser les évaluations d'un élève permet d'obtenir une mesure de son engagement : ceux qui notent le plus fiablement sont les plus susceptibles de poursuivre le cours. Cela permet au professeur d'identifier les élèves en risque de décrochage et de les re-mobiliser.

2.1.4 Enrichir son expérience, changer de perspective

Mettre en place une évaluation par les pairs dans son cours permet à un professeur d'enrichir son expérience pédagogique. Cela peut être une motivation en soit.

Passer de celui qui corrige à celui qui articule le guide de correction fait évoluer subtilement le rôle du professeur de "juge" à "coach". Une conséquence est que les étudiants ne considèrent plus que le prof "a ses têtes" ([6]). D'autre part s'obliger à expliciter les critères de réussite de manière très précise fait parfois prendre conscience d'améliorations qu'on peut apporter au cours.

2.2 Pour l'étudiant placé en position d'évaluateur

Dans le MOOC [6] la très large majorité des étudiants a déclaré qu'évaluer des productions des autres étudiants a été une expérience formatrice (30%) voire très formatrice (43%). 20% des étudiants ont même volontairement évalué plus de productions que ce qui était demandé. Mais ces chiffres ne reflètent qu'un sentiment et ne mesurent pas de progrès effectifs. [9] rapporte les résultats d'une expérience consistant à soumettre des élèves à un test puis de les partager en 3 groupes : un groupe réalisant une évaluation par les pairs de ce test, un groupe réalisant une auto-évaluation, et un groupe de test. Une semaine plus tard et de manière inopinée les étudiants sont soumis au même test. Les étudiants ayant notés leurs pairs n'obtiennent pas des résultats significativement meilleurs que ceux du groupe de contrôle (en revanche ceux ayant réalisé une auto-évaluation progressent largement). [10] avance que les étudiants utilisant l'outil CPR ont en moyenne des performances 10% supérieures au autres étudiants mais comme CPR permet de faire à la fois de l'auto-évaluation et de l'évaluation par les pairs il faudrait plus de détails sur l'obtention de ce chiffre de 10%. Ainsi il n'est pas clair que l'évaluation par les pairs profite réellement à l'évaluateur. Dans la suite nous imaginons pourtant quelques bénéfices.

2.2.1 Consolider ses connaissances

Les auteurs de [6] ne démontrent pas que les "fortune cookie feedbacks" améliorent le feedback mais ils y trouvent malgré tout l'intérêt pour l'évaluateur de réviser à nouveau le cours sous un angle un peu différent. Demander à des étudiants d'évaluer leurs pairs peut être considéré comme une activité pédagogique supplémentaire liée à une partie du cours. C'est une autre manière d'aborder le même problème, une autre manière de travailler la même matière, le même sujet.

Par exemple si l'exercice consiste à écrire un texte au passé simple, l'évaluateur va nécessairement réviser sa leçon de conjugaison pour décider si une phrase est correcte ou non.

De plus l'évaluation par les pairs porte le plus souvent sur des questions ouvertes dont la correction ne peut pas être automatisée en raison de la grande variété des réponses qui peuvent être proposées. Cela signifie que l'évaluateur va être confronté à différentes manières de résoudre un problème ou de présenter une solution. Un professeur n'a pas toujours le loisir de présenter plusieurs approches et l'évaluation par les pairs peut être l'occasion pour l'évaluateur d'en découvrir de nouvelles.

Lorsque l'évaluation se réalise en groupe plutôt qu'individuellement, des discussions peuvent apparaître entre pairs évaluateurs ayant des opinions différentes au sujet de la production à évaluer, ou au sujet de l'interprétation de la grille d'évaluation. Ces discussions sont certainement bénéfiques à l'appropriation des concepts mobilisés dans le devoir. Il est également possible qu'un des pairs évaluateurs soit plus compétent ou plus expérimenté que les autres sur un point donné et en fasse profiter le groupe à l'occasion de l'activité d'évaluation.

2.2.2 Développer son esprit critique et son auto-critique

Plusieurs articles mentionnent que de nombreuses situations de la vie professionnelle (on pourrait ajouter personnelle) font appel à notre capacité à évaluer les autres et à donner un feedback constructif. L'exercice d'évaluer ses pairs pourrait permettre de développer cette compétence. Rappelons que [6] indique qu'en indiquant à l'évaluateur comment il se situe par rapport aux autres ses évaluations ultérieures deviennent plus justes.

