TRAVAIL ET MÉTHODES

REVUE TECHNIQUE DE LA DIRECTION ET DE L'ORGANISATION DES ENTREPRISES

NOVEMBRE 1948

ÉDITIONS SCIENCE ET INDUSTRIE 6, Avenue Pierre-de-Coubertin PARIS XVI

La fidélité des tests

par **M. P. SCHÜTZENBERGER**

Il n'est pas besoin de revenir une fois de plus, dans ces colonnes, sur l'avantage que présente dans les problèmes de sélection l'emploi des méthodes rationnelles. Outre le fait qu'elles sont le plus souvent seules utilisables dans le cas où un nombreux personnel doit être sélectionné pour un poste d'embauche, elles présentent en outre un caractère de rigueur qui permet leur emploi et leur comparaison à divers moments et en divers endroits.

Mais les méthodes rationnelles reposent essentiellement sur une mesure de certaines caractéristiques : traits ou aptitudes ou connaissances de l'individu examiné, et il est bon de connaître la précision avec laquelle cette mesure est faite par tel ou tel dispositif psychométrique donné.

Nous n'étudierons ici que le cas le plus simple où la mesure est un *score*, c'est-à-dire une note conventionnelle obtenue en attribuant au sujet un certain nombre de points, déterminé une fois pour toutes, pour chacune des réponses aux différentes questions (ou plus fréquemment « items ») du test.

Le score se trouve donc ainsi être une véritable conciliation dialectique entre le *jugement qualitatif* porté sur un individu au moyen de chaque item (« capable de répondre à cette question ou non ») et une *mesure quantitative* déduite de la totalité de l'épreuve.

Ce processus de mesure, qui n'est d'ailleurs pas spécial et de loin, à la psychologie, présente cependant dans ce domaine une particularité extrêmement importante, celle de comporter, du fait même de son objet, des erreurs ou fluctuations très larges et presque toujours du même ordre de grandeur que les différences inter-individuelles.

Si, par exemple, l'on opère des pesées au moyen d'une balance de type courant, il sera possible de peser des poids allant de 10 grammes à 5 kilos avec une précision d'environ 10 à 15 grammes. Par conséquent l'étendue de l'instrument de mesure est un très grand nombre de fois plus grand que son erreur propre.

Bien au contraire, en psychologie, si la mesure d'une aptitude s'effectue sur une échelle telle que la quasi totalité des sujets voit son score compris entre + 50 et — 50 il ne sera pas rare, hélas, que chacun de ces scores ne soit connu qu'avec une précision d'environ 10 points.

La nécessité où se sont trouvés les psychologues spécialisés dans ce domaine, la psychométrie, de n'utiliser que de semblables bases numériques pour leur travail, a conduit tout naturellement à un approfondissement et un renouvellement de la théorie de la mesure.

Des outils statistiques nouveaux ont donc dû être forgés permettant de traiter des données peu précises.

Par suite, l'étude des tests en a reçu une nouvelle impulsion et ces méthodes permettent maintenant de construire des épreuves dont la précision dépasse de beaucoup celle que l'on aurait pu espérer il y a seulement une quinzaine d'années.

Il n'est pas inutile de souligner que cette étude de la précision ou *fidélité des tests* n'est pas purement académique. Comme nous le verrons ultérieurement, et comme d'ailleurs le simple bon sens l'exige, la valeur pratique d'une épreuve pour le pronostic de la réussite professionnelle dépend en grande partie de la précision de cette même épreuve et ne peut dépasser une certaine limite, fonction de l'importance des fluctuations.

Comment mesure-t-on la fidélité d'un test ?

