Micro Centre
Pythagore inc.
L'évaluation sur mesure

Télécharger
Fichier d' installation
du logiciel EduStat

Contrôle de qualité

Les qualités docimologiques des épreuves

L'évaluation des apprentissages se réalise à l'aide d'épreuves constituées de questions que l'on désigne habituellement sous le terme « item ». Les résultats obtenus à la suite de l'administration d'une épreuve sont utilisés pour prendre des décisions variées. Il peut s'agir de promouvoir un élève, de le sélectionner pour son admission dans une filière scolaire ou un emploi; les résultats peuvent aussi permettre de planifier des programmes d'intervention scolaire. On pourrait énumérer plusieurs autres utilisations des résultats obtenus à la suite d'une évaluation des apprentissages. Aussi, faut-il que les instruments utilisés pour recueillir les renseignements qui serviront de base à l'évaluation respectent de hauts standards de qualité.

Il existe des techniques d'analyses statistiques permettant d'apprécier les qualités docimologiques d'une épreuve. Nous allons faire la présentation de quelques-unes. De plus, nous rappellerons certains concepts reliés aux techniques présentées. Il sera question de l'analyse d'items par rapport à leurs qualités métriques ainsi que par l'absence de biais pénalisant certains sous-groupes d'élèves touchés par l'évaluation.

A. L'analyse d'items

Il est important d'examiner les qualités d'un instrument de mesure et des items qui le composent. Il s'agit du procédé statistique que l'on désigne habituellement sous le vocable d'analyse d'items. Plusieurs techniques peuvent être utilisées pour s'assurer des qualités métriques d'une épreuve. Nous présentons brièvement ce que l'on entend par *analyse classique d'items* et par les techniques basées sur la *théorie des réponses aux items*.

1. L'analyse classique

L'*analyse classique d'items* concerne les réponses à choix multiple ou sur une échelle continue ou à crédits partiels. Elle permet l'estimation de la qualité métrique de chaque item faisant partie d'un test ou d'une épreuve. Ce type

d'analyse permet l'estimation de la qualité métrique de chaque item faisant partie d'un test ou d'une épreuve. Trois séries d'indices statistiques sont ordinairement retenues.

- *La difficulté* – Pour chaque item faisant partie d'une épreuve, on doit en calculer la difficulté relative; une épreuve doit contenir des questions faciles, moyennement faciles, plutôt difficiles afin de recueillir des renseignements sur l'ensemble des élèves soumis à l'évaluation. Pour les items à choix multiple, il s'agit de la proportion d'élèves ayant répondu correctement à la question. Pour les items dits à crédits partiels, il faut calculer un résultat moyen.

- *La discrimination* – Les questions posées aux élèves évalués doivent permettre de bien identifier ceux qui maîtrisent ce qui est mesuré et ceux qui ne maîtrisent pas les notions, habiletés ou compétences évaluées; il s'agit de connaître certaines caractéristiques de ces deux groupes d'élèves. Des indices statistiques fournissent des renseignements sur la capacité des items à bien identifier les élèves dits « performants » et ceux moins performants.

- *La cohérence* – Comme on le sait, une épreuve est constituée de plusieurs items ou questions. On peut s'interroger sur la cohérence qui devrait s'établir entre les différents renseignements recueillis. Il s'agit de la cohérence interne de l'instrument.

Le logiciel EduStat permet de réaliser une telle analyse. C'est ainsi que pour chaque item d'une épreuve à choix multiple, un rapport présente le nombre d'individus, la proportion de ceux qui ont réussi l'item, le coefficient de corrélation item/test. Il est aussi possible d'avoir, en parallèle, d'autres statistiques souhaitées (par exemple, les pourcentages de réussite à l'ensemble du test de ceux qui ont réussi l'item et de ceux qui l'ont échoué, le coefficient Alpha recalculé en excluant l'item analysé, les statistiques de discrimination par rapport à chaque leurre, le regroupement des items par rapport au tableau de spécification, la description de chaque item). Par ailleurs, pour chaque item dont les réponses se situent sur une échelle continue (correction à crédits partiels), le rapport fournit le nombre d'individus, le minimum et le maximum observés, la moyenne obtenue, l'expression de cette moyenne sur 100 ainsi que le coefficient de corrélation item/test. Pour l'ensemble de l'épreuve et par rapport à chaque regroupement d'items, le rapport fournit de plus le coefficient de consistance interne sous la forme de l'Alpha de Cronbach, l'erreur type de mesure et l'écart type.

 © 2018 Micro Centre Pythagore Inc.