Évaluer ses pairs permet aussi de développer sa capacité à s'auto-évaluer, de prendre conscience de ses propres forces, progrès et manques. [7] cite un étudiant qui avait le sentiment d'être le seul à rencontrer des difficultés en classe et qu'évaluer ses pairs lui a permis de réaliser qu'il n'était pas un cas isolé.

2.2.3 Avoir un point de vue plus positif sur les tests et les enseignants

L'évaluation par les pairs pourrait permettre de modifier le regard que les élèves portent sur les professeurs et les évaluations. Comme je l'ai déjà mentionné, le rôle du professeur "glisse" de "juge" à "coach". Cela peut permettre de modifier le rapport entre l'élève et le professeur et constitue pour certains élèves des conditions d'apprentissage plus favorables.

[7] cite un élève pour qui évaluer ses pairs a été une révélation sur ce que doivent endurer les enseignants lorsqu'ils corrigent (copies peu soignées par exemple) et que constater l'effort que les enseignants mettent dans cette tâche l'a motivé à travailler davantage.

L'évaluation par les pairs permet aussi de donner un sens moins scolaire ou "sanction" aux examens, et un aspect plus constructif. Dans notre électif aucun élève n'est jamais venu consulter sa copie de questions de cours (corrigée par le professeur) pour savoir ce qui avait été réussi ou raté (mon expérience est que dans la quasi totalité des cas un étudiant qui vient consulter sa copie cherche juste à "gratter des points" et non pas à identifier sur quoi il doit progresser...).

En revanche plus de la moitié est venue consulter les remarques laissées par les pairs sur l'autre partie de l'examen.

2.2.4 Se rendre utile

Les évaluateurs peuvent se sentir valorisés de se voir confier ce rôle, et se sentir utiles et s'en réjouir.

[6] mentionne que des assistants d'enseignement peuvent être recrutés parmi les évaluateurs les plus actifs et fiables. Cet exercice peut donc être une manière pour l'évaluateur de se faire remarquer du professeur (pour obtenir une recommandation par exemple).

2.3 Pour l'étudiant évalué

[6], qui s'intéresse plus particulièrement à l'évaluation des productions dans des domaines dits "créatifs" où l'évaluation est présentée comme plus délicate, fait l'analogie avec "le studio", cet espace à l'école des Beaux-Arts partagé par plusieurs étudiants et dans lequel ils travaillent sur leurs projets. Dans cet espace, les étudiants peuvent observer le travail de leurs pairs, recevoir des commentaires de la part des professeurs ou des autres étudiants mais également entendre les commentaires adressés à leurs pairs. Cette disposition a un aspect pratique (plus simple que de fournir x petites salles bien éclairées, avec un point d'eau etc) mais a également des vertus pédagogiques : les étudiants ont beaucoup plus de feedback sur lequel se baser pour améliorer itérativement leurs productions.

L'évaluation par les pairs permet à un étudiant de recevoir davantage de feedback que si le professeur devait tout évaluer lui-même : avis plus nombreux par devoir, et fréquence des évaluations éventuellement plus élevée.

Dans le MOOC [6] la très large majorité des étudiants a déclaré que recevoir des évaluations des autres étudiants a été une expérience très formatrice.

D'autre part l'évaluation par les pairs peut catalyser le développement d'une communauté d'apprenants qui fournira des conditions favorables à l'apprentissage : les étudiants seront plus enclins à s'entraider, ils se respecteront davantage...

Un autre avantage de l'évaluation par les pairs est que dans de nombreux cas les critères de réussite sont bien mieux explicités que dans le cas d'un devoir évalué par le professeur seul.

Chapitre 3

Les freins, les craintes, les limites, les résistances

3.1 Du point de vue technique

[6] mentionne un problème de l'évaluation par les pairs que nous avons eu l'occasion de constater dans notre MOOC : pour fonctionner, il faut qu'un nombre suffisant de pairs aillent au même rythme. Sans ça il y a un risque de devoir attendre longtemps l'évaluation de son devoir, ou bien de n'avoir aucun devoir à corriger.

