(ISSN: 2805-413X) Evariste KAKULE MUKULU[1].*

https://ijojournals.com/ *Volume 07 || Issue 08 || August, 2024 ||*

Les apports de quelquestechniquesd'analysedes itemspourl'optimisationdesmesuresscolaires dans les approches de l'évaluation critérielle et normative.

✍ Evariste KAKULE MUKULU*

Résumé.

La psychométrie classique propose la réduction de la longueur d'un instrument de mesure scolaire par la sélection des items enfin d'optimiser sa fidélité. Quant à la théorie de généralisabilité ; soit l'allongement de l'instrument de mesure, soit la réduction de l'instrument, soit encore limiter la portée de la conclusion seulement aux instruments considérés dans l'étude sont les voies d'optimisation de la fidélité de mesure.

Dans le présent article, nous avons voulu vérifier si en réduisant les instruments de mesurepar la sélection des items à l'aide de certaines techniques d'analyse des items, nous pouvons optimiser la fidélité. Sixdifférentes technique de sélection d'items notamment ; la technique de F. DAVIS, la technique de J. BROWN, la technique de GRUNLUND, la technique de RIGAUX, la technique de corrélation item test, la technique d'analyse des facettes ; celles-ci nous ont permis de constituer cinq différentes épreuves.

Après analyse, un constat est que, les coefficients de fidélité(généralisabilité) calculés sont satisfaisant dans l'approche de l'évaluation critérielle, bien qu'aucune technique d'analyse des items n'a permis une optimisation remarquable, car l'épreuve de départ avait déjà une fidélité satisfaisante. Quant à l'approche de l'évaluation normative, ces techniques d'analyse des items ont réalisé certaines améliorations de la fidélité. Cependant, ces améliorations n'ont pas été satisfaisantes, car inférieures à .80 ; seule la technique d'analyse des facettes a permis de réaliser des améliorations plus ou moins remarquables que les techniques d'analyse de la psychométrie classique.

Mots clés : techniques d'analyse des items, optimisation de mesure, généralisabilité, évaluation critérielle, évaluation normative.

Abstract.

Classical psychometrics proposes reducing the length of a school measurement instrument by selecting items in order to optimize its reliability. As for the theory of generalizability; either extending the measurement instrument, or reducing the instrument, or limiting the scope of the conclusion only to the instruments considered in the study are the ways to optimize measurement reliability.

In this article, we wanted to verify whether by reducing the measurement instruments by selecting items using certain item analysis techniques, we can optimize reliability. Six different item selection techniques in particular; the F. DAVIS technique, the J. BROWN technique, the GRUNLUND technique, the RIGAUX technique, the item test correlation technique, the facet analysis technique; these allowed us to constitute five different tests.

After analysis, it is found that the calculated reliability coefficients (generalizability) are satisfactory in the criterion-referenced assessment approach, although no item analysis technique has allowed remarkable optimization, because the initial test already had satisfactory reliability. As for the normative assessment approach, these item analysis techniques have

*KAKULE MUKULU EVARISTE est enseignant chercheur à l'Université Officielle de Ruwenzori à BUTEMBO/NORD KIVU en République Démocratique du Congo.

(ISSN: 2805-413X) Evariste KAKULE MUKULU[1].*

https://ijojournals.com/ Volume 07 || Issue 08 || August, 2024 ||

achieved some improvements in reliability. However, these improvements were not satisfactory, because they were less than .80; only the facet analysis technique has allowed more or less remarkable improvements than the analysis techniques of classical psychometrics.

Key words: ***item analysis techniques, measurement optimization, generalizability, criterion-referenced assessment, normative assessment.***

I. INTRODUCTION.

Selon De KETELE, cité par CARDINET (1986, p5), l'évaluation est reconnue actuellement comme un des points d'entré privilégié de l'étude du processus d'enseignement apprentissage. Aborder le problème de l'évaluation, c'est nécessairement touché à tous les problèmes fondamentaux de la pédagogie. Mais l'élaboration des instruments de mesure pour une évaluation objective soufre encore de certains maux.

Des chercheurs de plus en plus nombreux ont démontré les sources d'erreurs dans le processus d'évaluation traditionnelle (F. BACHER, 1969) distingue trois sources d'erreur :

La première source d'erreurs est due à l'évaluateur lui-même, lui qui note autour d'une moyenne plus ou moins élevée et qui disperse plus ou moins ses notes au tour de cette moyenne.

La seconde source d'erreurs tient au choix du sujet même des examens traditionnels (matière), le choix du contenu de la matière, le nombre des questions par rapport au contenu global. Ainsi il est peu satisfaisant de généraliser à l'ensemble des candidats une constatation ponctuelle fondée sur l'un seulement des innombrables sujets (matières) qui auraient pu lui être proposé.

La question peut être mal posée, et donner des confusions, et les réponses à cette question seraient différentes. La manière de poser la question, de présenter le problème servant à l'évaluation, va influencer sur la réponse. Des études en psychologie cognitive «Elisabeth LOFTUS» (2003), en donnent des explications claires.

Normand BAILLARGEON, dans «Petit cours d'autodéfense intellectuelle» ; présente un film d'un accident à plusieurs personnes, et demande la vitesse à laquelle s'est produit le choc et s'il y avait des bris de vitre. Quand elle utilise le mot «fracassé » (smashed) dans l'énoncé de la question ; la vitesse est estimée plus rapide que si elle utilise le mot «percuté» (hit) dans l'énoncé de la question. Comment donc coter les sujets à des questions pareilles ? Quand l'énoncé peut induire en erreur :

-il peut comporter des erreurs ; comment alors évaluer la réponse à une question erronée ?

-l'énoncé peut être inadapté à la formulation d'un niveau trop simple ou au contraire trop élevé, ou bien présentant une situation que l'apprenant ne peut pas gérer, car les connaissances (savoir), savoir-faire, ou savoir être nécessaires ne font pas partie des prérequis à l'examen.

(ISSN: 2805-413X)
Evariste KAKULE MUKULU[1].*

https://ijojournals.com/ Volume 07 || Issue 08 || August, 2024 ||

La troisième source d'erreurs peut venir également des élèves. D'un jour à l'autre, d'un moment à l'autre se produisent des fluctuations aléatoires de la capacité qu'il s'agit d'évaluer ou mesurer. De plus, une certaine forme d'évaluation peut défavoriser de façon systématique un type d'élève.

C'est ainsi que MISENGA (1988), ainsi que MASANDI(2016), signalent deux qualités essentielles qui sont exigées d'un instrument de mesure ou d'évaluation en éducation : la validité ou la justesse et la fidélité. La première exigence est qualitative ; elle cherche à répondre à la question de savoir si : l'instrument utilisé pour l'évaluation, mesure-t-il réellement ce pourquoi il a été construit et rien que cette caractéristique-là? C'est la capacité de l'instrument à donner la valeur vraie de la grandeur mesurée. Par opposition à cette exigence purement qualitative, la seconde se rapporte plutôt à un aspect quantitatif. Elle correspond à la capacité d'un instrument d'observation à fournir pratiquement la même mesure pour un même objet d'étude si on estime que l'attribut mesuré n'a pas changé pendant l'intervalle de temps qui sépare les deux prises de mesure.

Il faut remarquer que l'étude et l'amélioration de la fidélité opposent la psychométrie classique à la théorie de généralisabilité. L'étude de la fidélité dans la psychométrie classique se fait essentiellement à partir de la corrélation entre deux séries des mesures aussi équivalentes que possibles. Pour la théorie de généralisabilité, la corrélation interclasse (intra classe) est étroitement associée au calcul de la fidélité (CARDINET et TOURNEUR, 1985, p20).

Au niveau de l'amélioration de la fidélité d'un instrument de mesure, la psychométrie classique recommande d'allonger l'épreuve et de sélectionner les items. Cette sélection des items conduit à éliminer certaines questions peu discriminantes. Mais cette sélection intervient avant l'étude de la fidélité. Cette procédure paraît contradictoire avec la sélection car les questions sont sélectionnées soigneusement, elles ne sont plus sources des fluctuations d'échantillonnage.

Dans la théorie de généralisabilité, trois procédures nous sont offertes pour optimiser la fidélité de mesures : soit il faut procéder par l'allongement des instruments de mesure, soit encore limiter la portée de la conclusion seulement aux instruments considérés dans l'étude, soit enfin éliminer les éléments atypiques des instruments de mesure ou des objets d'étude pour l'analyse des facettes.

Ainsi donc, pour optimiser la fidélité de mesure dans le sens de la réduction des instruments de mesure, les techniques de sélection des items peuvent être exploitées ; mais après l'étude de fidélité.

Considérant une épreuve quelconque, à laquelle peuvent être appliquées certaines techniques d'analyse d'items, celles-ci nous permettent-elles d'améliorer la fidélité de l'épreuve réduite ?

