AlphaPlus

Approches statistiques sur la détermination des seuils d'attribution des notes finales

Résumé

Octobre 2022

Résumé

Contexte

Le Baccalauréat International (IB) est une fondation internationale à but non lucratif majeure proposant un ensemble de programmes éducatifs aux élèves de 3 à 19 ans, qui peuvent être envisagés à la place des programmes « nationaux » et possèdent leur propre programme d'études et leurs propres évaluations. Pour assurer le bon fonctionnement de ses programmes, l'IB doit notamment déterminer et maintenir les normes des évaluations afin de garantir leur équité et leur comparabilité d'année en année.

À propos de cette étude

Cette étude s'intéresse à l'un des aspects de la manière dont l'IB convertit les notes des évaluations en notes finales et maintient les normes d'une année à l'autre. Plus spécifiquement, elle se concentre sur les différentes approches possibles lors de la création des seuils recommandés statistiquement pour le Programme du diplôme, le Programme à orientation professionnelle (POP) et le Programme d'éducation intermédiaire (PEI), leurs points forts et points faibles, et les situations dans lesquelles une approche est préférable à une autre. Pour les besoins de cette étude, les seuils recommandés statistiquement sont définis comme les estimations pour les seuils d'attribution des notes finales (le nombre minimum de points à partir duquel chaque note finale est attribuée), sur la base d'une comparaison statistique du niveau général de la cohorte par rapport à celle de l'année précédente (IBO, 2018).

À cette fin, l'étude s'attache à :

- réaliser une revue de la littérature actuelle et majeure concernant les différentes approches utilisées dans le calcul des seuils recommandés statistiquement, en examinant leurs points forts et points faibles, ainsi que les cas pour lesquels elles sont les plus pertinentes. Par ailleurs, la revue présente également d'autres approches statistiques visant à déterminer les seuils lorsque les seuils recommandés statistiquement ne sont pas appropriés, par exemple en l'absence d'une cohorte évaluée précédemment ;

- recenser différents contextes concrets de détermination des seuils d'attribution des notes finales, pertinents pour l'IB, en se basant sur des caractéristiques comme la taille de la cohorte, le taux de croissance, ou d'autres caractéristiques des matières individuelles ;

- analyser la précision et le bien-fondé de ces différentes approches de calcul des seuils recommandés statistiquement dans différents contextes jugés pertinents pour l'IB.

Revue de littérature

La revue de littérature constituait la première étape de ce projet, qui vise à examiner les procédures de l'IB en matière de détermination des seuils recommandés statistiquement dans le but de les améliorer. Dans l'idéal, grâce à ce projet, les seuils recommandés statistiquement fourniront une estimation plus précise des seuils d'attribution des notes finales afin qu'il ne soit pratiquement plus nécessaire de les ajuster (ou tout du moins pour que les modifications requises soient moindres que celles effectuées actuellement).

Compte tenu de ce qui précède, la revue de littérature s'emploie à atteindre les objectifs qui suivent.

1. Dresser une cartographie des procédures de détermination des normes statistiques, notamment :

   a. l'ensemble des conditions requises à leur utilisation ;
   b. leurs avantages et inconvénients par rapport aux autres approches.
2. Se faire une première opinion sur les procédures qui semblent les plus ou les moins adaptées aux contextes de l'IB.

Une littérature abondante a été revue dans le but de rassembler des informations sur les méthodes de détermination des normes statistiques utilisées. Elles appartiennent généralement à l'une des deux catégories suivantes : mise en équivalence et prédiction. Nous résumons ici les techniques relevant de chaque catégorie, qui comprennent notamment :

- mise en équivalence :
  - techniques basiques de mise en équivalence (moyenne, linéaire, percentiles égaux, etc.),
  - techniques de lissage,
  - designs avec des groupes non équivalents,
  - théorie de réponse à l'item ;
- prédiction :
  - façons de calculer une prédiction,
  - indicateurs externes de différences entre les cohortes ;
- manières de combiner plusieurs approches.

Sur la base de ces données, nous tirons de premières conclusions sur les approches de détermination des normes qui pourraient être utilisées (ou non) dans les contextes de l'IB. Les approches fondées sur les résultats obtenus précédemment ne semblent pas réalisables en raison d'un manque d'informations. Il ne parait pas non plus possible d'appliquer les designs de groupes non équivalents, car les éléments d'ancrage compromettraient la sécurité des évaluations de l'IB.

