



Les examens QCM.

Comment lire et interpréter les rapports d'analyse d'items (Contest) en vue d'améliorer la qualité de vos examens ?

Résumé :

De nombreux enseignants optent pour un examen par questionnaire à choix multiples (QCM).

Si ce type d'examen peut sembler facile à réaliser et permettre des économies en temps de correction, il suscite également nombre de questions, dont :

- Mon QCM est-il bien conçu ?
- Dois-je augmenter le nombre d'items ? Puis-je me permettre de le réduire ?
- L'examen est-il trop simple ou trop difficile ?
- Les scores des étudiants sont-ils fiables ?
- Les leurres sont-ils efficaces ?
- Quels items pourrais-je conserver pour un autre examen ?
- Quels sont les items à éliminer ?

Répondre à ces questions n'est pas si compliqué lorsque l'on utilise correctement les outils à disposition des enseignants ayant recours au système Contest via le service EVA et que l'on exploite les règles de base de composition d'un bon QCM.

Cette note, vous explique comment utiliser ces outils.

Introduction :

L'analyse d'items reprend l'ensemble des procédés statistiques dont le but est d'évaluer la qualité d'un examen et des items (c'est-à-dire les questions) qui le composent.

Ce type d'analyse présente plusieurs avantages :

- Obtenir des indications sur la précision et la fiabilité des résultats observés.
- « Valider » un examen en retirant, au besoin, les questions qui présentent des lacunes.
- Calibrer les examens futurs.
- Créer une « banque d'items ».
- Améliorer la formulation des questions.

Nous vous proposons 3 rapports d'analyse d'items, grâce à l'outil Contest (Consilium, 2011) :

- Le test analysis report : ce rapport présente en une page les indices statistiques qui permettent d'évaluer les résultats de l'examen pris globalement.
- L'item analysis report présente les indicateurs statistiques qui permettent d'évaluer les résultats de chacune des questions (items) qui composent votre examen, ainsi que la répartition des choix de réponses cochées par les étudiants.

- L'item advice report permet de visualiser la place qu'occupe chaque question de votre examen en fonction de leur niveau de difficulté relative et de leur capacité à bien différencier les étudiants entre eux (indice de discrimination).

Ces deux derniers rapports vous aident à repérer les questions qui posent problème, celles qu'ils mériteraient d'être révisées ou remplacées en vue d'une prochaine session d'examen. Ils permettent aussi d'identifier les "meilleures" questions, celles qui devraient vous inspirer lors de la rédaction de nouvelles questions.

Nous vous proposons une démarche d'analyse progressive pour vous aider à lire et à interpréter les résultats de ces rapports.

1. Analyse globale de l'examen : Test analysis report

Ce 1er rapport à examiner comporte deux volets :

Des informations générales (*General details*) :

- le nombre de questions
- le nombre d'étudiants ayant participé à votre examen
- les pourcentages de réussite et d'échec à votre examen
- le score le plus bas que peut obtenir un étudiant
- le score probable d'un étudiant répondant totalement au hasard (chance score)
- le score le plus haut que peut obtenir un étudiant
- etc.

Entire Test		
General details		
Questions	20	
Excluded questions	0	
Participants	212	
Participants passed	76	35,85 %
Participants failed	136	64,15 %
Minimum score	-7,50	
Chance score	3,13	
Maximum score	20,00	
Minimum grade	0,00	
Maximum grade	20,00	
Pass grade	10,00	

Des données statistiques (*Statistical details*) :

Sans entrer dans les détails à ce stade-ci, nous vous conseillons de regarder en priorité les valeurs de deux indices importants (**l'Item-rest correlation** et **l'Alpha de Cronbach**) et de vérifier s'ils atteignent les normes de référence que l'on trouve dans la littérature.