Laquelle des différentes techniques d'analyse d'items optimise t- elle au mieux la fidélité dans les approches critérielle et normative d'évaluation ?

(ISSN: 2805-413X) Evariste KAKULE MUKULU[1].*

https://ijojournals.com/ *Volume 07 || Issue 08 || August, 2024 ||*

Pour répondre à cette préoccupation, notre recherche s'appuie sur l'hypothèse fondamentale selon laquelle les différentes épreuves constituées des items sélectionnés, auront une fidélité supérieure à celle de l'épreuve de départ. Notre hypothèse secondaire stipule que la technique d'analyse des facettes serait la mieux adaptée dans les deux approches d'évaluation pour optimiser la fidélité de mesure.

II. METHODOLOGIE.

Avant de présenter la méthodologie proprement dite, nous allons d'abord passer à la clarification conceptuelle pour permettre aux lecteurs d'avoir une bonne compréhension de notre objet d'étude ; ensuite nous allons présenter comment avait évoluéla théorie de l'évaluation, et nous allons présenter quelques techniques d'analyse des items, pour finir avec les considérations expérimentales de l'étude.

2. 1. Clarification conceptuelle.

Souvent, dans le langage pédagogique, nous utilisons certains concepts clés tels que : item, mesure, évaluation, optimisation et psychométrie que nous allons présenter dans la suite.

a. Item.

Selon PIERON (1968, p.230) l'item signifie un article, un élément, une question d'un test. Le sens que nous utilisons ici pour le concept item, est celui de question d'une épreuve.

Pour d'autres auteurs, le mot item signifie une difficulté quelconque à résoudre. cette difficulté est donnée à titre expérimental pour voir comment un individu pourra réagir devant elle.

Une question que l'enseignant pose à l'élève dans le cadre du processus didactique établit un certain stimulus et attend une réaction de la part de l'élève. cette question est dite item lorsqu'elle se trouve présentée dans une situation de test.

Dans une épreuve, les items peuvent prendre plusieurs formes. Selon De KETELE (1984, pp.15 16) on peut classer les différentes formes d'items de multiples façons. Comme critère de classification, nous pouvons prendre l'opération exigée pour répondre : ou bien le répondant doit produire lui-même sa réponse, ou bien il doit choisir parmi plusieurs solutions possibles.

On peut combiner dans un même item les deux opérations ; ainsi il y a à distinguer deux principaux types d'items avec des catégories différentes qui sont : les items à production à réponse courte qui consistent en des questions classiques de messages à compléter et des textes lacunaires. Les items de sélection quant à eux, se distinguent en items vrai faux, items à choix multiple et items d'appariement.

b. Mesure.

En éducation, la mesure prend le sens du processus de contrôle de l'acquisition et de l'apprentissage. ainsi pour DOTTRENS (1971, pp.123 125), la mesure signifie en langage scolaire, prendre toute précaution pour que la note attribuée exprime objectivement la valeur d'un travail ou celle d'un

(ISSN: 2805-413X) Evariste KAKULE MUKULU[1].*

https://ijojournals.com/ *Volume 07 || Issue 08 || August, 2024 ||*

comportement. C'est aussi agir objectivement en éliminant tous risque d'erreurs.

Pour De KETELE (1984, p.8) mesurer signifie attribuer des nombres à des choses selon des règles bien déterminées et en respectant le principe d'isomorphisme (similarité entre les propriétés des objets mesurés et les propriétés du système de mesure).

Au sens strict du terme, une mesure sera toujours quantitative ; au sens large, elle pourra être qualitative, dans ce cas, les chiffres n'ont pas la valeur des nombres, car ce nombre ou ce chiffre n'est que simple étiquette. Ainsi l'évaluation peut recourir à une mesure, mais la mesure n'est pas nécessairement un acte d'évaluation.

c. Evaluation.

Le concept évaluation peut être utilisé dans des contextes différents et ces contextes lui confèrent des significations diverses telles que : ***appréciation***, calcul, estimation, examen, détermination, ***jugement***, mesure, etc.

Selon DOTTRENS (op. Cit., p.123), dans le concept d'évaluation deux idées se trouvent combinées : l'une est celle de mesure et l'autre celle d'appréciation.

Pour De KETELE (op. Cit., p.9), évaluer signifie examiner le degré d'adéquation entre un ensemble d'information et un ensemble de critères adéquats à l'objectif fixé en vue de prendre une décision.

Dans la littérature pédagogique, plusieurs définitions ont été élaborées. Mais il ressort que toutes tournent autour de deux concepts complémentaires qui sont : le contrôle et l'appréciation des connaissances acquises par les élèves. Ce contrôle et cette appréciation sont guidés par les objectifs déterminés au départ en vue de détecter les insuffisances et les faiblesses dans l'acquisition d'une part et d'y apporter remède d'autre part.

Pour THERER(1985), l'évaluation est une appréciation quantitative et ou qualitative d'un apprentissage en fonction d'objectifs préalablement définis, en fonction d'une décision à prendre. C'est pourquoi, il faut retenir que :

- L'évaluation transcende la simple mesure des capacités parce qu'elle peut être qualitative et diagnostique.
- L'évaluation suppose des objectifs d'apprentissage explicites, sinon il n'y aurait rien à mesurer.
- L'évaluation concerne tout autant l'enseignant que l'apprenant, elle vise à optimiser les apprentissages.
- L'évaluation constitue une aide à la décision ; sans elle, elle superflue.

d. Optimisation.

(ISSN: 2805-413X) Evariste KAKULE MUKULU[1].*

https://ijojournals.com/ Volume 07 || Issue 08 || August, 2024 ||

D'après le Larousse (1996), l'optimisation est l'action de placer dans les meilleures conditions d'utilisation, de rendement, résultat de cette action.

Dans le domaine de l'évaluation, l'optimisation de mesure est le fait de mettre au point un dispositif de mesure optimal, permettant de fonder des décisions ultérieures ou en tout cas d'ajuster au niveau désiré la précision de la fidélité (CARDINETet TOURNEUR, 1985, p.326).

Pour CARDINET, il n'est possible d'optimiser que lorsque l'on connaît l'importance relative des diverses sources de variances d'échantillonnage. ainsi on peut modifier le dispositif d'observation, d'estimation et de mesure pour mieux contrôler les fluctuations indésirables.

En résumé, par optimisation nous entendons l'amélioration de la fidélité de mesure.

e. Psychométrie.

Dans le langage psychologique, on appelle psychométrie, l'ensemble des méthodes de mesures utilisées en psychologie. Elle comprend la sensorimétrie et constitue la psychotechnique (PIERON,1968, p.353).

Selon MUCCHELLI (1985, p90), la psychométrie est l'ensemble des méthodes de mesure appliquées aux fonctions psychologiques et aux comportements humains.

La théorie psychométrique relève de SPEARMAN et BROWN qui, en s'intéressant dès le début du vingtième siècle au problème des mesures, ont donné le coup d'envoi au courant de recherche désigné aujourd'hui sous le nom de psychométrie classique. Dans ce courant de recherche, les auteurs ont surtout porté attention sur l'estimation et la réduction de la part que prennent les sources d'erreurs dans la prise de mesure.

2.2. Développement de l'évaluation.

La pratique évaluative fait partie intégrante de notre culture, surtout scolaire. C'est pour quoi vous entendrez parler aujourd'hui par exemple : d'hôtels à 3 ou à 5 étoiles, à des biens de consommation on attribue des labels, pour prouver davantage que l'évaluation reste incontournable. Et pourtant depuis les années 1968, certaines doctrines pédagogiques prônaient la suppression pure et simple des examens, pour vouloir démontrer comment les problèmes soulevés par l'évaluation sont multiples et interpellent, non seulement les formateurs, mais aussi les psychologues, les statisticiens, les moralistes,…. Et ceci nousa permis de distinguer deux conceptions différentes de l'évaluation : la conception classique et la conception moderne.

a. Selon la conception classique de l'évaluation.

Dans la pédagogie traditionnelle, le rôle des examens est la sélection des élèves ou étudiants. Cette sélection s'opère à partir d'épreuves qui privilégient les performances restitutives des candidats. Les

(ISSN: 2805-413X) Evariste KAKULE MUKULU[1].*

https://ijojournals.com/ Volume 07 || Issue 08 || August, 2024 ||

notes élevées restent exceptionnelles et minoritaires. Les épreuves sont d'ailleurs construites pour aboutir à un tel classement. Une telle évaluation est dite normative ou normée.

b. Selon la conception nouvelle de l'évaluation.