Pour généraliser, nous pouvons donc dire que trois approches globales sont en apparence prometteuses pour l'IB.

a. Techniques basiques de mise en équivalence
b. Approches fondées sur les résultats obtenus en simultané
c. Approches cherchant à maintenir les résultats obtenus précédemment (centres communs)

Les techniques basiques de mise en équivalence, dans leur ensemble, sont pertinentes lorsque les deux cohortes possèdent des aptitudes comparables. Cependant, ce n'est le cas que dans la moitié environ des contextes de l'IB. Il est important de mentionner que les techniques basiques de mise en équivalence **peuvent** être appliquées dans pratiquement toutes les situations (elles ne nécessitent qu'un petit échantillon) et que, par conséquent, elles peuvent représenter la seule option envisageable dans certains cas. La question est de savoir si cela est opportun (notamment lorsque les cohortes sont susceptibles d'être différentes) ou s'il est préférable de se fier à des approches reposant sur le jugement.

Les approches de mise en équivalence simultanées, comme la méthode « Instant summary of achievement without grades » (ISAWG) développée par Benton (2017), sont performantes et conviennent aux programmes de l'IB du fait qu'elles peuvent inclure un large éventail de matières. De plus, cette méthode constitue (de loin) l'approche de mise en équivalence la plus convaincante pour les contextes les plus délicats, notamment pour les matières choisies par peu d'élèves, lorsque le niveau de la cohorte change complètement et pour les toutes nouvelles matières. Cependant, les approches ISAWG sont extrêmement complexes et comptent un nombre très important d'options et de modifications (en comparaison notamment avec les autres approches présentées dans ce document). Selon toute vraisemblance, l'ISAWG est une méthode qui **peut** convenir aux contextes les plus difficiles de l'IB, mais qui demande des efforts importants pour la tester et la mettre en œuvre correctement : de tels efforts pourraient être disproportionnés par rapport aux avantages retirés.

L'approche a d'autres inconvénients, puisqu'elle est délicate à mettre en œuvre et peut être très difficile à expliquer à des personnes qui ne la maîtrisent pas.

Les approches qui cherchent à maintenir les résultats obtenus précédemment pour un sous-ensemble de la cohorte (comme l'approche des « centres communs ») sont des manières bien établies d'essayer de tenir compte des changements au sein des cohortes. Elles sont viables pour autant que la cohorte soit suffisamment grande et que suffisamment de centres choisissent la matière d'une année sur l'autre. Bien qu'elles ne permettent pas de maintenir les résultats aussi efficacement que les approches fondées sur les résultats obtenus précédemment, elles sont néanmoins plus efficaces que de nombreuses autres approches, car elles s'appliquent à prendre en compte tous les changements qui affectent les aptitudes des cohortes au fil du temps. Elles conviennent également dans la plupart des contextes de l'IB, à l'exception des cohortes très petites et des toutes nouvelles matières (bien qu'il soit possible d'utiliser les centres communs pour faire le lien avec une matière existante similaire, aussi discutable que cela puisse être).

Les phases ultérieures du projet pourront s'appuyer sur cette revue afin de déterminer les approches pour lesquelles il serait intéressant de mener une modélisation supplémentaire, dans le but d'évaluer leur pertinence pour les différents contextes d'attribution des résultats de l'IB.

Analyse de simulation des seuils recommandés statistiquement

Ce rapport résume les constatations tirées de l'analyse et de la modélisation qui constituent les dernières étapes d'un projet visant à examiner les procédures de détermination des seuils recommandés statistiquement de l'IB, dans le but de pouvoir plus facilement les affiner et les améliorer. Nous avons repris les objectifs de cette analyse ci-dessous en tenant compte de ce contexte.

- Simuler où se situeraient les seuils d'attribution des notes finales en fonction de certaines approches potentiellement réalisables pour déterminer les seuils recommandés statistiquement, en tenant compte d'un large éventail de matières qui représentent l'ensemble des contextes importants de l'IB.
- Examiner dans quelle mesure les résultats de chaque approche simulée s'alignent les uns avec les autres, mais aussi avec les seuils recommandés statistiquement et les seuils d'attribution des notes finales définis dans la pratique, puis déduire quelles seraient les procédures les plus ou les moins adaptées aux contextes de l'IB.

Douze matières ont été sélectionnées pour la modélisation, couvrant les nombreux contextes d'attribution des résultats de l'IB.