Statistical details				
	Average	Stdev	Lowest	Highest
Score	8,58	3,33	0,00	16,00
P	0,57	0,18	0,24	0,88
Corrected p	0,33	0,21	-0,02	0,75
Item-rest correlation	0,15	0,14	-0,15	0,34
Grade	8,58	3,33	0,00	16,25

Cronbach's Alpha	0,45
Standard error of measurement	1,09
Estimated Alpha for similar test with norm length 40 items	0,62
Cronbach's Alpha (norm)	0,10 0,20 0,30 0,40 0,50 0,60 0,70 0,80 0,90 0,99
Required replications (K)	0,13 0,30 0,52 0,80 1,21 1,81 2,81 4,82 10,85 119,32

- L'item-rest correlation (indice de discrimination global du QCM – voir aussi section suivante) devrait atteindre une valeur égale ou supérieure à **0,20**
- L'alpha de Cronbach (indice de cohérence interne encore appelé d'homogénéité interne) ne devrait pas être inférieur à 0,70. De préférence, il doit être égal ou supérieur à **0,80**.

Pour en savoir plus :

Coefficient alpha de Cronbach. Indice statistique variant entre 0 et 1 qui permet d'évaluer l'homogénéité (la consistance ou cohérence interne) d'un instrument d'évaluation ou de mesure composé par un ensemble d'items qui, tous, devraient contribuer à appréhender une même entité (ou dimension) "sous-jacente": le niveau de connaissance ou de compétence sur un thème donné; le niveau d'aptitude, d'attitude, de motivation, d'intérêt dans tel domaine ou par rapport à tel objet, etc.

Cet indice traduit un degré d'homogénéité (une consistance interne) d'autant plus élevé(e) que sa valeur est proche de 1. Dans la pratique, on considère généralement que l'homogénéité de l'instrument est satisfaisante lorsque la valeur du coefficient est au moins égale à 0.80.

(Gianreto Pini, *Petit lexique pour l'évaluation de la mesure en éducation : termes et expressions du langage méthodologique.*

http://www.irdp.ch/edumetrie/lexique/coef_alpha.htm)

2. Analyse des items de votre examen

LES INDICATIONS REPRISES DANS CETTE SECTION NE SONT VALABLES QUE POUR UN QCM POUR LESQUELS UNE SEULE REPONSE CORRECTE EST ATTENDUE PAR QUESTION

Ceci se fait au moyen de deux rapports : l'**Item analysis report** et l'**Item advice report**.

- L'*Item analysis report* présente les indicateurs statistiques qui permettent d'évaluer la **qualité de chacune des questions (items)** qui composent votre examen.
- L'*Item advice report* en propose un résumé graphique pour une lecture rapide.

Ces deux rapports se basent sur **deux types d'indicateurs importants** :

- Les indices de facilité (P, P')
- Les coefficients de discrimination (Rir, Rit)

2.1. Les indices de facilité des items : P et P'

La question traitée ici est celle de la proportion d'étudiants qui ont répondu correctement à cette question.

L'**indice de difficulté P** d'une question donne le pourcentage de réussite à la question.

Pour en savoir plus :

Indice de difficulté P. Différentes procédures peuvent être utilisées pour exprimer le degré de difficulté d'un item à un groupe (...). Dans le cas d'items donnant lieu à une appréciation dichotomique du type juste – faux (et cotées 0 ou 1), la difficulté de l'item est exprimée par la proportion des étudiants qui ont sélectionné la bonne réponse à cette question : il s'agit du nombre d'étudiants qui ont répondu correctement, divisé par le nombre total d'étudiants.

(Gianreto Pini, Petit lexique pour l'évaluation de la mesure en éducation : termes et expressions du langage méthodologique.

http://www.irdp.ch/edumetrie/lexique/indice_difficulte.htm)

Par exemple, si pour une question, l'indice $p=0,70$, cela signifie que 70 % des étudiants ont donné la bonne réponse à cette question :

- **Plus ce pourcentage est élevé, plus la question est facile.**
- **Plus ce pourcentage est faible, plus la question est difficile.**

Exemples de questions à se poser (Serge Normand, Université de Montréal, 2001)

- P est inférieur à 0,50 : pourquoi la moitié des étudiants échouent-ils à cette question ?
- P est inférieur à 0,30 : est-ce normal qu'autant d'étudiants n'atteignent pas l'objectif visé par la question ? Combien d'autres questions y-a-t-il avec des indices aussi faible dans le test ?
- P est supérieur à 0,90 : est-ce que cette question correspond à un objectif important ? Faut-il conserver, pour une session suivante, les questions réussies par la quasi-totalité des étudiants ?
- P est égal à 1,00 : tout le monde réussit : faut-il conserver cette question ?