En pédagogie moderne, l'essentiel n'est pas la sélection d'une minorité de surdoués, mais bien la promotion de tous. Un maximum d'apprenants doit pouvoir maitriser un maximum d'objectifs d'apprentissage. En quelque sorte, l'enseignant doit combattre la distribution normale caractérisée par une faible taux de résultats supérieurs. Dans ce sens, on parle de la pédagogie de maitrise et de l'évaluation critériée. L'évaluation devient partie intégrante de l'apprentissage ; elle constitue une information qui aide l'apprenant à progresser et qui permet à l'enseignant d'améliorer ses propres interventions. Cette évaluation postule des exigences précises fondées sur des objectifs comportementaux explicites, c'est-à-dire des compétences dument définies. Cette évaluation suppose aussi une différenciation pédagogique qui respecte la démarche mentale (style cognitif) et le rythme de travail des apprenants. Cette nouvelle conception de l'évaluation offre d'indéniables avantages :

- réduction sensible du nombre d'échecs ; en principe 80% des apprenants devraient maitriser 80% des objectifs prévus.
- motivation accrue ; l'apprenant entre en compétition avec lui-même plutôt que de rivaliser avec les condisciples.
- importance accrue des savoir-faire par rapport aux simples savoirs.

2.3. Etude et optimisation de la fidélité des mesures scolaires.

L'étude et l'amélioration de la fidélité des mesures ont considérablement évolué, de la psychométrie classique à la théorie de la généralisabilité.

2.3.1. Etude de la fidélité de mesure.

Dans la littérature psychométrique, la fidélité des mesures est construite sur la théorie de SPEARMAN(1904). Ce dernier étudia la fidélité en partant d'une série de mesures d'un même trait.

L'étude et le contrôle de la fidélité d'une épreuve sur l'estimation et la réduction de la part d'erreurs due aux diverses sources dans la prise de mesure, n'a pas été facile à faire. C'est ainsi que les psychométriciens classiques ont proposé un certain nombre des procédures pour estimer la fidélité de mesure. De ces procédures nous citerons : la méthode de test retest, la méthode de GUILFORD, la méthode de bipartition du test (ou the split half method), la méthode de KUDER et RICHARDSON, la méthode de correction par plusieurs juges et la méthode de l'analyse de variances. Chacune de ces méthodes propose une procédure à suivre propre à elle.

2.3.2.**Optimisation de la fidélité.**

(ISSN: 2805-413X) Evariste KAKULE MUKULU[1].*

https://ijojournals.com/ *Volume 07 || Issue 08 || August, 2024 ||*

Notons que pour CARDINET ET TOURNEUR (op. cit. p.17), la psychométrie classique est fondée sur les postulats suivants :

* les erreurs ont une moyenne nulle, une variance uniforme, sont indépendantes les une des autres et indépendantes du score vrai.
* tous les tests parallèles ont même moyenne et mêmes variances vraie et d'erreur.

De ces postulats découle la formule de l'amélioration de la fidélité en fonction de l'accroissement du nombre de mesure. La première voie de l'optimisation de la mesure est donc l'allongement du test ; l'autre voie ou procédure consiste à sélectionner les items non discriminatifs.

2.4. La théorie de généralisabilité.

Le recourt au modèle de l'analyse de la variance pour l'estimation de la fidélité des épreuves a ouvert des nouvelles possibilités à la théorie de l'évaluation. Des auteurs comme CRONBACH et ses collaborateurs (1964, 1965, 1969,1972) ont forgé un cadre conceptuel et une méthodologie qui assurent une meilleure exploitation de ce modèle d'analyse de variance (MISENGA, 1988). Ce cadre conceptuel est celui de la généralisabilité.

La généralisabilité selon De LANDSHEERE (1979, p.132), est le degré auquel on peut généraliser une observation particulière à la valeur théorique rechercher. En effet, lorsqu'on utilise un instrument de mesure, le score que l'on obtient n'a d'intérêt que s'il nous renseigne au moins sur la valeur attendu d'autres mesures prises dans les conditions identiques. C'est cette exigence que l'on appelait traditionnellement la fidélité. Il doit nous informer aussi sur les comportements attendus du sujet dans les classes de situations plus larges que celles qui ont été observées. C'est ce qu'on appelait la validité. Les deux concepts (fidélité et validité) se fondent en un seul, celui de généralisabilité (CARDINET et TOURNEUR, op. cit. p.322).

L'étude de la généralisabilité comprend en principe quatre étapes principales à savoir : le plan d'observation, le plan d'estimation, le plan de mesure et le plan d'optimisation que MASANDI MILONDO (2016, p139) présente en deux étapes regroupées deux à deux. Toutefois les formules utilisées dans le processus de généralisabilité varient selon l'optique de l'évaluation envisagée.

a. Plan d'observation.

Au niveau de l'analyse, **le plan d'observation** est simplement descriptif. Il identifie les facettes, leurs interrelations et le nombre des niveaux échantillonnés sur chaque facette. A l'aide des procédures d'analyse de la variance, on calcule le carré moyen de chaque source de variation du plan utilisé.

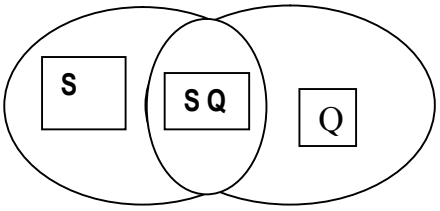
Ainsi pour aborder l'étude de la fidélité de l'épreuve concernée par cette étude, nous tenons à informer d'avance que la matrice de données dont nous disposons comporte les résultats de 150 sujets à 7

(ISSN: 2805-413X)
Evariste KAKULE MUKULU[1].*

https://ijojournals.com/ Volume 07 || Issue 08 || August, 2024 ||

questions. Compte tenu de ces données, nous avons distingué deux facettes à savoir : la facette « sujets » et la facette « questions ». Comme signaler plus haut, dans l'analyse de la fidélité par l'étude de la généralisabilité, le nombre de facettes retenues ainsi que les relations qui les unissent permettent de choisir un plan déterminé. Pour notre cas, nous utilisons un plan plus simple qui consiste à croiser tous les sujets avec toutes les questions de l'épreuve (S x Q). Schématiquement ce plan comporte trois sources de variations : les sujets «S», les questions «Q » et l'interaction sujets-questions «SQ » qui se présente comme suit :

Diagramme n° 1. Plan d'observation S x Q.



b. Plan d'estimation.

Au niveau plan d'estimation, on cherche à préciser la taille des populations et des univers, le nombre de niveau admissibles, le mode de sélection des données observées qu'il s'agisse de l'échantillonnage purement aléatoire, aléatoire fini ou tout simplement fixe. Cette seconde phase permet ainsi de déterminer la variance de chaque source de variation dans le dispositif.

A ce niveau d'estimation, il est recommandé dans une étude de généralisabilité de travailler uniquement avec des composantes de variances exprimées selon le modèle mixte. Cependant, il est plus souhaitable de calculer d'abord les composantes de variances aléatoires, quittes à les transformer en composantes mixtes. Mais il faut remarquer que le passage des composantes aléatoires aux composantes mixtes dépend de la nature de l'échantillonnage utilisé pour choisir les niveaux de chaque facette du plan. (MISENGA 1988 p.81).

c.Plan de mesure.

Le plan de mesure explicite le rôle de chaque facette dans la mesure et classe celle-ci suivant le mode d'échantillonnage et son rôle en quatre types : les facettes de sondage (S) ou facette de différenciation aléatoire, les facettes d'examen (E) ou facette de différenciation fixées, les facettes de contrôle (C) ou facette d'instrumentation fixées, et les facettes de généralisation (G) ou facette d'instrumentation aléatoire. Ces types sont définis d'après le mode d'échantillonnage de la facette et son rôle : l'appartenance à l'objet d'étude ou aux instruments de mesure.

Etant donné que dans notre étude nous devons considérer seulement deux facettes (les sujets et les questions), lorsque nous devons considérer les questions comme objets d'étude, les sujets constituent

(ISSN: 2805-413X) Evariste KAKULE MUKULU[1].*

https://ijojournals.com/ Volume 07 || Issue 08 || August, 2024 ||

à leur niveau, l'instrument de mesure (il s'agit dans ce cas d'une épreuve critérielle) et lorsque nous devons considérer les sujets comme objets d'étude, les questions deviennent l'instrument de mesure (il s'agit d'une épreuve normative).

C'est pour cela que dans cette étude, nous devons considérer deux plans de mesure : le plan de mesure pour l'épreuve critérielle et le plan de mesure pour l'épreuve normative.

Ce plan nous amène au calcul du coefficient de généralisabilité qui en effet, permet de déterminer la précision de la différenciation (classement) que l'on obtient sur la facette de différenciation lorsqu'on accepte des fluctuations sur la facette d'instrumentalisation (CARDINET et Tourneur, op. cit., p.313).