3.2 Du point de vue de l'enseignant

Voici une liste d'objections, certainement non exhaustive, qu'un professeur peut opposer à l'évaluation par les pairs :

- C'est mon travail, c'est moi qui suis payé pour évaluer, on va me traiter de feignant si je le fais faire par les étudiants.
- C'est mon travail, c'est moi qui suis payé pour évaluer, c'est à moi de le faire.
- Je ne peux pas me fier aux notes attribuées par les étudiants. Ces craintes sont-elles fondées ? Une des principales conclusions que l'on peut tirer de la bibliographie est les résultats des études sur la fiabilité des notes attribuées par les pairs sont contradictoires (comme le note par exemple [1] qui s'intéresse particulièrement à cette question). L'étude [8] conforte les professeurs opposés à donner une vote via ce biais en indiquant que dans le MOOC considéré 43% des productions se sont vues attribuer une note s'éloignant de plus de 10% de la note attribuée par un professeur). Les raisons qui peuvent faire craindre un manque de fiabilité sont multiples :
  — les étudiants ne sont pas compétents pour s'évaluer mutuellement
  — les étudiants ne vont pas faire cette évaluation sérieusement
  — les étudiants vont se mettre d'accord entre eux
- Les étudiants ne tireront rien de cet exercice.
- Cela va me demander plus de travail (création des barèmes détaillés, des devoirs étalon, méta-évaluation, anonymisation...)

- La manière dont l'atelier Moodle calcule la note finale est obscure ([7]).

3.3 Du point de vue de l'évaluateur

Lorsque les évaluations ne sont pas anonymisées, certains étudiants peuvent craindre des formes de "représailles".

L'évaluation peut être perçue comme non utile à l'apprentissage et donc comme une perte de temps. La gestion de plusieurs dates limite est aussi mentionnée comme une difficulté ([7]).

[7] cite un étudiant qui dit qu'aider un pair en l'évaluant ne lui pose pas de problème qu'à la condition qu'il ait le sentiment que le dit pair a mis tous ses efforts dans son devoir.

Certains étudiants peuvent se montrer réticents à endosser une responsabilité traditionnellement réservée au professeur, soit par esprit d'opposition soit car ils ne se sentent pas légitimes.

3.4 Du point de vue de l'évalué

L'élève évalué peut ressentir de la frustration de ne pas avoir un retour du professeur si l'avis des autres apprenants est perçu comme ayant moins de valeur que l'avis d'un professeur. Les évalués peuvent ne pas avoir confiance dans l'évaluation, et craindre qu'elle ne soit pas faite avec sérieux : [6] indique que plus du quart des étudiants ont eu le sentiment que leurs pairs avaient mis moins d'efforts à les évaluer que ce qu'ils en avaient mis eux-mêmes et [7] rapporte que les étudiants accordent plus de valeur aux évaluations qu'ils ont fournies qu'à celles qu'ils ont reçues (71% contre 44%).

[7] cite des étudiants désemparés devant des évaluations très différentes reçues pour leur devoir.

Lorsque les productions ne sont pas anonymisées, un évalué peut craindre les moqueries des évaluateurs (c'est d'ailleurs ce qui a conduit une mère à porter plainte contre une école utilisant l'évaluation par les pairs (cas Falvo v. Owasso School en 2001)).

Lorsqu'un classement est en jeu, l'évalué peut craindre que les évaluateurs le notent volontairement mal afin d'"éliminer la concurrence".

[5] souligne que la culture nationale, mais également la culture de l'organisation, peut conduire à des attitudes différentes envers l'évaluation et la critique.

L'introduction de [7] mentionne le problème de la propriété intellectuelle. Et en effet un étudiant peut faire confiance à son enseignant pour ne pas divulguer ses idées de scénario par exemple mais être plus réticent à les confier à un pair avec qui il est en compétition pour réaliser le port-folio qui lui permettra de décrocher le stage de ses rêves avec un metteur en scène réputé...

Chapitre 4

Les outils

4.1 Intérêts d'une solution en ligne

L'utilisation d'un outil informatique pour l'évaluation par les pairs présente plusieurs avantages (selon les possibilités de la plate-forme choisie) :

- collecte automatique des devoirs
- relance des étudiants en retard
- ventilation des copies
- gestion de l'anonymat
- collecte des évaluations
- calcul de la note finale
- identification des copies litigieuses à faire vérifier par un enseignant
- consultation aisée par les évalués de leurs évaluations
- ...