Pour en savoir plus :

Le corrected P (indice P') correspond à l'indice **P** corrigé pour tenir compte de l'effet du hasard. Dans un QCM, il faut en effet tenir compte de la probabilité de réussir des questions sans vraiment en connaître la réponse correcte.

Il arrive que l'on enregistre des valeurs négatives pour l'indice **P'** . Cela arrive lorsque le taux de réussite (indice **P**) est inférieur au résultat que peut espérer un étudiant en répondant au hasard. Un **P'** négatif permet de repérer les questions particulièrement difficiles à résoudre.

A retenir

P et P' : *quelles leçons en tirer ?*

Le principal objectif du calcul des indices de facilité est d'identifier les questions dont le niveau de difficulté est approprié à l'objectif de l'évaluation.

- *Les questions où tous les étudiants réussissent ($p = 1,00$) ou échouent ($p = 0,00$) ne sont pas suffisamment informatives car elles ne permettent pas de discriminer les étudiants.*
- *Si l'examen ne comprend que des questions très faciles ou très difficiles, il ne permet pas de discriminer deux étudiants moyens puisque les deux réussiront toutes les questions faciles et échoueront à toutes les questions très difficiles, ils auront donc tous les deux la même cote.*
- *A l'inverse, si l'examen ne comporte que des questions de difficulté moyenne ($p = 0,50$), il permettra de discriminer correctement les sujets moyens, mais ne permettra pas de discriminer les sujets faibles entre eux (car ils échoueront à toutes les questions) ni les sujets forts entre eux (car ils réussiront toutes les questions).*

*Habituellement, dans un examen certificatif, on cherche à identifier les différences individuelles. Dans cette situation classique, un examen bien équilibré est donc un examen qui comporte une **majorité de questions de facilité moyenne ($p=0,50$)**, quelques questions faciles ($p=0,90$), et quelques questions difficiles ($p=0,10$). Ainsi, les étudiants moyens, qui sont les plus nombreux, seront bien discriminés, et les étudiants faibles, ou forts, trouveront également de quoi s'exprimer. Tout dépend, en fait, de l'objectif de votre examen.*

- *Dans le cas d'un examen qui s'apparente à un concours ou à un test de sélection, on voudra sélectionner les étudiants les plus forts. On privilégiera donc les questions difficiles ($p = 0,10$).*
- *Dans le cas où l'examen a pour but de vérifier que la majorité des étudiants ont bien atteint les objectifs minimaux ou essentiels d'un cours, il est normal que l'examen soit plus facile et on choisira donc majoritairement des questions avec un indice de facilité élevé ($p = 0,80$ ou $0,90$).*

2.2. Les coefficients de discrimination des items Rir et Rit

La question traitée ici est double :

- est-ce que les étudiants qui réussissent cette question ont également un score élevé pour l'ensemble des autres questions ?
- est-ce que les étudiants qui échouent à cette question ont aussi un score plus faible à l'ensemble des autres questions ?

Pour en savoir plus :

L'objectif d'une évaluation certificative est de vérifier que les étudiants ont bien atteint un niveau de connaissance et de compétence suffisant. Pour atteindre cet objectif, il est primordial que le questionnaire d'évaluation soit valide, c'est-à-dire qu'il permette de révéler le niveau de compétence réel de chacun des étudiants : seuls les étudiants ayant atteint le niveau requis par l'enseignant devraient pouvoir obtenir le score maximal ; un score faible doit idéalement être le reflet d'une connaissance et compétence insuffisante. (Morissette, 1993, p. 310).