La formule qui permet de calculer ce coefficient de généralisabilité s'exprime en fonction de la variance introduite par les facettes de différenciation par rapport à la variance attendue des scores observés relatifs à ces facettes dans le dispositif choisi. Ainsi, la variance de différenciation est la variance qui est due à des différences véritables entre les objets que l'on étudie. Elle est une estimation des scores univers. Elle est formée des composantes des variances attribuables aux différences entre les scores univers des objets d'étude. La variance de généralisation quant à elle, est la variance d'erreur introduite par les fluctuations d'échantillonnage qui interviennent dans les conditions de mesure. Elle peut être définie de deux façon selon le type de différenciation que l'on désire (CARDINET et al, 1976, cité par Misenga, 1988, p.23). il y a une variance d'erreur absolue et une variance d'erreur relative.

Selon qu'on s'intéresse à une mesure absolue ou relative, on peut donc calculer pour chaque plan de mesure deux coefficients de généralisabilité : le coefficient de généralisabilité absolu et le coefficient de généralisabilité relatif. Ces coefficients sont déterminés par les ***équations suivantes*** :

$$\hat{E}_{P^2}(\Delta) = \frac{\delta^2{}_{(\tau)}}{\delta^2{}_{(\tau)} + \delta^2{}_{(\Delta)}} \quad \text{et} \quad \hat{E}_{P^2}(\delta) = \frac{\delta^2{}_{(\tau)}}{\delta^2{}_{(\tau)} + \delta^2{}_{(\delta)}}$$

Dont : $\hat{E}_{P^2}(\Delta) = coefficient.de.généralisabilité.absolu$

$\hat{E}_{P^2}(\delta) = coefficient.de.généralisabilité.relatif$

$\delta^2{}_{(\Delta)} = \text{var}\, iance.d'erreur.absolue$

$\delta^2{}_{(\tau)} = \text{var}\, iance.de.différenciation$

$\delta^2{}_{(\delta)} = \text{var}\, iance.d'erreur.relative$

Il convient de noter que le coefficient de généralisabilité est un coefficient de corrélation intra classe.

1°. Plan de mesure pour l'épreuve critérielle.

(ISSN: 2805-413X)
Evariste KAKULE MUKULU[1].*

https://ijojournals.com/ Volume 07 || Issue 08 || August, 2024 ||

Les questions étant purement aléatoires, elles constituent de ce fait une facette de différenciation aléatoire ou facette de sondage. Ces sujets sont alors la facette de généralisation ou d'instrumentation aléatoire. Ce plan de mesure peut être symbolisé comme suit : M (Q/- /- / S)

La question que nous nous posons ici est de savoir si l'épreuve permet de différencier fidèlement les questions. Pour différencier les questions de façon relative et absolue, nous avons fait recours aux coefficients de généralisabilité relative et absolue.

Pour le faire, nous allons utiliser les formules de la variance de différenciation et de la variance d'erreur que nous proposent CARDINET et TOURNEUR (1985, pp. 127 128).

2°. Plan de mesure pour l'épreuve normative.

Ce plan de mesure cherche à différencier les sujets quelles que soient les questions. Ce plan de mesure peut aussi être symbolisé comme suit : M(S/ / / Q).

Ainsi, nous allons calculer les coefficients de généralisabilité du classement des élèves suivant leur degré de réussite aux différentes questions. Procédant au calcul des variances de différenciation et d'erreur à l'aide des formules appropriées

d. Plan d'optimisation.

Après le plan de mesure vient enfin **le plan d'optimisation** qui consiste à établir sur base des informations obtenues aux phases précédentes, un dispositif optimal pour un type de décision souhaitée. A ce niveau, on peut envisager par exemple, des modifications, soit du plan d'observation, soit celui d'estimation, soit de plusieurs de ceux ci à la fois.

Ainsi,comme pour notre étude, nous voulons savoir quelle serait la procédure à utiliser pour rendre notre épreuve plus fiable?

A cette question nous avons essayé de répondre en nous référant à la fois à la thèse formuléepar la théorie de la généralisabilité qui stipule que pour optimiser la fidélité des instruments de mesure, on peut procéder par la réduction de ces instruments de mesure et au postulat de la psychométrie classique qui recommande la sélection des items pour améliorer la fidélité de mesure.

Comme nous venons de le dire ci-dessus, l'optimisation de mesure dans le modèle de la généralisabilité se fait dans plusieurs voies. On peut soit éliminer les niveaux atypiques des objets d'études et de l'univers de généralisation de mesure, on peut même fixer les niveaux des conditions sur les quelles porte la mesure.

Voilà pourquoinous devons procéder à l'application des différentes techniques d'analyse d'items qui vont nous permettre d'éliminer certaines questions et donc de réduire l'épreuve (instrument de mesure).

(ISSN: 2805-413X) Evariste KAKULE MUKULU[1].*

https://ijojournals.com/ Volume 07 || Issue 08 || August, 2024 ||

2.5. Quelques techniques d'analyse des items.

En psychométrie classique, l'analyse des items consiste à examiner la contribution apportée par chaque item de l'épreuve dans le résultat général.

De ce fait, elle permet de réduire le nombre des questions relatives à l'épreuve en mettant en évidence les items dont le degré de difficulté et celui de discrimination réponde mieux aux critères.

Selon MEURIS (1965 pp. 17-19) pour faire cette analyse, deux méthodes peuvent être utilisées : l'analyse quantitative qui consiste à demander aux sujets de décrire les processus psychologiques qui accompagnent la solution des problèmes posés et d'énoncer en détails le cheminement des processus mentaux ; l'autre méthode consiste à calculer les degrés de difficulté et de discrimination de chaque item ; c'est la méthode quantitative (méthode ou technique de la psychométrie classique).

Parmi les techniques de la psychométrie classique, nous avons considéré : la technique de F. DAVIS, la technique de James BROWN, la technique de GROUNLUND, la technique de RIGAUD et la technique de Corrélation item test. Et en plus nous avons considéré la technique d'analyse des items par la théorie de généralisabilité ou analyse des facettes proposée par CARDINET et TOURNEUR.

2.5.1. La technique de F. DAVIS.

a. **Fondement de la technique**.

La technique de DAVIS repose sur deux notions de base à savoir, le degré de difficulté et le pouvoir discriminatoire des items. La difficulté de l'item est déterminée ici en fonction du groupe expérimental. Quant au pouvoir discriminatoire de l'item, il peut être considéré comme sa capacité à distinguer les sujets forts des faibles (BAMWISHO, 1972 p. 41)

b. **Procédure.**

L'analyse des items par la technique de DAVIS se réalise en quatre étapes principales :

La première étape revient à déterminer les groupes supérieur et inférieur à partir des 27% du groupe total.

Dans la seconde étape, on calcule pour chaque item le pourcentage de réussite du groupe supérieur qui est symbolisé par PH et le pourcentage de réussite du groupe inférieur qui est symbolisé par PL. Ces pourcentages sont calculés par les formules suivantes :

$PH = \frac{RH}{NH} * 100$

$PL = \frac{RL}{NL} * 100$

RH= nombre de sujets du groupe supérieur ayant répondu correctement à l'item

NR= nombre de sujets dans le groupe supérieur.

(ISSN: 2805-413X) Evariste KAKULE MUKULU[1].*

https://ijojournals.com/ *Volume 07 || Issue 08 || August, 2024 ||*

RL= nombre des sujets du groupe inférieur ayant répondu correctement à l'item.

NL= nombre des sujets dans le groupe inférieur.

A partir des pourcentages calculés (PH et PL), on aborde la troisième étape qui consiste à déterminer le degré de difficulté et le pouvoir discriminatoire de l'item à partir de la table des valeurs critiques élaboré par DAVIS.

Enfin, à la quatrième étape, la décision se prend sur l'acceptation ou le rejet de l'item. Pour ce faire, on se sert de l'intervalle de confiance tant bien pour l'indice de difficulté que pour le pouvoir discriminatoire. A cet effet, plusieurs auteurs ont proposé des limites de sélection différentes.

Pour DAVIS, un degré de difficulté acceptable varie toujours entre la limite de 0.25 et 0.75 ; le pouvoir discriminatoire varie de 0.20 à 1.00. D'autre part DOROTHY propose l'intervalle de 0.15 à 0.85 pour le degré de difficulté. De ces limites de sélection, nous allons nous servir chaque fois de celle proposées par DAVIS.

Toutefois au cours de l'analyse des items par la technique de DAVIS certains cas particuliers peuvent surgir et poser des difficultés «énormes. C'est le cas des valeurs PL ou PH négatives. Si PH et PL sont tous négatifs, on change les signes des indices obtenus et ces valeurs sont considérées comme positives. Dans le cas où PH est positif et PL négatif ou égale à zéro, on transforme PL en PL corrigé selon la deuxième table statistique de DAVIS (1966 p.39) . Les valeurs de PH qui sont supérieur à 0.99 ou égale à 1.00 doivent être corrigées avant la consultation de la table. Lorsque les valeurs de PH et PL sont de signes différents, c'est- à- dire PH positif et PL négatif ou PH négatif et PL positif, on peut obtenir des indices de difficultés corrigés en lisant la table IV de DAVIS (DAVIS 1966, p.40).

