Méthodes d'évaluation en psychologie de développement

H. Fischer[1]

Le problème des échelles de développement en psychologie expérimentale est fortement discuté depuis la publication de quelques ouvrages américains de psychométrie. Le passage d'échelles qualitatives aux échelles quantitatives pose des problèmes nouveaux, signalés surtout depuis les travaux de Yule et de Kendall[2]. Les épreuves statistiques concernant les variables qualitatives ou *attributs* se distinguent des épreuves statistiques concernant les variables quantitatives. Nous avons signalé l'existence de cette différence dans notre manuel de statistique élémentaire destiné aux étudiants en psychologie et en pédagogie[3]. Ces discussions mettent en évidence un aspect particulier dès qu'on pose la question: les échelles de développement sont-elles également à repenser et à rediscuter? Or, ces quelques lignes représentent un effort de relier le point de vue génétique aux discussions plus générales sur les mesures en psychologie.

1. Rappel des méthodes classiques d'échelles de développement

Notre tâche n'est pas d'exposer ici les méthodes classiques de Binet-Simon et d'autres pour juger du niveau d'intelligence ou des quotients d'âges des enfants. René Zazzo a donné l'historique des techniques et des révisions des méthodes préconisées par Binet, Simon, Terman, Merrill[4]. Rappelons la technique sous sa forme originale[5]: On pose, par exemple, cinq questions aux enfants d'un âge chronologique fixe, puis cinq autres questions aux enfants d'un âge chronologique fixe suivant et ainsi de suite pour chaque âge. Ces tâches sont chaque fois caractéristiques pour l'âge chronologique considéré, c'est-à-dire l'étalonnage précédant l'application du test a montré une réussite moyenne suffisante, par exemple de 80%. Autrement dit: un enfant normalement doué devrait pouvoir résoudre les tâches correspondant à son âge.

Cependant, en réalité, on rencontre des enfants ne pouvant pas résoudre l'ensemble des questions posées. C'est pour cela qu'on commence toujours par

[1] Adresse: PD Dr. Hardi Fischer, Forschungsstelle für Arbeitspsychologie an der ETH., Zürichbergstraße 18, Zürich 7/32

[2] Yule et Kendall: An Introduction to the Theory of Statistics. London 1950.

[3] Fischer Hardi: Les méthodes statistiques en psychologie et en pédagogie. Neuchâtel/Paris 1955.

[4] Zazzo René: Intelligence et quotient d'âges. Paris 1946.

[5] Binet Alfred et Simon Th.: La mesure du développement de l'intelligence chez les jeunes enfants. Paris 1931.

poser des questions d'âges précédents jusqu'à ce qu'on ne rencontre que des réponses fausses à un niveau supérieur à un âge dont toutes les réponses sont justes. On peut rencontrer le schéma suivant pour un enfant de $5\frac{1}{2}$ ans, en considérant toujours cinq questions par âge chronologique:

Tâches pour l'âge chronologique de	1	2	3	4	5
4 ans	+	+	+	+	+
5 ans	+	+	—	+	+
6 ans	+	+	+	—	—
7 ans	+	—	+	—	—
8 ans	—	—	—	—	—

Voici comment on juge l'enfant: on constate qu'il a résolu toutes les questions d'un enfant de 4 ans, puis $^4/_5$ des questions pour 5 ans, $^3/_5$ des questions pour 6 ans, $^2/_5$ des questions pour 7 ans, aucune question pour 8 ans, etc. Le total des fractions à ajouter à la base de 4 ans (toutes les réponses justes) est ainsi de $^4/_5 + {}^3/_5 + {}^2/_5 = {}^9/_5 = 1^4/_5$ ans. On appelle *âge mental* le total, soit $5^4/_5$ ans.