Considérons 2 publics :

- d'une part les étudiants ayant obtenu un score élevé sur l'ensemble de l'examen (dits étudiants compétents ou forts);
- et d'autre part les étudiants ayant obtenu un score faible sur l'ensemble de l'examen (dits étudiants faibles),

on s'attend naturellement à ce que chaque question soit réussie par une proportion plus importante des étudiants de cette première catégorie, comparativement à la proportion d'étudiants de la seconde catégorie.

Pour en savoir plus :

Dans cette perspective, un "bon" item est un item qui serait réussi par une plus grande proportion de sujets ayant obtenu un score élevé à l'examen que par des sujets ayant obtenu un score faible. (Laveault & Grégoire, 1997, p. 231).

Cette capacité à trier les deux catégories d'étudiants ce qu'on appelle le **pouvoir discriminant** d'une évaluation. Il se mesure au moyen des indices de corrélation (ou indice de discrimination) suivant :

- Rir (Item Rest correlation) : disponible pour l'ensemble du test (cf section 1) et pour chaque question ;
- Rit (Item Test correlation) : disponible pour chaque question .

2.2.1. Le Rir d'un item

Cet indice permet de vérifier si la question remplit bien sa fonction, à savoir : différencier les étudiants en fonction de leur niveau de connaissance ou de compétence.

Pour en savoir plus :

L'item rest correlation (Rir) calcule, pour chaque question, la corrélation entre le score à cette question, et le score total calculé sur base des résultats obtenus aux autres questions de l'examen. Ce coefficient varie entre -1,00 et 1,00.

- Une question qui a un Rir de **1,00** est une question pour laquelle on observe que tous les étudiants qui ont donné la bonne réponse ont également donné la bonne réponse à toutes les autres questions de l'examen et que tous les étudiants qui ont donné une réponse incorrecte ont également donné une réponse incorrecte aux autres questions de l'examen.
- A l'inverse, un Rir de **-1,00** signifie que les étudiants qui ont donné la bonne réponse à cette question ont donné des réponses incorrectes à toutes les autres questions de l'examen et ceux qui ont donné une réponse incorrecte à cette question ont donné la réponse correcte à toutes les autres questions de l'examen.

Un item au pouvoir discriminant élevé apporte beaucoup d'information sur la compétence des étudiants, un item peu discriminant renseigne peu sur la compétence du sujet.

Le Rir d'un item permet de répondre aux questions qui introduisent cette section :

- un indice positif indique que les meilleurs étudiants réussissent la question et que les plus faibles échouent à la question ;
- un indice négatif signale une incohérence : les meilleurs étudiants échouent à la question, les plus faibles la réussissent ;
- un indice proche de zéro indique qu'il n'y a aucune relation entre le score à cette question et le total des autres questions ;
- de manière générale, la valeur 0,20 est prise comme balise pour indiquer un début de discrimination.

2.2.2. Le Rit d'un item

L'*item-test correlation (Rit)* fonctionne sur le même principe, à la nuance près qu'il calcule, pour chaque question, la corrélation entre le score obtenu à cette question et le score total obtenu pour l'ensemble de l'examen (la question analysée y compris). Il s'interprète de la même manière que le Rir. Le Rir est souvent préféré au Rit car sa valeur n'est pas influencée par le résultat obtenu à la question analysée.

Le nombre de questions et le nombre d'étudiants influencent le calcul des indices de corrélation, qui sont dès lors tributaires de la composition de l'examen et de son contexte de passation.

A retenir

Les questions ayant un Rir à valeur négative doivent absolument être analysées.

Deux cas de figures peuvent se présenter :

- Il est possible qu'il y ait eu une erreur dans votre déclaration.
La bonne réponse attendue n'a pas été encodée comme telle dans Contest.
- Si la réponse correcte a bien été encodée dans Contest, c'est probablement que la question comporte une absurdité, une incohérence, ou qu'elle repose sur un énoncé incomplet. La question devrait dans ce cas être retirée de l'examen.