[7] met cependant en garde sur les risques de s'appuyer sur un outil en ligne et suggère d'avoir un plan B en réserve si le serveur tombe en panne au plus mauvais moment...

4.2 Inventaire des outils disponibles

L'étude des différents outils disponibles sur le marché fera l'objet d'un document séparé. La liste suivante n'est pas exhaustive :

- fonction "atelier" de Moodle. [7] et [4] utilisent cette fonction. L'objectif de [7] est de discuter l'utilisation de cet outil. Essayé à CentraleSupélec.
- Web PA (`https://github.com/WebPA`), essayé à CentraleSupélec.
- MEC, le module d'évaluation des co-équipiers, plugging Moodle développé par Eric Francoeur, une présentation est disponible à cette adresse : `https://docs.google.com/presentation/d/1K2tKNP7XrVjl396584-ubWPHqqU0nK932aLUUBX0efU/edit\#slide=id.p`
- iPeer : `http://itlal.org/index.php?q=node/283`
- outil Coursera
- outil PeerMark de la plate-forme Turnitin, utilisé par exemple à UCLA ; voir `http://my.ucla.edu/turnitin/PeerMark_manual.pdf` pour le manuel et `https://dspace.lboro.ac.uk/dspace-jspui/handle/2134/4559` pour un article décrivant un cas d'usage.

- Aropä (http://www.dcs.gla.ac.uk/~hcp/aropa/index.html), dont l'usage est discuté dans un article de 2007 d'Hamer, Kell et Spence (http://dl.acm.org/citation.cfm?id=1273678)
- Computerized Assessment with Plagiarism / Computerised Assessment by Peers, décrit sur la page de son auteur Davies : https://at-web1.comp.glam.ac.uk/staff/pdavies/caa.htm
- Calibrated Peer Review : http://cpr.molsci.ucla.edu/Overview.aspx
- PASS or Peer Assessment Support System : http://www.unm.edu/~ehk1/pdf/Peer%20Assessment%20of%20WebQuests.pdf
- PeerScholar.com : http://peerscholar.com/

[11] dresse un inventaire comparatif récent des outils disponibles.

Bilan

Les conclusions de cette étude préalable :

- Le terme d'évaluation par les pairs recouvre en fait des pratiques et des objectifs (et des résultats) relativement divers. Les enseignants ne pensent pas forcément à inclure ce mode pédagogique dans leur boîte à outils, ou en ont une idée préconçue, il peut être bon de les informer davantage.
- Dans le cas où l'évaluation par les pairs a pour objectif d'attribuer une note, les études sur la fiabilité présentent des résultats contradictoires. En se limitant au périmètre de Paris Saclay, il sera peut-être plus facile d'identifier les facteurs influençant cette fiabilité. D'autre part il existe de nombreuses manières de calculer la note finale. Il faudrait les étudier davantage, avec un regard d'expert en statistiques.
- La conception de l'activité d'évaluation par les pairs (rédaction des consignes des devoirs, de la grille de correction, mais aussi tous les réglages : anonyme ou pas, évaluation à la volée ou séquentielle etc) doit être réalisée avec beaucoup de soin pour maximiser l'intérêt de l'exercice. Les professeurs gagneraient vraisemblablement à disposer de l'aide de conseillers pédagogiques sur ce point. Par exemple, dans les classes où plusieurs évaluations par les pairs sont prévues, [9] conseille de mettre en place l'évaluation par les pairs de manière progressive : dans un premier temps les évaluateurs se contentent de commenter certains aspects des productions, sans tout juger et sans noter.
- L'exercice est souvent nouveau et déroutant pour les élèves. Eux aussi ont besoin d'accompagnement (explication sur l'intérêt, évaluations-étalons, feedback sur leurs évaluations...).
- Disposer d'un outil pour dématérialiser le processus est essentiel car le nombre de documents à collecter et croiser ne croît pas linéairement avec la taille de la classe. L'atelier de Moodle et WepPA ne couvrent pas tous nos besoins.