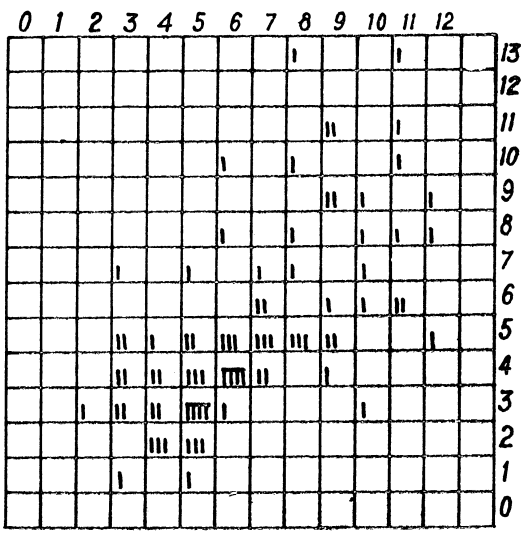
Il convient d'abord de préciser ce que l'on entend par fidélité d'un dispositif quelconque de mesure. Soit par exemple une centaine d'objets de poids différents que l'on pèse avec une balance peu sensible, d'une fidélité de 25 grammes pour fixer les idées, les objets à comparer extrêmes ne différant que d'environ 50 grammes. Une autre estimation de la fidélité de la balance serait la corrélation existant entre les poids indiqués par une première série de mesures et les poids indiqués par une deuxième série ; pour chaque objet en portant en « ordonnée » le poids obtenu à la première mesure et en « abcisse » le poids à la seconde, on pourrait représenter graphiquement l'ensemble de ces deux séries d'opérations par le tableau ci-joint (v. tabl. 1).

L'on appellera alors corrélation un coefficient pouvant varier de — 1 à + 1 et décrivant la manière dont ces points se répartissent autour de la diagonale principale sur laquelle ils seraient presque tous exactement situés si la fidélité de la balance était très grande (de l'ordre de 1 gramme). On dirait alors que la corrélation est 1. Dans le cas présent, elle serait de $r = 0{,}63$.

Enfin la corrélation entre les deux séries de mesure serait 0 si les différents objets ne différaient entre eux que par une quantité inférieure, par exemple à la fidélité de la balance ; les points se répartissant alors en un nuage où n'apparaîtrait aucune direction principale.

De même pour un test quelconque, en supposant que seules sont en jeu les erreurs de mesure dues à son imperfection propre et non à des modifications des aptitudes de chaque sujet d'une passation à l'autre, on définirait la fidélité par la corrélation entre les mesures dans deux séances consécutives.

Le coefficient ainsi obtenu s'appelle coefficient : « test-retest » et donne certainement, quand il est applicable, l'estimation la meilleure de la fidélité du test.



Dans chaque case : nombre d'objets observés la première fois avec le poids p_1 (ordonnée) et la seconde fois avec le poids p_2 (abscisse.

Les chiffres indiqués désignent conventionnellement des unités de 5 grammes.

Ici r = 0,63

Cependant, son emploi n'est en général pas possible. En effet, pour la plupart des tests, l'entraînement dû à une première passation de l'épreuve modifie complètement les résultats de la seconde.

Si l'on imaginait de remplacer le test de la deuxième séance par une épreuve très voisine de la première, on obtiendrait certes un résultat meilleur. Mais encore il convient de voir que l'on a mesuré ensemble deux variations bien distinctes :

1°) *La fluctuation propre due au manque de fidélité de l'appareil de mesure.*

2°) *La fluctuation des sujets d'une expérience à l'autre.* Cette deuxième partie peut être fort importante, comme le montre l'exemple ci-dessous où l'on a corrélé entre eux et de toutes les manières possibles les notes obtenues par un groupe important de sujets dans 6 tests pratiquement identiques passés à des intervalles de temps différents : les tests 1 et 2 dans la même séance, un jour, les tests 3 et 4 dans une même séance un autre jour, et les tests 5 et 6 dans une même séance un troisième jour. La diminution des corrélations est parlante par elle-même.

	1.	2.	3.	4.	5.	6.
1	—	.9457	.9086	.8937	.8643	.8932
2	.9457	—	.9081	.8953	.8609	.8969
3	.9086	.9081	—	.9393	.8806	.8993
4	.8937	.8953	.9393	—	.8527	.8764
5	.8643	.8609	.8806	.8527	—	.9278
6	.8932	.8969	.8993	.8764	.9278	—

Le coefficient à l'intersection de la colonne x et de la ligne y indique la corrélation entre les notes x et y obtenues au test par le sujet.

On remarque la rapide décroissance de cette corrélation à mesure que l'intervalle de temps augmente.

On remarque aussi que les coefficients dans les cases 1,2 — 3,4 et 5,6 correspondants à deux tests passés dans la même séance sont nettement plus élevés que les autres.

SPLIT HALF RELIABILITY. — Pour obvier aux inconvénients présentés par la méthode précédente, l'on a imaginé de diviser au hasard le test en deux moitiés équivalentes et de comparer les scores qui seraient attribués au sujet en ne considérant alternativement que l'une ou l'autre des moitiés.