Par ailleurs la discrimination est très fortement influencée par la facilité ou la difficulté d'un item. En règle générale, des items très faciles ou très difficiles discriminent habituellement peu.

« Si tous les étudiants, à la limite, réussissent ou ratent l'item, la discrimination devient nulle, car il n'y a plus aucune différence entre les étudiants du groupe supérieur et ceux du groupe inférieur. »

(Morissette, 2005, p. 313).

Pour en savoir plus :

Valeurs repères pour interpréter les coefficients de discrimination :

- 0,40 et plus Question qui discrimine très bien
- 0,30 à 0,39 Question qui discrimine bien
- 0,20 à 0,29 Question qui discrimine peu
- 0,10 à 0,19 Question limite, à améliorer
- **Moins de 0,10 Question sans intérêt réel pour une évaluation certificative**

(Ebel, 1965)

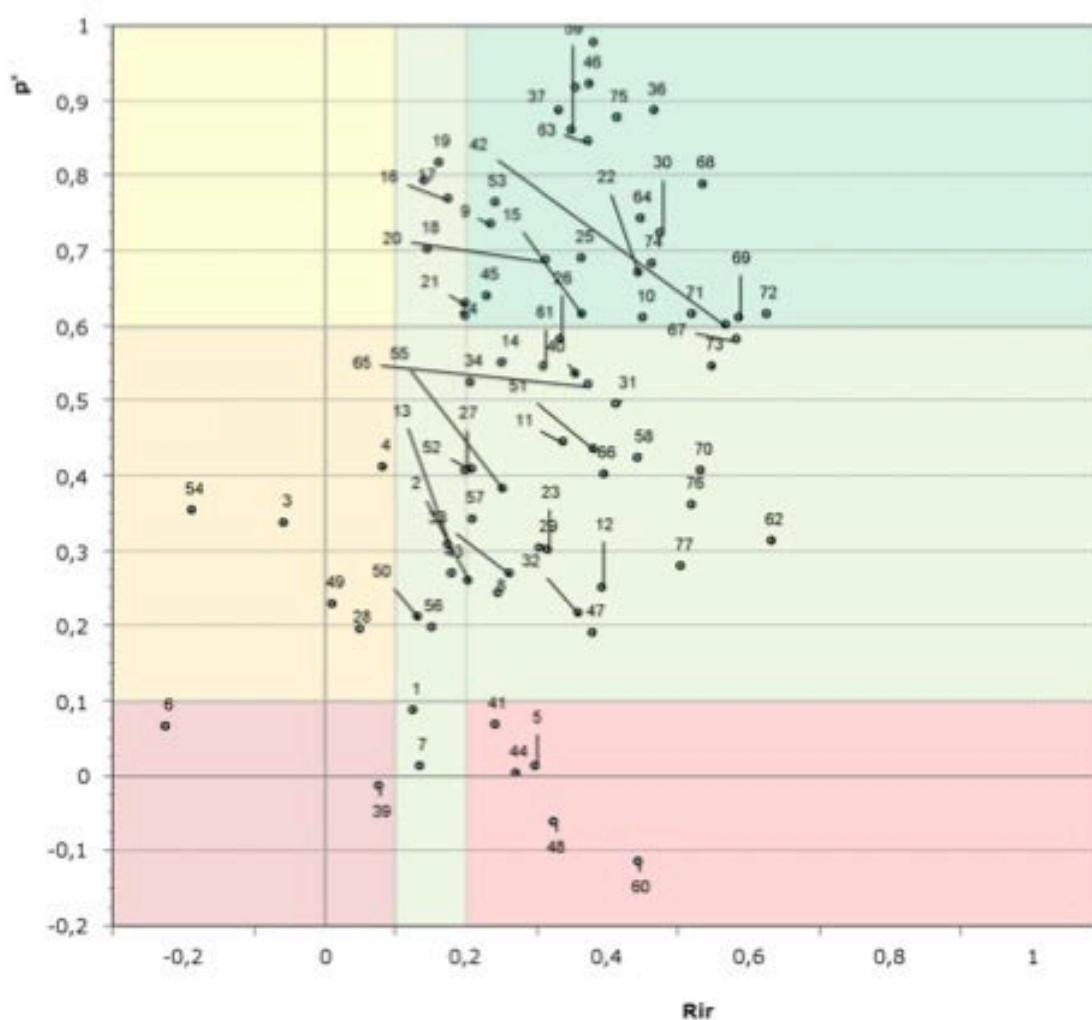
2.3 Utilisation de l'Item advice report

Après consultation du *Test analysis report* (qui vous donne une idée générale de la qualité de votre examen), nous vous conseillons d'examiner l'*Item advice report* pour commencer l'analyse de la qualité des questions qui le composent.

L'*Item advice report* permet de visualiser la place qu'occupe chaque question de votre examen en fonction des deux indicateurs décrits ci-dessus :

- Le coefficient de discrimination (Rir) sur l'axe horizontal ;
- L'indice P' sur l'axe vertical.

Chacun des chiffres figurant dans ce graphe correspond au numéro de la question de votre examen.



A retenir

L'Item advice report : quelles exploitations en faire ?

- Avoir une représentation visuelle du niveau de difficulté des questions de l'examen :
 - si la majorité de vos questions sont situées dans la partie supérieure du graphe cela signifie que votre examen est "facile".