L'évaluation par les pairs est utilisée dans d'autres contextes que celui sur lequel ce document se concentre. Dans le futur, on pourra étudier l'extension de notre étude et de notre plate-forme à ces domaines :

- l'évaluation des co-équipiers
- en médecine, pour atteindre ou maintenir l'excellence de la pratique et apprendre via l'observation des autres
- en travail social, car les "clients" acceptent parfois mieux les conseils de leurs pairs que ceux des professionnels
- comptabilité, pour vérifier la conformité
- dans le monde de la recherche, l'évaluation par les pairs est massivement répandue pour décider du sort d'un article proposé à publication dans

une revue ou dans une conférence. Cette pratique est largement débattue depuis des années mais existe toujours. L'évaluation par les pairs est également utilisée pour constituer des jurys de recrutement, ou encore des comités de projets décidant des aides financières qui seront attribuées (ou non) à différentes propositions.

Références

Remarques :

- [6] propose une bibliographie très fournie mais pas très récente.
- Beaucoup d'études portent sur des cours de "creative writing", l'évaluation par les pairs semble avoir été moins utilisée dans des cours de science. Une explication effleurée par [7] est que des QCM peuvent permettre de vérifier qu'un étudiant sait appliquer une formule par exemple mais que seule une rédaction (dont la correction n'est pas automatisable) peut permettre de vérifier qu'un étudiant sait rédiger un poème romantique. [1] montre que l'évaluation par les pairs est très fiable dans des cours d'informatique ou d'électronique.
- Je n'ai pas (pas encore ?) trouvé d'étude de l'utilisation en groupe au sens "évaluer la production d'un groupe" ou "évaluer de manière collégiale une production" à part [4].

Bibliographie

[1] L'hadi Bouzidi and Alain Jaillet. Can online peer assessment be trusted ? *Educational Technology and Society*, 12(4) :257–268, 2009.

[2] Coursera Partner Help Center. Optimizing assessments. https ://partner.coursera.help/hc/en-us/articles/203597939-Optimizing-Assessments.

[3] Matthieu Cisel, August 2013. http ://blog.educpros.fr/matthieu-cisel/2013/08/08/mooc-comment-concevoir-une-evaluation-par-les-pairs/.

[4] Pascal Da Costa and Florence Labord. Mooc développement durable centralesupélec : Pourquoi l'évaluation par les pairs ? pour quels désordres et quelles opportunités ? 2016.

[5] Maria Gutknecht-Gmeiner. Peer review in education. Technical report, öibf – Österreichisches Institut für Berufsbildungsforschung, 2005.

[6] Chinmay Kulkarni, Koh Pang Wei, Huy Le, Daniel Chia, Kathryn Papadopoulos, Justin Cheng, Daphne Koller, and Scott R. Klemmer. Peer and self assessment in massive online classes. *ACM Trans. Comput.-Hum. Interact.*, 20(6) :33 :1–33 :31, December 2013.

[7] Markus Mostert and Jen D. Snowball. Where angels fear to tread : online peer-assessment in a large first-year class. In H. Blackey, L. Habib, A. Jefferies, and M. Johnson, editors, *The 17th Association for Learning Technology Conference (ALT-C 2010)*, September 2010.

[8] Chris Piech, Jonathan Huang, and Zhenghao Chen. Tuned models of peer assessment in moocs. In *Proceedings of the 6th International Conference on Educational Data Mining.* http ://web.stanford.edu/ cpiech/bio/papers/tuningPeerGrading.pdf.

[9] Philip M. Sadler and Eddie Good. The impact of self- and peer-grading on student learning. In *Educational Assessment*, January 2006. https ://www.cfa.harvard.edu/sed/staff/Sadler/articles/Sadler

[10] Los Angeles University of California. Calibrated peer review : A writing and critical-thinking instructional tool. EDUCAUSE Learning Initiative, September 2005. https ://net.educause.edu/ir/library/pdf/eli5002.pdf.

[11] Usman Wahid, Mohamed Amine Chatti, and Ulrik Schroeder. The State of Peer Assessment : Dimensions and Future Challenges. *International Journal on Advances in Systems and Measurements*, 9(3 and 4), 2016. http ://www.iariajournals.org/systems_and_measurements/.