La comparaison de l'âge chronologique AC et de l'âge mental AM nous fournit les renseignements suivants:

$AC = AM$: développement normal
$AC > AM$: développement retardé
$AC < AM$: développement avancé

Le degré d'intelligence peut se mesurer par le quotient intellectuel QI qui est défini par le rapport entre l'âge mental et l'âge chronologique, donc

$$QI = \frac{AM}{AC}$$

En termes de QI la discussion précédente devient: si

$QI = 1$: développement normal
$QI < 1$: développement retardé
$QI > 1$: développement avancé

Nous ne croyons pas que les auteurs de ce genre de tests se soient posés la question de savoir comment on a calculé le QI. Ils ont fait l'hypothèse que chaque question était équivalente et n'ont pas fait de pondération. Supposons maintenant deux enfants de 9 ans exactement; les deux obtiennent un âge mental de $9^1/_5$. Mais tandis que le premier a été calculé sur la base de 6 ans (toutes les réponses justes), le second a été calculé sur la base de 9 ans. Il y a donc bien une différence qu'on peut exprimer sous forme d'indice de dispersion. Ce qui pourrait intéresser à la rigueur, c'est la nature psychologique des réponses justes fournies. Or, les auteurs des tests classiques de développement, étant donné qu'ils ont accepté l'hypothèse de l'équivalence des questions en jeu, n'ont pas eu une raison particulière d'étudier ces phénomènes et du point de vue statistique ils ont peut-être eu raison. Cela était d'autant plus justifié qu'il n'ont jamais considéré, à vrai dire, des stades, de telle sorte qu'une seule question réapparaît à des âges chronologiques différents, pouvant ainsi donner lieu à des réponses différentes d'un âge à l'autre.

Gesell a d'ailleurs établi un graphique avec les deux variables suivantes: âge mental et dispersion de l'âge mental. Il obtient ainsi un zéro absolu, semblable au zéro absolu dans la mesure des températures, se situant à un âge mental de —66 jours et ceci grâce à une représentation de la droite comme fonction entre les deux variables. —66 jours correspond à 66 jours avant la naissance, date minimum à laquelle un bébé né prématurément est viable. Ce raisonnement de Gesell suppose l'hypothèse vraie selon laquelle les deux variables sont liées par une fonction linéaire, ce qui n'est nullement prouvé.

Le choix des différentes tâches est tout d'abord dicté par des considérations psychologiques. A part cela, une fois constituée une série de questions typiques pour un âge donné, on peut calculer la corrélation entre la répartition des réponses à une question avec le QI. En effet, on peut, par exemple, calculer le pourcentage de réponses positives pour trois classes de QI (disons A: audessous de 0,95, B: entre 0,96 et 1,05, C: au dessus de 1,06), ce qui permet de calculer déjà le coefficient de contingence de Pearson ou le coefficient de corrélation selon Coumétou[1]. Ces calculs présupposent cependant des distributions normales. Or, par le fait même qu'un QI de 1,40 ne peut pas être considéré comme le double d'un QI de 0,70, on est amené à rejeter l'hypothèse d'une distribution continue des QI. En acceptant cependant quand même l'hypothèse de distributions normales, ce qui nous paraît justifié après des recherches approfondies de L. L. Thurstone[2], on applique les formules mentionnées ou bien, si la différenciation est plus grande (nombre de classes > 6, par exemple) le quotient de corrélation des rangs selon Spearman[3], ne présupposant pas nécessairement des distributions normales. Ainsi Terman et Merrill ont publié quelques-unes de ces corrélations entre une tâche et l'ensemble (QI). Les plus fortes corrélations étaient de l'ordre de 0,90, les plus faibles de l'ordre de 0,40[4]. Afin d'éviter de mesurer les mêmes facteurs psychologiques, il serait indiqué de

[1] Fischer Hardi: Les méthodes statistiques en psychologie et en pédagogie. Neuchâtel/Paris 1955. pp. 85–87.

[2] Thurstone L. L.: A method of scaling psychological and educational tests. Journal of Educational Psychology, 16/1925.

[3] Fischer Hardi: op. cit. pp. 102–104.