Mais à ce niveau, on recommande d'utiliser la moyenne des pourcentages de réponses correctes après correction du hasard dans le groupe supérieur.

2.5.2. La technique de JAMES BROWN.

a. Fondement.

La technique de J. BROWN repose sur les mêmes notions de base que celle de DAVIS, à savoir le degré de difficulté et le degré de discrimination des items. L'indice de difficulté de l'item est déterminé par le rapport du nombre des sujets qui ont répondu correctement à l'item, au nombre des sujets qui ont essayé d'y répondre (réussite ou échec). BROWN a introduit dans sa théorie la notion d'omission, c'est à dire les sujets qui ont sauté un item donné.

B .Procédure.

(ISSN: 2805-413X)
Evariste KAKULE MUKULU[1].*

https://ijojournals.com/ Volume 07 || Issue 08 || August, 2024 ||

La procédure de BROWN distingue deux cas : le premier concerne les échantillons de taille inférieure ou égale à 100 et le second concerne les échantillons grands c'est à dire de taille supérieure à 100.

Pour les petits échantillons, la technique se déroule en quatre étapes.

1. Premièrement, on ordonne les copies selon les mérites, ensuite on divise l'échantillon en trois parties égales et chaque groupe est constitué par 33% des sujets.
2. Deuxièmement, on met de côté le groupe moyen pour NE rester qu'avec les tiers supérieur et inférieur.
3. A la troisième étape, on cherche à déterminer le degré de difficulté et le pouvoir discriminatoire qui sont en appliquant les formules suivantes :

$$F_s = \frac{R_s}{T - U_s} \qquad D = F_s - F_i \qquad F_j = \frac{R_j}{R_j + w_j + O_j} \qquad F_j = \frac{R_j}{R_j + w_j + O_j}$$

$$F_i = \frac{R_i}{T - U_i} \qquad P_i = \frac{R_i}{TT_i - U_i}$$

Tel que :

Ou Fj =indice de difficulté

Rj = nombre de sujets des tiers supérieur et inférieur qui ont répondu correctement à l'item.

Wj = nombre de sujets qui ont donné une fausse réponse à l'item.

Oj = nombre de sujets qui ont sauté (omis) l'item.

D = indice de discrimination.

Fs = facilité de l'item dans le groupe supérieur.

Fi = facilité de l'item dans le groupe inférieur.

Rs = nombre de sujets qui ont répondu correctement à l'item dans le groupe supérieur.

T = nombre total de sujets du groupe (inférieur ou supérieur).

Us = nombre de sujets du groupe supérieur qui n'ont pas atteint l'item.

Ri = nombre de sujets qui ont répondu correctement à l'item dans le groupe inférieur.

Ui = nombre des sujets du groupe inférieur qui n'ont pas atteint l'item.

4. Enfin, on procède à la prise de décision pour l'acceptation ou le rejet de l'item. Pour ce faire, on se sert de l'intervalle de confiance souhaité par le chercheur Ainsi pour notre cas nous nous servirons de même indice pour l'intervalle de confiance proposé par DAVIS.

Dans le deuxième cas des échantillons grand ; c'est à dire avec plus de 100 sujets, la

(ISSN: 2805-413X) Evariste KAKULE MUKULU[1].*

https://ijojournals.com/ Volume 07 || Issue 08 || August, 2024 ||

procédure d'analyse se déroule en étapes suivantes :

i. On classe les copies dans l'ordre de mérite.
ii. Ensuite on divise les copies en trois groupes égaux.
iii. On détermine le nombre des sujets des groupes supérieur, inférieur et moyen.
iv. On calcule l'indice de facilité de l'item par la formule suivante :

$$F = \frac{R}{T - U}$$

v. On calcule la proportion de réussite du tiers supérieur du tiers supérieur par la formule :
vi. On calcule On calcule également la proportion de réussite du tiers inférieur par la formule suivante :

$$Ps = \frac{R_s}{T_s - U_s}$$

vii. On calcule ensuite la proportion de réussite de deux groupes de tiers supérieur et moyen ensemble Cette proportion est déterminée par la formule ci-après :

VII Enfin, on calcule la proportion de r 130ussite de deux tiers moyen et inférieur par la formule suivante :

$$P_{sm} = \frac{R_{sm}}{(T - - U_s) + (T - U_m)} \qquad P_{mi} = \frac{R_{mi}}{(T - U_m) + (T - U_i)}$$

IX A partir des proportions de réussite calculées, on arrive à déterminer l'indice de discrimination grâce à la table de corrélation double tétrachorique. Pour entrer dans cette table, il faut considérer d'abord les proportions de réussite du tiers supérieur et de l'ensemble des tiers supérieur et moyen d'une part et les proportions de réussite du tiers inférieur et de l'ensemble des tiers moyen et inférieur d'autre part. Ensuite il faut prendre la moitié de la somme des résultats de la table, trouvés pour déterminer cet indice de discrimination.

X. Enfin, les indices trouvés sont appréciés selon les critères suivants :

- pour l'indice de discrimination, l'intervalle de confiance est de 0.40 à 1.00 comme proposé par l'auteur.

- pour l'indice de difficulté, il est apprécié de la même façon que chez DAVIS.

2.5.3. **La technique de GROUNLUND.**

a. Fondement.

Comme les deux techniques précédentes, la technique de GROUNLUND est fondée sur le calcul de l'indice de discrimination et de l'indice de difficulté. Au lieu d'utiliser les 27% supérieur et inférieur de la population d'enquête comme le recommande F. DAVIS, GROUNLUND lui, préfère travailler avec les tiers comme BROWN.

(ISSN: 2805-413X) Evariste KAKULE MUKULU[1].*

https://ijojournals.com/ Volume 07 || Issue 08 || August, 2024 ||

b. Procédure.

Pour calculer ces indices, l'auteur propose d'appliquer les formules suivantes :

1.

$$P = \frac{R}{T}.100$$

P = degré de difficulté de l'item

R = nombre de sujets des tiers supérieur et inférieur qui ont réussi l'item.

T = Total de sujets des tiers supérieur et inférieur ayant essayé l'item.

2.

$$IDS = \frac{R_u - R_l}{\frac{1}{2}T}$$

Où : IDS = indice de discrimination !

Ru = nombre de sujets du tiers supérieur ayant réussi l'item.

Rl = nombre de sujets du tiers inférieur ayant réussi l'item.

T = total de sujets des tiers supérieur et inférieur ayant essayé l'item.

Enfin, concernant les critères d'appréciation du degré de difficulté et de l'indice de discrimination, l'auteur n'a pas proposéun seul critère d'appréciation des résultats. Les critères dépendent de la nature et de l'objectif de la recherche.

2.5.4. TECHNIQUE DE RIGAUX.

a. Fondement.

Cette technique se fonde également sur les notions de difficulté et de discrimination. Toutefois, la difficulté n'est plus exprimée directement en termes de pourcentage de réussite à l'item. Selon l'auteur, l'utilisation des pourcentages de réussite ne donne pas une gradation suffisamment claire de la difficulté intrinsèque des items. Ainsi, l'auteur estime que les indices de difficulté permettent par contre, non seulement l'ordination des items sur base de leur difficulté réelle, mais encore, un calcul justifié de la difficulté de l'ensemble des items.

Concernant l'indice de discrimination, l'auteur le considère comme étant le coefficient de corrélation entre les réponses à un item déterminé et les réponses au test entier.

b. Procédure.

L'analyse des items par la technique de RIGAUX se réalise en trois étapes. La première consiste à déterminer le degré de difficulté de chaque item. Pour cela, on calcule d'abord le pourcentage de réussite à chaque item. Ensuite, avec ces pourcentages on entre dans la table critique de RIGAUX qui donne les indices de difficulté. Ces indices varient entre 0 et 100.

Deuxièmement, on passe à la détermination du pouvoir discriminatoire des items. Pour y

$$r_p bis = \frac{m_p - m_q}{\delta t}\sqrt{pq}$$

(ISSN: 2805-413X)
Evariste KAKULE MUKULU[1].*

https://ijojournals.com/ Volume 07 || Issue 08 || August, 2024 ||

parvenir, RIGAUX utilise le coefficient de corrélation point bis serial (rp. Bis.) dont la formule est la suivante :

M p = moyenne de la catégorie (1).

M q = moyenne de la catégorie (0)

P = proportion de réussite pour l'item considère.

Q = proportion d'échec pour l'item considéré.

Toutefois, il faut savoir que cette corrélation est biaisé par le fait que l'item considéré fait partie de l'épreuve à laquelle il est par ailleurs en corrélation. Ainsi, ce fait surestime le coefficient de corrélation r.p. bis obtenu par la formule ci haut.