 - à l'inverse, votre examen est "difficile" si la majorité de vos questions se concentrent au bas de la figure.
- Repérer les questions qui ne jouent pas en harmonie avec les autres :
 - ce sont en quelque sorte des "électrons libres", on les repère car ces questions se situent en orbite de la majorité des autres questions qui devraient normalement former une "constellation" assez serrée (indicateur de l'homogénéité de l'examen).
- Identifier les questions qui discriminent peu ($Rir < 0,10$) ou encore négativement ($Rir < 0,00$).

Ces questions sont sans doute problématiques et

- **devraient être revues voire supprimées ou**
- **vous inciter à revoir la manière dont les scores sont transformés en grade**

2.4. Utilisation de l'Item analysis report

Après avoir repéré les questions problématiques dans *l'Item advice report*, nous vous conseillons de vous tourner vers *l'Item analysis report*.

Lorsque des questions discriminent peu ou discriminent négativement ($Rir < 0,10$), il est nécessaire de regarder ces questions de plus près afin de mieux comprendre ce qui a pu se passer.

Dans le cas de QCM où une seule réponse par question est attendue, il est possible d'identifier quel pourcentage d'étudiants forts ou faibles a opté pour chacune des options de réponse.

A partir de ces résultats, Contest calcule un indice de discrimination non seulement pour la bonne réponse (cf. ci-dessus) mais aussi pour les distracteurs (les réponses incorrectes).

- Ces coefficients de discrimination pour les distracteurs devraient tous être négatifs car ils sont censés être choisis par une très faible proportion d'étudiants ayant eu un bon résultat aux autres questions (dits compétence).
- À l'inverse, le coefficient de discrimination de la bonne réponse devrait être positif.