[4] Terman L. M. et Merrill M. A.: Measuring intelligence. Boston.

trouver des corrélations ou des associations faibles entre les diverses questions, car une forte association signifierait qu'une question pourrait aussi bien être remplacée par une autre.

Si la méthode de Binet-Simon consiste en une évaluation qualitative qui n'est quantifiée que par un procédé reposant sur une convention d'équivalence et non pas sur une hiérarchisation des questions ou sur une pondération des diverses parties, les échelles quantitatives d'aptitudes mettent en évidence un autre aspect du développement. Dès qu'on utilise une variable quantitative, on sous-entend la normalité de la distribution des caractères étudiés dans la population. En outre on suppose l'additivité simple des résultats formant un ensemble et par conséquent la proportionnalité de ces résultats (dès qu'on travaille dans une échelle continue et absolue).

Les épreuves statistiques de significativité concernant les variables quantitatives sont déjà classiques: t-test, F-test, etc. Or, dès que deux distributions normales concernant deux âges différents nous fournissent des différences significatives, nous établissons deux étalonnages distincts et ainsi de suite pour tout âge chronologique. On observe ainsi une augmentation du rendement pour les catégories d'âges différents.

Nous nous sommes posés les questions suivantes:

1) Y aurait-il un moyen d'éviter les significations multiples dans les échelles de Binet-Simon? En effet, un âge mental de $9^1/_5$ peut être atteint, non pas seulement à partir de bases d'âge chronologique différentes, mais par l'intermédiaire de réponses différentes. L'âge mental de $9^1/_5$ ne nous dit rien sur la nature des questions pour lesquelles l'enfant a échoué ou réussi.

2) Y aurait-il moyen de relier une qualification selon des stades du développement et une quantification suivant un modèle mathématique rigoureux?

A ce propos, nous avons étudié les possibilités existantes en nous basant sur deux méthodes de quantification: la comparaison par paires et l'analyse hiérarchique.

2. Comparaison par paires

Peter R. Hofstätter et J.-M. Faverge mentionnent dans leurs manuels la méthode de comparaison par paires, et en donnent les caractéristiques essentielles[1]. Nous nous poserons la question de savoir si la méthode de comparaison par paires peut trouver une application pour les échelles de développement.

Prenons un exemple fictif pour expliquer la méthode: on aimerait savoir si la préférence pour une couleur change en fonction de l'âge chronologique. A cet effet on choisira quatre couleurs: rouge – bleu – jaune – brun.

[1] Hofstätter Peter R.: Einfuhrung in die quantitativen Methoden der Psychologie. Munich 1953. Faverge J.-M.: Méthodes statistiques en psychologie appliquée. Paris 1954.

On ne les montre aux enfants que deux par deux, mais suivant toutes les combinaisons possibles. Il y en a:

$$C_4^{(2)} = \frac{4 \cdot 3}{1 \cdot 2} = 6 \text{ possibilités différentes.}$$

Aucune réponse intermédiaire n'est possible; le sujet doit chaque fois se décider pour l'une ou pour l'autre. Nous pouvons établir un tableau à double entrée de la façon suivante: dans les colonnes on mettra la préférence donnée à la couleur correspondante étudiée par rapport à celle indiquée en ligne (+). Les diagonales restent ainsi vides.

	rouge	bleu	jaune	brun
rouge		—	—	+
bleu	+		—	+
jaune	+	+		—
brun	—	—	+	

Ce tableau à double entrée est inversement symétrique par rapport à la diagonale. En effet, si le sujet préfère ici le rouge au bleu, ce que nous avons représenté par le symbole + dans le deuxième casier de la première colonne (rouge), on peut en déduire que bleu n'est pas préféré au rouge, représenté par le symbole — dans le premier casier de la deuxième colonne (bleu), etc.

On forme le total des + pour chaque colonne et on constate que le sujet examiné préfère les couleurs rouge et brune (total 2) aux couleurs bleu et jaune (total 1). On aurait pu trouver également les sommes de 3, 2, 1, 0, ceci dans n'importe quelle combinaison.