A ce propos, L. D'HAAINAUT cité par MISENGA(1988) fait remarquer que « quand on cherche la corrélation entre l'échec ou la réussite d'un item et la note global au test dont l'item fait partie, on commet une erreur dite de recouvrement due au fait que la présence de l'item dans le test augmente artificiellement la corrélation de l'item avec le test. »

Etant donné que cette corrélation de l'item au test est biaisée, une formule de correction est proposée par l'auteur L. D'HAINAUT. Cette formule permet de réduire la part d'erreur qui fait augmenter considérablement la corrélation. Elle se présente de la manière suivante :

$$r_{12} = \frac{r_1 s - s_1}{s_1^2 + s^2 - 2 r_1 s s_1}$$

.r12 = coefficient de corrélation rpbis corrigé de la partie avec l'ensemble.

r1 = coefficient de corrélation rp bis biaisé.

S = écart - type pour l'ensemble des notes dans le test

S1= écart - type de la partie considérée

Apres le calcul de l 'indice de discrimination et du degré de difficulté, il convient de fixer dans la troisième étape le seuil de sélection des items.

Pour le degré de difficulté qui est exprimé ici en termes des rangs (Rg), nous avons maintenu l'intervalle de confiance proposé par DAVIS ; .25 à .75, ce qui correspond à l'intervalle de 37.00 à 63.00 des indices de RIGAUX. Tandis que pour les discriminations, nous avons recouru au test de signification des coefficients de corrélation rp bis qui est déterminé par le test 't' de student. Ainsi la signification de la corrélation rpbis se calcule par la formule suivante :

$$t = rpbis \frac{N-2}{1-R^2 PBIS}$$

OU t= test t de student.

(ISSN: 2805-413X)
Evariste KAKULE MUKULU[1].*

https://ijojournals.com/ Volume 07 || Issue 08 || August, 2024 ||

Rpbis = coefficient de corrélation point bis serial

N = nombre de sujets.

2.5.5. Technique de corrélation item - test.

a**. Fondement**.

La technique de corrélation item - test est un mode d'analyse qui a sélectionné les items sur base de leur sensibilité au pouvoir discriminatoire sera influence par la clarté dans la rédaction de l'item, le soin avec lequel on a évité les expressions équivoques et l'exactitude de la réponse correcte.

Comme l'exige cette technique, pour déterminer la corrélation item - test, plusieurs voies sont ouvertes parmi celles-ci ; nous pouvons distinguer le coefficient de corrélation bis serial, le coefficient de corrélation de Bravais Pearson et le coefficient Phi (ϕ). A cette liste nous pouvons joindre d'autres coefficients tels que le coefficient de rapport critique proposé par Alexandre (1971) et la corrélation point bis serial.

b. Procédure.

L'étude de la corrélation item - test se déroule en 3 étapes.

1. On calcul d'abord le coefficient de corrélation entre les réponses à chaque item et au test entier. Mais l'utilisation du coefficient d'une variable dichotomique et une variable métrique. C'est-à-dire la réponse à l'item doit être sanctionnée par la réussite ou l'échec sans une situation intermédiaire. Pour calculer cette corrélation, la formule se présente de la manière suivante :

$$Rpbis = \frac{Mp - Mq}{\delta t}\sqrt{pq}$$

tel que signalé dans la technique précédente, la corrélation item - test obtenu à partir des notes individuelles des sujets dans l'épreuve globale et la réussite ou l'échec des sujets à un item particulier de l'épreuve, souffre d'une erreur dite de recouvrement. Cette erreur est due au fait que l'item appartient à l'épreuve.

Comme signale par la technique précédente, une corrélation est apportée à cette erreur qui surestime la corrélation item - test.

Après calcul des coefficients de corrélation et la corrélation envisagée, on teste la signification de ceux-ci. Pour cela le test «t» nous est utile, il se présente comme suit :

$$t = rpbis \quad \frac{N-2}{1-R^2 PBIS}$$

Enfin, il convient de prendre une décision d'acceptation ou de rejet de l'item à partir d'un seuil donné. Pour plus de rigueur nous préférons travailler avec le seuil de 1% (0.1).

2.5.6. La technique d'analyse d'item par la théorie de l'analyse des facettes

(ISSN: 2805-413X) Evariste KAKULE MUKULU[1].*

https://ijojournals.com/ *Volume 07 || Issue 08 || August, 2024 ||*

(généralisabilité)

a. **Fondement.**

Se distinguant des techniques de la psychométrie classique qui sont fondées sur le degré de difficulté et l'indice de discrimination ; la technique d'analyse des facettes quant à elle, se fonde sur le model de l'analyse des variances.

L'objectif poursuivi par ce modèle est de sélectionner les données qui accroissent le degré de généralisabilité de l'épreuve. Cette sélection se fait d'abord au niveau des instruments de mesure qui paraissent applicables aux objets d'études et en suite au niveau des objets d'études aux quels peuvent s'appliquer les mesures proposées.

Procédure.

La sélection des données par la technique d'analyse des facettes se fait dans deux directions : d'abord sur la facette de généralisation et ensuite sur la facette de différenciation.

1° **Sélection des instruments de mesure ou des éléments de la facette de généralisation.**

Les éléments à considérer pour cette étape diffèrent selon la nature de l'épreuve. Dans tous les cas, il est d'abord question de dresser la matrice des données brutes en respectant les sujets sur lignes et les questions sur les colonnes. Ensuite il convient d'estimer les composantes d'interaction de chaque cellule de la matrice originale en soustrayant de chaque élément les moyennes de la ligne et de la colonne correspondante.

Dans un troisième temps, il s'agit de calculer la variance des résidus par colonne (pour l'épreuve normative), par ligne (pour l'épreuve critérielle).

En fin, il est question de rejeter les éléments de la facette de généralisation qui présentent des larges variances résiduelles et qui réduisent ainsi la généralisabilité des mesures.

2° **Sélection des objets d'études ou des éléments de la facette de différenciation.**

Le point de départ de cette seconde direction est constitué par la matrice des données brutes purifiées des éléments atypiques. L'analyse se fait ainsi mutatis - mutandis comme à la première direction.

L'avant dernière étape consiste à calculer les variances résiduelles des éléments de la facette de différenciation.

Dans la dernière étape, on rejette les données qui présentent des très larges variances résiduelles.Il convient de signaler ici que le rejet des données se fait à partir d'un seuil fixé par le chercheur lui-même compte tenu de la configuration des carrés moyens.

(ISSN: 2805-413X) Evariste KAKULE MUKULU[1].*

https://ijojournals.com/ *Volume 07 || Issue 08 || August, 2024 ||*

2.6. Considérations expérimentales.

a. Population d'enquête et échantillon d'étude.

Notre population d'enquête est constituée des élèves de 4ème année secondaire des sections scientifiques des écoles conventionnées protestantes de la CBCA. Cette population comprend 8écoles. Ces écoles organisent 13 classes de 4ème année au courant de l'année scolaire 2015- 2016. Ces classes sont fréquentées par 427 élèves.

Pour constituer notre échantillon, nous avons d'abord dressé une liste alphabétique de toutes les écoles secondaires de la CBCA/Butembo organisant la section scientifique. Etant donné que le nombre des classes organisées n'est pas le même dans les différentes écoles, nous avons attribué à chaque classe un numéro que nous avons ensuite déposé dans une urne. En appliquant la méthode de tirage au hasard, nous avons sélectionné 7 classes dans l'ensemble.

Notre échantillon idéal comprend 228 sujets. Alors que nous comptions travailler avec tous ces sujets, nous avons pu rencontrer le jour de passation de l'épreuve que 150 sujets.Les problèmes qui ont pu rabattre les effectifs de notre échantillon peuvent s'expliquer par le fait que, quand nous sommes arrivés dans les écoles pour la passation de l'épreuve, certains élèves étaient absents. Ces absences peuvent se justifier soit par le cas de maladie, soit par le cas d'exclusion temporaire pour raison de frais scolaire. Certains autres élèves ont tout simplement refusé de passer l'épreuve.

b. Construction de l'épreuve.

Après analyse du contenu du programme de physique au niveau de 4èmescientifique, nous avons consulté les prévisions des matières de certains professeurs. Dans le contenu du programme, nous avons relevé 27 unités de contenu de la matière. De ces unités de contenu, nous avons tiré au hasard 25 unités pour lesquelles nous avons construit 25 questions ouvertes portant sur les chapitres de thermométrie, de dilatation et de changement d'états de la matière.

Après réflexion et essai de réponse à cette première forme des questions, nous avons constitué 15 questions à choix multiples. Celles-ci ont été soumises à l'appréciation des professeurs du cours de physique qui, après les avoir critiquées, nous ont permis de ne considérer que 7 questions pour la forme finale de l'épreuve.Cette limitation de l'épreuve a été dictée par l'abondance d'analyse auxquelles nous allons procéder dans le traitement de données. Cependant il faut reconnaître par ce même fait, la réduction de la validité du contenu de notre épreuve.

c. Forme des questions.