Programme	Matière	Contexte de l'attribution des notes finales
Programme du diplôme	Mathématiques	Matières stables, comptant beaucoup d'élèves
Programme du diplôme	Arménien A : littérature	Matières comptant peu d'élèves
Programme du diplôme	Suédois A : littérature	Matières stables, comptant peu d'élèves
Programme du diplôme	Anglais A : langue et littérature	Matières en expansion : **croissance progressive**

Programme du diplôme	Anglais A : littérature	Matières en recul
Programme du diplôme	Politique mondiale	Matières en expansion : **croissance significative**
Programme du diplôme	Technologie de l'information dans une société globale (TISG)	Matières en expansion : **croissance soudaine**
Programme du diplôme	Cinéma	Changements dans le programme d'études et les modèles d'évaluation
Programme du diplôme	Science du sport, de l'exercice et de la santé	Nouvelles matières : **lancement du NS**
PEI	Mathématiques	Nouvelle cohorte dans une matière existante
Programme du diplôme	Théâtre	Modèle de « vérification »
Programme du diplôme	Chinois B	Distribution biaisée

Cinq approches de détermination des seuils recommandés statistiquement ont été modélisées.

1. **Maintenir les normes précédentes** : les seuils pour l'année en cours sont déterminés de façon à se rapprocher le plus possible de la distribution des notes finales de l'année précédente.
2. **Centres communs** : au lieu de conserver les résultats pour l'ensemble de la cohorte, la cohorte de référence et celle de l'année en cours sont d'abord réparties en sous-ensembles dans un groupe défini (dans le cas présent, un groupe de centres présents les deux années). Ce sont les résultats de ce groupe qui sont ensuite utilisés pour comparer la cohorte de référence à celle de l'année en cours (Pinot de Moira, 2019).
3. **Centres communs stables** : cette approche est une variante de la méthode ci-dessus. Le groupe de centres communs est restreint davantage par l'application d'autres critères. Par définition, le critère permettant de préciser la « stabilité » des centres est donc la seule exigence supplémentaire pour cette méthode.
4. **Mise en équivalence par arc de cercle** : utilise un graphique montrant les notes de l'évaluation actuelle par rapport à une évaluation de référence. Un arc de cercle passant par trois points (la note maximale, la note médiane obtenue et la note minimale) est ensuite tracé (Livingston et Kim, 2009).
5. **Instant summary of achievement without grades (ISAWG)** : cet aperçu instantané de la réussite sans notes finales est une approche prédictive qui utilise les résultats obtenus en simultané comme indicateur externe des différences entre les cohortes (Benton, 2017). Autrement dit, elle utilise un amalgame des notes de toutes les composantes pour en déduire un indicateur général de l'aptitude des élèves de toute la série, pour l'année de référence et pour l'année en cours. Ces deux indicateurs sont ensuite mis en équivalence pour établir une relation interannuelle. La mesure ISAWG qui en résulte peut être utilisée à la place des résultats obtenus précédemment pour prédire les résultats.

Conclusions

Nous pouvons résumer ainsi nos conclusions générales concernant les méthodes les plus adaptées aux différents contextes.

- Pour les matières de très faible ampleur comprenant au maximum entre 30 et 50 élèves, la seule méthode viable est celle de la mise en équivalence par arc de cercle.
- Dans les matières en expansion ou en recul qui présentent une croissance ou une diminution non négligeable du nombre d'élèves (différence de 25 à 33 % par an minimum), l'ISAWG semble être la meilleure approche. Cependant, les approches des centres communs peuvent être envisagées dès lors que des données suffisantes soutiennent ce sous-ensemble (compte tenu de la taille des cohortes de l'IB, les centres communs sont plus facilement viables).
- Pour les nouvelles évaluations ou celles qui présentent des changements, l'essentiel consiste à définir la référence avec laquelle comparer la matière : la méthode spécifique de détermination des seuils recommandés statistiquement est de moindre importance.
- Dans les autres contextes non mentionnés, les différences entre les méthodes sont minimes ; en l'absence d'une « vérité » objective, il est difficile de déterminer quelle est la « meilleure ». La distance et la direction par lesquelles les méthodes s'éloignent des limites définies sont souvent similaires. Cependant, si une méthode différente est appliquée pour les matières en expansion ou en recul, il serait sans doute judicieux de l'utiliser également dans les autres contextes afin de réduire l'effet qu'un changement inattendu dans l'aptitude des cohortes pourrait avoir sur la norme.
- Dans l'approche ISAWG ou dans les autres approches, les seuils recommandés statistiquement présentent des similitudes, et par conséquent, leur choix repose généralement sur le pragmatisme ou la rigueur méthodologique. L'ISAWG est la seule méthode qui autorise explicitement les changements d'aptitude « au sein des centres communs » pour une matière, mais elle est nettement plus complexe à mettre en œuvre. Il est possible d'utiliser l'approche des centres communs pour obtenir des résultats rapidement (en se basant idéalement sur des centres communs stables pour autant que le modèle inclue suffisamment de centres et d'élèves), puis d'explorer la méthode ISAWG.