Il est clair que l'on élimine ainsi, *ipso facto*, la fluctuation propre des sujets d'une séance à l'autre.

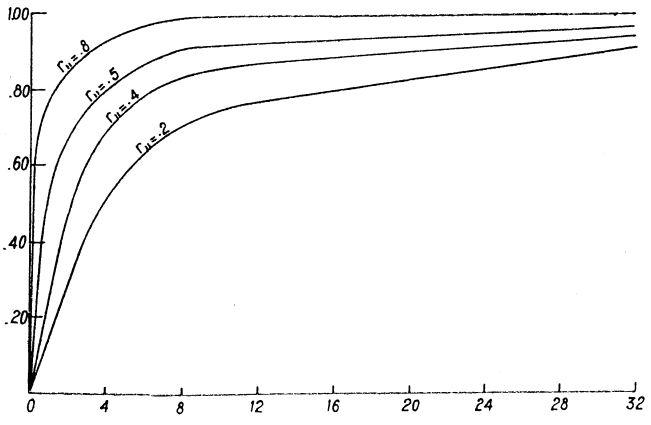
Cependant de nouvelles difficultés apparaissent. En effet, si l'on peut jusqu'à un certain point, lorsque le test est long, se prémunir contre le risque de voir l'une des moitiés mesurer une aptitude différente de celle qui influe sur la réussite dans la première moitié, il n'en est pas de même lorsque le nombre des items est restreint (une dizaine par exemple). Dans la pratique et pour éviter en outre que l'une des moitiés ne comprenne systématiquement des items situés soit à la fin soit au début des tests pour lesquels on pourrait penser que l'attention ou l'entraînement du sujet influe de manière sensible, on a l'habitude de comparer les scores du sujet :

1) dans les questions paires ;

2) dans les questions impaires.

Il est évident, en outre que la fidélité de la note totale d'un test est fonction de la longueur de celui-ci.

Chaque item représentant une sorte de mesure élémentaire, il est assez intuitif que les erreurs sur chacune de ces opérations simples peuvent se compenser en moyenne quand leur nombre augmente et que la note qui résulte de 50 items, par exemple, est plus fidèle et précise que celle qui résulterait de 10 items seulement.



Cette figure donne la fidélité d'un test que l'on allonge N... fois.

En « abscisses » est indiqué le nombre de fois dont on allonge le test ;

En « ordonnée » la fidélité du test allongé. Les courbes tracées correspondent à la fidélité du test initial.

En particulier, la corrélation r que nous avons considérée plus haut entre deux passations théoriques consécutives permet d'en déduire la corrélation r_n analogue pour un test semblable n fois plus long.

C'est la fameuse formule de Spearman-Brown :

$$r_n = \frac{n \cdot r}{1 + (n - 1)\, r}$$

En particulier lorsque nous avons considéré les notes dans les deux moitiés du test, nous n'avons estimé que la fidélité d'un test de longueur deux fois plus faible que le test réel et la fidélité réelle est donnée par :

$$r = \frac{2\, r'}{1 + r'}$$

Mais cette méthode est encore insuffisante car elle a toujours présupposé que chacun des items du test mesurait bien la même aptitude (ou trait, ou connaissance) des sujets. On a donc été amené à imaginer des méthodes plus élaborées, sur le détail desquelles nous ne pouvons entrer ici, et qui consistent essentiellement à comparer le classement que donne chacun des items avec le classement total du test : si le test est homogène, c'est-à-dire mesure bien un trait ou une aptitude unique, ces classements doivent être semblables en gros.

Dans la pratique cependant, il semble bien qu'il n'y ait que rarement avantage à employer cette méthode d'analyse complexe qui nécessite des calculs très longs, et que l'on ait avantage à se concenter le plus souvent de la split half reliability de Spearman-Brown.

D'autres aspects influent encore sur la fidélité que nous ne pouvons traiter entièrement ici : un des plus importants est ce que l'on pourrait appeler le degré de fermeture du test. On désigne par ce terme le nombre de réponses possibles qui sont offertes au sujet pour chaque question. On distingue ainsi les questions auxquelles il ne peut être répondu que par oui ou non, les questions où 2, 3, 4, 5 ou même 6 alternatives sont possibles.