Comme signalé ci haut, nous avons choisi pour notre épreuve les questions à choix multiples. Ce

(ISSN: 2805-413X) Evariste KAKULE MUKULU[1].*

https://ijojournals.com/ *Volume 07 || Issue 08 || August, 2024 ||*

type de questions bien qu'elles soient critiquées par les tenants de la méthode traditionnelle, leur emploi se justifie ici par la taille de notre échantillon et aussi par le souci d'atteindre l'objectivité dans la cotation.

d. Passation et cotation de l'épreuve.

La passation de l'épreuve a été réalisée au mois d'Avril. Elle durait 50 minutes.

Concernant la cotation, elle était binaire car chaque réponse exacte valait un point (1) et chaque fausse réponse (échec ou omission) valait zéro point (0).

III. RESULTAT.

A. Etude de la fidélité de l'épreuve.

L'étude de la fidélité de l'épreuve de départ s'est faite suivant le modèle de la généralisabilité.

En principe, l'étude de la fidélité par la généralisabilité se réalise en 4 étapes suivantes regroupées deux à deux : le plan d'observation, le plan d'estimation (analyse de variance), le plan de mesure et le plan d'optimisation (étude de généralisabilité proprement dite).

1°. Plan d'observation.

Pour aborder l'étude de la fidélité de notre épreuve, nous tenons à rappeler que la matrice de données dont nous disposons comporte les résultats de 150 sujets à 7 questions. Compte tenu de ces données, nous distinguons deux facettes à savoir : la facette « sujets » et la facette « questions ». Comme signaler plus haut, dans l'analyse de la fidélité par l'étude de la généralisabilité, le nombre de facettes retenues ainsi que les relations qui les unissent permettent de choisir un plan déterminé. Ainsi pour notre cas, nous avons utilisé un plan plus simple qui consiste à croiser tous les sujets avec toutes les questions de l'épreuve (S x Q). Schématiquement ce plan comporte trois sources de variations : les sujets «S», les questions «Q » et l'interaction sujets-questions «SQ » qui se présente comme suit :

Après calcul statistique de ces différentes sources de variation, nous avons abouti aux résultats que nous présentons au niveau de plan d'estimation.

2°. Plan d'estimation.

Ce plan consiste à estimer les variances de différents effets testés existant dans le plan d'observation. On supporte travailler avec toutes les données possibles des univers de généralisation et de différenciation.

Concernant l'échantillonnage de 150 sujets de notre enquête, ils ont été tirés aléatoirement d'une population finie tandis que les 7 questions sont supposées choisies également aléatoirement d'une population infinie.

Après calcul, les estimations obtenues pour les composantes de variances aléatoires et mixtes sont données dans le tableau ci-après.

(ISSN: 2805-413X)
Evariste KAKULE MUKULU[1].*

https://ijojournals.com/ Volume 07 || Issue 08 || August, 2024 ||

Tableau n° 1. Résultats des données des plans d'observation et d'estimation.

Source de variation	Somme des carrés	Dégrée de liberté	Carré Moyen	Variance aléatoire	Variance mixte.
Sujets S	34.69	149	.23	.007	.007
Questions Q	59.86	6	9.98	.065	.0654
SxQ	163.28	894	.18	.18	.18
Total	257.83	1049			

3°. Plan de mesure.

Le plan de mesure explicite le rôle de chaque facette dans la mesure. Comme signalé plus haut, on distingue généralement 4types de facettes dans l'étude de la généralité. Ces types sont définis d'après le mode d'échantillonnage de la facette et son rôle : l'appartenance à l'objet d'étude ou aux instruments de mesure.

Etant donné que dans notre étude nous avons considéré seulement deux facettes (les sujets et les questions), lorsque nous considérons les questions comme objets d'étude, les sujets constituent l'instrument de mesure (il s'agit dans ce cas d'une épreuve critérielle) et lorsque nous considérons les sujets comme objets d'étude, les questions deviennent l'instrument de mesure (il s'agit d'une épreuve normative). Ainsi dans notre étude, nous avons considéré deux plans de mesure : le plan de mesure pour l'épreuve critérielle et le plan de mesure pour l'épreuve normative.

a. Plan de mesure pour l'épreuve critérielle.

Les questions étant purement aléatoires, elles constituent de ce fait une facette de différenciation aléatoire ou facette de sondage. Ces sujets sont alors la facette de généralisation ou d'instrumentation aléatoire. Ce plan de mesure peut être symbolisé comme suit : M (Q/- /- / S)

La question que nous nous posons ici est de savoir si l'épreuve permet de différencier fidèlement les questions. Pour différencier les questions de façon relative et absolue, nous avons fait recours aux coefficients de généralisabilité relative et absolue.

Pour le faire, nous avons utilisé les formules de la variance de différenciation et de la variance d'erreur que nous proposent CARDINET et TOURNEUR (1985, pp. 127 128).

Après le calcul, les résultats suivants ont été obtenus pour notre épreuve.

Tableau n°2 Résultats du plan de mesure pour l'épreuve critérielle.

Sources de variation	Variance de différenciation	Variance d'erreur absolue	Variance d'erreur relative	Coefficient de généralisabilité absolu	Coefficient de généralisabilité relatif.

(ISSN: 2805-413X)

Evariste KAKULE MUKULU[1].*

https://ijojournals.com/ Volume 07 || Issue 08 || August, 2024 ||

S		.00003		.987	.988
Q	.0654				
S x Q		.00078	.00078		
Total	.0654	.00081	.00078		

En examinant les résultats obtenus, nous constatons que les coefficients de généralisabilité absolu et relatif sont supérieur au seuil minimum acceptable de .80 . Ces coefficients indiquent que notre dispositif permet de classer de façon fiable les questions de l'épreuve de physique quels que soient les élèves de 4ème année scientifique de Butembo.

b. Plan de mesure pour l'épreuve normative.

Ce plan de mesure cherche à différencier les sujets quelles que soient les questions. Ce plan de mesure peut aussi être symbolisé comme suit : M(S/ / / Q).

Ainsi, nous avons calculé les coefficients de généralisabilité du classement des élèves suivant leur degré de réussite aux différentes questions. Procédant au calcul des variances de différenciation et d'erreur à l'aide des formules appropriées, nous avons abouti aux résultats suivants :

Tableau n°3 Résultats du plan de mesure pour l'épreuve normative.

Sources de variation	Variance de différenciation	Variance d'erreur absolue	Variance d'erreur relative	Coefficient de généralisabilité absolu	Coefficient de généralisabilité relatif
S	.007			.16	.21
Q		.0093			
S x Q		.0257	.0257		
Total	.007	.0350	.0257		

En observant les résultats ci-dessus, nous constatons que notre épreuve ne permet pas de classer les élèves de façon fiable. Car les coefficients de généralisabilité absolue et relative sont inférieurs au seuil minimum de .80.

Après ce plan vient un quatrième, celui d'optimisation de mesure.

4°. Plan d'optimisation.

Le plan d'optimisation a pour but de préparer un dispositif optimal pour les décisions envisagées. Ainsi pour notre étude, nous voulons savoir quelle serait la procédure à utiliser pour rendre notre épreuve plus fidèle. A cette question nous avons essayé de répondre en nous référant à la fois à la thèse formulée

(ISSN: 2805-413X)
Evariste KAKULE MUKULU[1].*

https://ijojournals.com/ Volume 07 || Issue 08 || August, 2024 ||

par la théorie de la généralisabilité qui stipule que pour optimiser la fidélité des instruments de mesure, on peut procéder par la réduction de ces instruments de mesure et au postulat de la psychométrie classique qui recommande la sélection des items pour améliorer la fidélité de mesure.

Ainsi nous avons procéder à l'application des différentes techniques d'analyse d'items qui vont nous permettre d'éliminer certaines questions et donc de réduire le nombre de question dans l'épreuve (instrument de mesure).

L'application de ces six techniques d'analyse d'items aux mêmes données, en respectant les principes d'analyse de chaque technique, n'a pas abouti à la sélection de mêmes items. C'est ainsi que par exemplela première technique, celle de F. DAVIS a rejeté aucune question ;la technique de J. Brown a retenu quatre questions (2, 3, 5, 7), et en a rejeté trois ; et la technique de GROUNLUND a retenu les mêmes quatre questions (2, 3, 5, 7), que la technique de J. Brown. La technique de RIGAUXquant à elle, a retenu seulement deux questions (5, 7) et a rejeté toutes les autres.La technique de corrélation item test de même, a retenu quatre questions (1, 5, 6, 7) ; et la technique d'analyse des facettes dans l'approche d'évaluation critérielle a retenu 5 questions (1 3 4 6 7) et dans l'approche d'évaluation normative, elle a retenu six questions (1 3 4 5 6 7).

Ainsi, nous avons pu constituer 5 différentes épreuves selon les différentes techniques d'analyse d'items. Car la technique d'analyse des items de F. DAVIS n'a rejeté aucune question, elle a validé l'épreuve de départ toute entière.