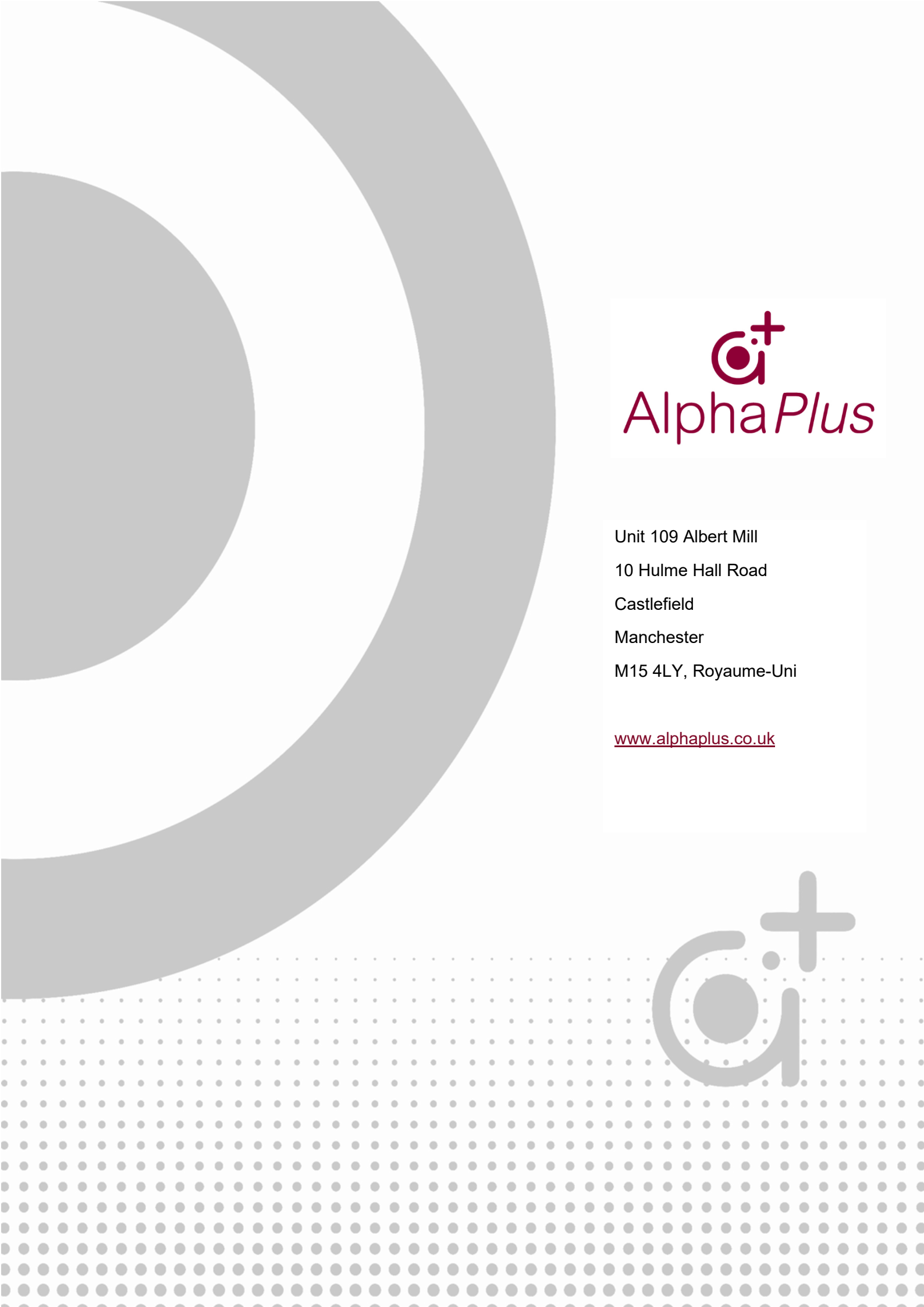
Références bibliographiques

BENTON, T. *Pooling the totality of our data resources to maintain standards in the face of changing cohorts*. Travaux présentés lors de la 18e conférence annuelle AEA-Europe, Prague, République tchèque. Novembre 2017.

IBO. 2018. *Principes et pratiques de l'évaluation – Des évaluations de qualité à l'ère du numérique.* Cardiff, Royaume-Uni : Organisation du Baccalauréat International.

LIVINGSTON, S. A. et KIM, S. The Circle-Arc Method for Equating in Small Samples. *Journal of Educational Measurement.* 2009. Volume 46, numéro 3, p. 330 – 343.

PINOT DE MOIRA, A. 2019. *Common Centres: In the context of maintenance of standards for the GCSE.* Rapport non publié pour le WJEC et le CCEA.



Unit 109 Albert Mill
10 Hulme Hall Road
Castlefield
Manchester
M15 4LY, Royaume-Uni

www.alphaplus.co.uk