2. L’analyse en utilisant la théorie des réponses aux items

La théorie des réponses aux items est un ensemble de procédés statistiques dont le but est d'évaluer la qualité d'un instrument de mesure et des items qui le composent. Il est possible de retenir des modèles tenant compte de l'un ou l'autre paramètre suivant (ou de l'ensemble des trois) :

- La *discrimination*, c’est-à-dire le pouvoir de l’item de bien identifier les élèves qui maîtrisent ce qui est mesuré et ceux qui ne maîtrisent pas les notions, habiletés ou compétences évaluées
- La *difficulté* de l’item, c’est-à-dire la plus ou moins grande facilité de l’item pour les élèves visés par l’épreuve.
- La *pseudo chance*, c'est-à-dire la possibilité pour l'élève d'obtenir une réponse correcte au hasard sans posséder réellement la connaissance requise.

Il existe des logiciels spécialisés pour réaliser ces calculs. Le logiciel EduStat permet la préparation de fichiers pouvant être utilisés par de tels logiciels. Par ailleurs, le logiciel EduStat peut utiliser les résultats calculés par un autre logiciel (il s’agit de XCalibre) pour tracer des courbes illustrant les résultats obtenus à l’analyse réalisée pour chaque item d’une épreuve.

B. L’analyse du fonctionnement différencié des items

Il est important de s’assurer que certaines questions faisant partie d’une instrumentation d’évaluation ne favorisent pas un sous-groupe d’élèves au détriment des autres. Il s’agit de l’examen de biais qui peuvent toucher plusieurs dimensions socio-économiques ou culturelles. On peut vouloir examiner s’il y a présence de biais par rapport à deux sous-groupes d’élèves (par exemple, les garçons et les filles) ou bien plusieurs (par exemple, les régions administratives). Les techniques d’analyses seront différentes pour chacune de ces situations. Les items identifiés comme « biaisés » à partir de la valeur des indices calculés devraient être retranchés de l’épreuve ou bien modifiés.

1. L’analyse « omnibus »

La technique dite « Omnibus » permet le calcul d’indices par rapport à plusieurs groupes d’enregistrements traités de façon simultanée. Le rapport produit contient un ensemble de statistiques sur chaque item retenu pour l’analyse. De plus, un graphique illustre l'adéquation de la distribution théorique à la distribution empirique qui s'évalue par les coordonnées des valeurs Rang (l'axe des «X») et des valeurs DR (l'axe des «Y»). Une bonne adéquation entre les valeurs

 © 2018 Micro Centre Pythagore Inc.

«Rang» et «DR» est obtenue par un graphique dans lequel les points (x, y) tracent une droite linéaire. Tout point s'écartant de manière importante de cette droite linéaire s'interprète comme une donnée potentiellement marginalisée.

2. L'analyse par rapport à deux groupes

Si l'on doit examiner le fonctionnement différencié des items par rapport à deux groupes, la technique de Mantel-Haenszel peut alors être utilisée. Le rapport produit par le logiciel EduStat à la suite de l'utilisation des options de calcul des coefficients « M-H » contient les éléments suivants :

- l'identification des groupes de référence et comparé;
- pour chaque item, les indices Alpha, Delta et l'erreur type rattachée au Delta;
- pour chaque item, l'intervalle de confiance du Delta;
- la classification des items par rapport à trois catégories de biais :
    - C : item à examiner prioritairement par rapport à la possibilité de fonctionnement différencié;
    - B : deuxième catégorie d'items susceptibles de posséder un fonctionnement différencié;
    - A : item qui n'est pas identifié comme susceptible d'avoir un fonctionnement différencié.

Voici quelques notes permettant d'interpréter les statistiques produites lors de l'analyse du fonctionnement différencié des items par la méthode du Mantel-Haenszel.

> Les items identifiés par l'étiquette « C » sont ceux qui risquent de comporter des biais affectant leur fonctionnement. La valeur absolue du Delta est alors égale ou supérieure à 1,5 et est significativement différente de 1.
>
> Quant aux items les moins susceptibles d'être biaisés (la catégorie « A »), la valeur absolue du Delta est inférieure à 1 ou n'est pas significativement différente de « 0 ».
>
> Les autres items dont la valeur absolue est entre 1 et 1,5 se retrouvent dans la catégorie « B » et peuvent être moyennement biaisés.

Une valeur positive du Delta indique que l'item a été plus difficile pour les individus appartenant au groupe de référence. Une valeur négative identifie donc un item plus difficile pour le groupe comparé au groupe de référence.

Adresse courriel : info@mcpythagore.com

 © 2018 Micro Centre Pythagore Inc.