B. **Etude de la fidélité des épreuves optimisées par les techniques d'analyse d'items.**

Telles que présentées dans la partie précédente, les différentes techniques à part celle de Davis qui a sélectionné toutes les questions, nous ont permis de mettre au point cinq différentes épreuves optimisées. L'étude de la fidélité de ces différentes épreuves s'est réalisée suivant le même modèle de la généralisabilité appliqué à l'épreuve du départ. Voici les différentes valeurs du coefficient de généralisabilité calculées pour toutes les épreuves optimisées.

Tableau n°4 Coefficients de généralisabilité des épreuves optimisées.

Approche de l'évaluation	Epreuve de départ	Technique de DAVIS	Technique de BROWN et GRUNLUND	Technique de RIGAUX	Technique de corrélation item test	Approche critérielle de l'analyse des facettes	Approche normative de l'analyse des facettes.

(ISSN: 2805-413X)
Evariste KAKULE MUKULU[1].*

https://ijojournals.com/ Volume 07 || Issue 08 || August, 2024 ||

	.98	.98	.97	.95	.99	.99	.99
	.98	.98	.97	.95	.99	.99	.99
	.16	.16	.07	.28	.22	.20	.44
	.21	.21	.08	.30	.32	.31	.57

IV. DISCUSSION DES RESULTATS

Il ressort de ces différentes valeurs calculées que, d'une manière générale, les coefficients de généralisabilité des épreuves optimisées sont supérieurs à ceux de l'épreuve de départ. Ceci répond affirmativement à notre première question et de même fait, nous affirmons notre première hypothèse selon laquelle les différentes épreuves constituées des items sélectionnés, auront une fidélité supérieure à celle de l'épreuve de départ.

Toutefois, l'épreuve optimisée par la technique de Brown et ou deGRUNLUND n'a apporté aucune amélioration mais plutôt elle semble sous-estimer la fidélité, si bien dans l'approche de l'évaluation critérielle que dans l'approche de l'évaluation normative. Il y a également l'épreuve optimisée par la technique RIGAUX qui a présenté ce même comportement sauf en évaluation normative.

Nous pensons que ces résultats obtenus peuvent être expliqué partiellement par le fait que la fidélité de l'épreuve de départ était déjà supérieur à .80 dans l'approche critérielle ; ainsi donc cette épreuve ne demanderait plus une quelconque optimisation car sa fidélité était déjà satisfaisante dans l'approche critérielle. D'autre part, l'épreuve de départ étant constituée de 7 questions seulement, elle est courte, l'amélioration de sa fidélité ne pouvait peut être obtenue que par la voie de l'approche d'allongement de fixation.

Concernant les coefficients de généralisabilité calculés pour une évaluation normative, nous avons constaté que les améliorations réalisées sont insatisfaisantes ; car aucune épreuve optimisée n'a réalisé un coefficient de généralisabilité supérieur ou égale à .80 qui constitue le seuil minimum d'acceptation des coefficients de généralisabilité. Cependant de toutes les techniques d'analyse d'items, seule la technique d'analyse de facettes a optimisé une épreuve qui a réalisé des coefficients de généralisabilité plus élevés que toutes les autres techniques d'analyse. Ceci, nous amène à affirmer partiellement notre deuxième hypothèse selon laquelle l'analyse des facettes serait adaptée dans les deux approches d'évaluation.

CONCLUSION.

Au terme de cet article sur les apports de quelques techniques d'analyse des items pour l'optimisation des mesures scolaires, dans lequel nous nous sommes intéressés à vérifier si les moyens

(ISSN: 2805-413X) Evariste KAKULE MUKULU[1].*

https://ijojournals.com/ *Volume 07 || Issue 08 || August, 2024 ||*

proposés d'une part par la psychométrie classique et la théorie de généralisabilité d'autre part, pour optimiser la fidélité de mesures scolaires offrent certaines garanties. Considérant que pour optimiser la fidélité de mesure, la psychométrie classique recommande entre autre la sélection des items, tandis que la théorie de généralisabilité recommande ; soit l'allongement des instruments de mesure, soit la réduction de ces instruments, ou soit encore limiter la portée de la conclusion seulement aux instruments considérés dans l'étude. Ainsi,pour optimiser la fidélité de mesure,nous avons procéder par la réductiondes instruments de mesure,en sélectionnant les items à l'aide de certaines techniques d'analyse des items.

Pour y parvenir, nous avons d'abord élaboré une épreuve de physique de niveau de quatrième secondaire. Nous avons ensuite administré cette épreuve à 150 élèves des différentes classes de 4ème scientifique, tirées au hasard parmi les classes des écoles secondaires protestantes CBCA de la ville de Butembo. Les données ainsi récoltées ont été soumises à une étude de généralisabilité en vue de déterminer la fidélité de l'épreuve. Cette étude de la généralisabilité a été orientée dans les perspectives de l'évaluation critérielle et normative.

Nous avons ensuite appliqué aux items de l'épreuve, quelques techniques de sélection des items notamment la technique de F. Davis, la technique de J. BROWN, la technique de GRUNLUND, la technique de RIGAUX, la technique de corrélation item test et la technique d'analyse des facettes. Celles-ci nous ont permis de constituer cinq épreuves différentes, à part celle optimisée par la technique de Davis ; car cette dernière est équivalente à l'épreuve de départ. Ces cinq épreuves optimisées par les techniques d'analyse des items ont été soumises également à l'étude de généralisabilité en vue de nous renseigner sur leurs fidélités.

A l'issue de nos analyses, nous avons constaté que les coefficients de fidélité de l'épreuve de départ ainsi que des épreuves optimisées, sont satisfaisants pour l'approche critérielle alors qu'ils sont insatisfaisants pour l'approche normative.

Bien que pour l'approche critérielle, les coefficients de fidélité sont satisfaisants, nous avons constaté qu'aucune technique d'analyse des items n'a permis une optimisation remarquable. Car déjà pour l'épreuve de départ, la fidélité était déjà satisfaisante. Ainsi, il y a lieu de dire que, lorsqu'une épreuve a une fidélité satisfaisante pour une approche d'évaluation quelconque (ici pour l'approche critérielle), il est superflu de vouloir encore chercher à optimiser sa fidélité par l'application des techniques d'analyse d'items. Des études ultérieures pourront se pencher sur cette hypothèse.

Concernant l'approche normative, nous avons constaté que les techniques d'analyse des items ont permis de réaliser certaines améliorations de la fidélité. Cependant, ces améliorations n'ont pas été satisfaisantes. Parmi les techniques d'analyse des items appliquées ici, seule la technique d'analyse de

(ISSN: 2805-413X) Evariste KAKULE MUKULU[1].*

https://ijojournals.com/ *Volume 07 || Issue 08 || August, 2024 ||*

facettes a permis de réaliser des améliorations plus ou moins remarquables que les techniques d'analyse de la psychométrie classique.

En terminant cet article, il nous semble qu'à la lumière de cette recherche, il nous est encore assez tôt pour une conclusion consistante sur les apports de ces techniques d'analyse des items pour l'optimisation de mesure. Ainsi nous souhaiterions que d'autres études soient menées pour savoir par exemple si les items sélectionnés par la technique dont les apports sont substantiels correspondent-il à la combinaison d'items qui optimise au mieux la fidélité. Par ailleurs, il serait également intéressant de déterminer les caractéristiques communes à certaines techniques en soumettant par exemple à l'analyse factorielle seulement les items qu'elles sélectionnent à la fois.

BIBLIOGRAPHIE.

ALEXANDRE, V., (1971). Les échelles d'aptitude, Paris 6e, et universitaire,

DOTTRENS, R., (1971). La crise de l'éducation et ses remèdes, Paris, DELACHAUX et NIESTLE.

DE LANDSHEERE, G., (1976). Evaluation continue et examens : précis de docimologie, Bruxelles, Labor.

DE LANDSHEERE, G., (1979). Dictionnaire de l'évaluation et de la recherche en éducation, Paris PUF.

DE KETELE, J.M., (1984). Docimologie, introduction aux concepts et aux pratiques. Kin C.R.P.

DAVIS, F., (1966). Analyse des items. Paris 6e. ed. Nauwelacet.

D'HAINAUT, L. (1978). Concepts et méthodes de la statistique. Vol 2 Labor.

CARDINET, J. et TOURNEUR, Y, (1985). Assurer la mesure, guide pour les études de généralisabilité. Berne, New York, Peter Lang, Francfort.

. MASANDI MILONDO Alphonse. (2016). Méthodes quantitatives et recherche scientifique en sciences sociales : aspects théorique et méthodologiques sur les traitement des données. éditions universitaires européennes, Saarbrûcken, Deutschland/ Allemagne.

THERER, J. (1999). Évaluer pour évoluer : élément de docimologie. IN EVALUATION ET DOCIMOLOGIE. ULG-LEM..