	p	p'	Rir	Rit	Open	A		B		C		D	
						f	z	f	z	f	z	f	z
1	0,93	0,91	0,27	0,32	5	<u>150</u>	<u>0,03</u>	4	-0,73	0	*	4	-0,57
2	0,46	0,29	0,34	0,44	83	0	*	<u>48</u>	<u>0,42</u>	30	-0,61	2	-0,94
3	0,95	0,94	0,64	0,67	4	<u>154</u>	<u>0,10</u>	3	-3,64	1	-2,63	1	-2,63
4	0,52	0,36	-0,18	-0,07	112	0	*	<u>47</u>	<u>0,10</u>	1	-0,40	3	-1,49
5	0,93	0,90	0,47	0,52	3	1	-3,15	6	-1,45	<u>150</u>	<u>0,09</u>	3	-0,75
6	0,89	0,86	0,55	0,60	10	4	-2,55	5	-0,86	2	-2,65	<u>142</u>	<u>0,14</u>
7	0,82	0,76	0,49	0,56	12	5	-1,11	12	-1,19	<u>130</u>	<u>0,21</u>	4	-1,90
8	0,91	0,89	0,55	0,59	6	3	-1,47	4	-2,75	3	-1,39	<u>147</u>	<u>0,13</u>
9	0,90	0,87	0,68	0,71	8	6	-2,65	5	-2,21	0	*	<u>144</u>	<u>0,19</u>
10	0,54	0,38	0,10	0,21	103	3	-0,69	2	-1,41	<u>53</u>	<u>0,08</u>	2	0,23
11	0,90	0,87	0,46	0,51	9	2	-1,65	<u>144</u>	<u>0,11</u>	5	-2,33	3	-0,08
12	0,63	0,50	0,36	0,46	55	<u>84</u>	<u>0,24</u>	14	-0,48	2	-1,68	8	-1,23
13	0,87	0,83	0,58	0,63	14	<u>137</u>	<u>0,19</u>	6	-1,42	3	-3,87	3	-1,95
14	0,75	0,66	0,14	0,24	45	4	-1,92	<u>107</u>	<u>0,11</u>	1	-0,59	6	-0,55
15	0,88	0,84	0,59	0,64	9	6	-1,30	3	-2,52	<u>140</u>	<u>0,19</u>	5	-2,33
16	0,50	0,33	0,26	0,37	102	6	-0,60	6	-0,84	2	-2,32	<u>47</u>	<u>0,28</u>
17	0,92	0,89	0,64	0,68	5	<u>148</u>	<u>0,16</u>	2	-1,89	4	-2,66	4	-2,37
18	0,53	0,37	0,26	0,37	90	<u>56</u>	<u>0,22</u>	10	-0,74	7	-0,68	0	*
19	0,62	0,49	0,13	0,24	80	2	-1,57	<u>74</u>	<u>0,12</u>	1	-0,58	6	-0,91
20	0,33	0,11	0,17	0,27	91	28	0,00	<u>24</u>	<u>0,37</u>	17	-0,38	3	-0,82

A retenir

L'Item analysis report : quelles exploitations en tirer ?

Dans le tableau ci-dessus

- Les propositions de réponse sont désignées par les lettres A, B, C, D, E, ...
- La bonne réponse apparaît **en gras soulignée**.
- Pour chacune des propositions de réponse, y compris la bonne réponse,

Contest réalise deux calculs :

- f = nombre d'étudiants ayant choisi la proposition de réponse en question
- z = coefficient de discrimination pour chacune des propositions

Une valeur z associée à la bonne réponse (en gras souligné) proche de zéro ou négative dénote un problème.

Une valeur z prenant une valeur positive pour un ou pour plusieurs leures révèle ce que l'on appelle des **distracteurs attirants car ils sont choisis majoritairement par des étudiants compétents sur l'ensemble de l'examen et trop peu par les étudiants faibles**.

Une telle situation n'est évidemment pas normale et suggère une analyse plus poussée des questions problématiques :

Deux cas de figure peuvent se présenter :

- Il est possible qu'il y ait eu une erreur dans votre déclaration. Un distracteur attirant est en fait la bonne réponse mais n'a pas été encodée comme telle dans Contest. L'encodage doit être adapté dans Contest.
- La question est mal conçue et il convient de considérer un ou plusieurs distracteurs attirants comme réponses acceptées. L'encodage doit être étendu dans Contest.

Dans les autres cas, il est nécessaire d'analyser la question pour essayer de comprendre ce qui a pu pousser les étudiants compétents à choisir cette mauvaise réponse avec pour résultat :

- de la réviser en vue d'un prochain examen (améliorer la question en entier, reformuler un nouveau distracteur, supprimer le distracteur attirant, voire abandonner la question en tant que telle)
- d'éventuellement l'abandonner dans la présente épreuve (modification de l'encodage dans Contest).