Il est assez immédiat (et le calcul le confirme) que la fidélité du test est augmentée avec le nombre des alternatives offertes au sujet, car la probabilité, pour qu'un sujet trouve la bonne réponse simplement en répondant au hasard, est diminuée dans des proportions considérables. Pour corriger cette influence du hasard, le score s d'une feuille de réponse où il a été bien répondu à B questions, mal à M et où N questions sont restées sans réponse est donné par convention par la formule :

$$s = (B - M)\ N/n$$

où n est le nombre des alternatives.

D'autre part, des expériences nombreuses ont montré que la fidélité d'une épreuve était sensiblement accrue lorsque les sujets possédaient un entraînement suffisant à résoudre ce type de problème. C'est pourquoi dans la plupart des tests bien construits, 5 à 10 minutes du temps avant la passation sont laissées au sujet pour s'entraîner sur des questions analogues à celles qui lui seront posées ultérieurement et dont on ne tient aucun compte dans la correction définitive.

Ces diverses remarques nous permettent de donner quelques brèves indications sur la manière dont peut s'étudier rapidement une épreuve nouvelle lorsque les exigences d'un poste de travail nécessitent l'élaboration d'un test spécial.

Il va sans dire que toutes ces indications sont données au sujet des tests « papier-crayon » que l'on tend à employer de plus en plus fréquemment mais qu'elles s'étendent, mutatis mutandis, aux autres types d'épreuves (tests d'appareils, etc.) :

1) Construire un nombre d'items environ deux fois supérieur à celui que l'on a l'intention d'inclure dans la forme définitive des tests ; utiliser de préférence des questions où les diverses alternatives possibles ont, *a priori*, la même chance d'être désignées par le sujet ; éviter les questions dont la solution est une véritable devinette et qui, de ce fait, font appel à des aptitudes différentes de celles que doit mesurer le test ; prévoir un nombre suffisant de questions pour exercer le sujet.

2) Administrer à un groupe homogène d'une centaine de sujets de même niveau que ceux auxquels est destinée ultérieurement l'épreuve, la première forme ainsi préparée.

3) Pour chaque item dénombrer le nombre de fois pour lequel la bonne réponse a été indiquée ; éliminer de la correction celles des questions auxquelles moins de 20 % ou plus de 80 % ont correctement répondu.

4) Attribuer à chaque sujet sur la base des items restant un score égal au nombre de ses bonnes réponses et vérifier que la courbe de répartition de ces scores a une forme suffisamment voisine de la normale.

5) Calculer la Split half reliability.

6) Si celle-ci est inférieure à la borne que l'on s'est fixée, il convient alors d'éliminer les items pour lesquels une forte proportion de sujets, considérés comme les meilleurs d'après le score total, a mal répondu (une limite commode est plus de 75 % des sujets classés dans les 25 % les meilleurs d'après le score total) ; rajouter de nouvelles questions et refaire passer sur une nouvelle population.

Cette opération doit souvent être répétée 2 ou 3 fois jusqu'à l'obtention d'une fidélité suffisante pour le but que l'on se propose.

Il nous faut maintenant examiner ce que signifient concrètement les coefficients de fidélité que nous venons d'étudier et quels sont les rapports de cette notion avec la validité ou efficacité des tests.

Une interprétation fort simple du coefficient de fidélité est la suivante :

Supposons que nous ignorions tout de la valeur de l'aptitude d'un sujet à un test dont les notes s'échelonnent environ de — 50 à + 50. *A priori* il nous est donc impossible de dire à quel point de cet échelle la note du sujet se trouvera.

Si nous connaissons déjà la note du sujet dans une épreuve très voisine de l'épreuve en question et dont la corrélation avec cette dernière se trouve être égale à r nous pouvons savoir que cette étendue se trouve diminuée dans la proportion de $1 - r$.

D'après ce qui précède il est donc clair qu'à moins que la fidélité n'atteigne une valeur extrêmement élevée, telle que 0,97 ou 0,98, l'amélioration de la connaissance de la valeur vraie du sujet dans un test n'est que très faible.

Ce point est important car les épreuves sont rares dont la fidélité dépasse 0,80 et notamment dans le domaine caractériel.

L'on a ainsi une idée plus précise des services que peut rendre la psychométrie susceptible d'apporter des renseignements précieux lorsqu'il s'agit de comparer des moyennes de notes relatives à un groupe de sujets, mais souvent encore trop imprécise pour pouvoir être utilisée dans la prédiction des possibilités d'un sujet donné, comme dans l'orientation professionnelle.