Une dernière donnée de l'Item analysis report est toutefois à prendre en compte, en lien avec la difficulté des questions (P) : les non réponses (colonne « Open »), sachant qu'il peut s'agir :

- d'une déclaration d'ignorance ou de non prise de risque par l'étudiant (en cas de pénalisation pour un choix d'une mauvaise réponse) ;
- d'une omission par manque de temps.

Pour vos prochains QCM :

Une façon d'améliorer vos prochains examens consisterait à analyser de plus près les questions qui discriminent bien et qui ont un niveau de difficulté adéquat. Ces questions sont précieuses car elles devraient vous inspirer lors de la rédaction de nouvelles questions.

Pour résumer :

Valider un examen revient à

- Analyser les résultats de chaque question :
 - Indice de facilité P
 - Indice de discrimination Rir
 - Répartition des choix de réponses (f et z)
- Analyser les résultats de l'ensemble du test :
 - Indice de discrimination global de l'examen (Rir)
 - Indice de cohérence interne (alpha de Cronbach)
 - La caractéristique de la distribution des scores (dans le Score Frequency Report)

Ce qui vous permet de

- Tirer un maximum de profit des résultats d'un examen
- Faire face aux éventuels recours
- Obtenir des indications sur la précision des résultats
- Retirer le cas échéant les questions problématiques
- Ajuster le calcul de la note finale (grade)
- Améliorer la formulation des questions pour une session ultérieure
- Calibrer de futurs examens
- Créer une banque d'items

Pour en savoir plus :

Validité d'un QCM : Degré d'adéquation entre ce que l'on déclare faire (évaluer telle ou telle dimension) et ce que l'on fait réellement, entre ce que l'outil mesure et ce qu'il prétend mesurer.

Fiabilité d'un QCM : Degré de confiance que l'on peut accorder aux résultats observés : seront-ils les mêmes si on recueille l'information à un autre moment, avec un autre outil, par une autre personne, etc. ?

(Laveault & Grégoire, 1997).

Bibliographie

Beech, J.R., Harding, L. & coll (1994). *Tests, mode d'emploi, guide de psychométrie*, (trad. de J.-Luc Mogenet), Paris, ECPA, les Ed. du Centre de psychologie appliquée.

Bernier, J-J. & Pirtulewicz, B. (1997). *La psychométrie. Traité de mesure appliquée*. Montréal, Paris, Casablanca : Gaëtan Morin Editeur.

Consilium Benelux (2011). *ConTest Statistics. Version 0.19. Mars 2012*.

EBEL, R.L. (1965). *Measuring Educational Achievement*. Upper Saddle River, New Jersey (USA) : Prentice-Hall education series.

Laveault, D. & Grégoire, J. (1997). *Introduction aux théories des tests en sciences humaines*. Paris, Bruxelles : De Boeck Université & Larcier s.a.

Morissette, D. (1997). *Guide pratique de l'évaluation sommative : gestion des épreuves et des examens*. Paris, Bruxelles : De Boeck & Larcier.

Morissette, D. (2005). *Les examens de rendement scolaire*. Sainte-Foy : Les Presses de l'Université Laval.

NORMAND, S. (2004). *Guide de lecture d'un rapport d'analyse d'items*. Université de Montréal, Faculté de Médecine, Bureau d'évaluation. Montréal.
<http://medecine.ulb.ac.be/tools/docimo/Guide%20de%20lecture%20AnItem.pdf>.

NORMAND, S. (2001). *L'analyse d'items, une technique docimologique pour valider des questions d'examen*, Diaporama du mode d'emploi du logiciel AnItem, Bureau d'évaluation, Faculté de médecine, Université de Montréal.
http://www.medbev.umontreal.ca/docimo/DocSource/Docimologie_AnItem.pp