En reprenant la même méthode avec un grand nombre de sujets, disons $N = 60$, on pourrait superposer les tableaux à double entrée de chaque sujet et on obtiendrait pour un âge chronologique fixe, par exemple:

	rouge	bleu	jaune	brun
rouge	—	20	30	40
bleu	40	—	10	50
jaune	30	50	—	30
brun	20	10	30	—
Total	90	80	70	120

En prenant le total de chaque colonne, nous avons une indication sur la préférence du groupe de sujets étudiés. Or, il est clair que la somme totale des quatre sommes partielles nous donne 360 ou tout simplement le produit $C_4^{(2)} \cdot N$, c'est-à-dire $6 \cdot 60 = 360$.

Dès qu'on désire faire des comparaisons, on a avantage à exprimer les résultats en pourcentage. Soit S la proportion sur une échelle de 100 (au lieu de 60) et soit la somme partielle de chaque colonne symbolisée par Σ, on aura alors la relation:

$$S = 100 \frac{\Sigma}{N\,(n-1)}$$

où n = nombre de critères (couleurs). Dans notre exemple on trouve:

rouge	bleu	jaune	brun	
50	44	39	67	%

Le total des sommes partielles est ainsi de $4 \cdot 50 = 200$ ou tout simplement de $n \cdot 50$.

Toutes les épreuves de significativité s'appliquent: entre garçons et filles, entre milieux différents et également entre âges différents. Supposons qu'on ait trouvé pour un autre âge les proportions sur une échelle de 100 de S':

rouge	bleu	jaune	brun	
58	47	40	55	%

On peut se demander si les deux séries S et S' se différencient significativement ou on peut encore comparer chaque série S et S' à une série attendue composée de quatre fois 50. Dans le premier cas $\chi^2 = 0{,}86$ n'est pas significatif pour un degré de liberté 3, ce qui veut dire que les deux séries ne se différencient que par l'effet du hasard. Dans le second cas $\chi^2 = 9{,}91$ indique une différence significative pour un degré de liberté 3 pour la série S, mais aucune différence significative ($\chi^2 = 3{,}96$) pour la série S', ce qui veut dire que la série S se différencie seule de la série attendue 50, 50, 50, 50, tandis que les différences observées de la série S' avec la série attendue ne sont explicables que par le hasard seul.

Si le nombre des critères est de beaucoup plus élevé, on peut même prévoir le calcul d'un coefficient de corrélation de Spearman, pour des expériences du type suivant fournissant deux séries S et S': notre exemple fictif cité pourrait être répété en variant certaines conditions. Ainsi, au lieu d'utiliser des couleurs vives, on pourrait répéter l'expérience par des teintes atténuées; ou bien encore on pourrait comparer l'influence des conditions du temps pendant lequel on fait l'expérience (matin et soir), ces schémas d'expériences se faisant toujours sur les mêmes sujets.

Si ces expériences permettent bien une analyse génétique des résultats, il convient maintenant d'envisager une identification des critères avec des stades de développement. Nous nous excusons d'emblée du caractère purement théorique de la présentation de ce problème dans le cadre d'une telle étude résumée. Nous nous rapprochons ainsi des échelles génétiques de Binet-Simon. Un enfant normalement développé devrait fournir une série S nettement décroissante, si les critères A, B, C, etc. sont représentatifs d'une théorie génétique.

A titre d'exemple nous choisissons quelques recherches de Jean Piaget et de Bärbel Inhelder concernant la conservation[1]. On sait que l'enfant possédant la notion de conservation du volume (11 ans environ) possède nécessairement les notions de conservation de poids (9 ans environ) et de conservation de la substance (environ 7 ans). Un enfant de 11 ans devrait donc se distinguer par la possession des trois types de conservation mentionnés. On appellera le «stade du volume» stade A, le «stade du poids» le stade B et le «stade de la substance» le stade C. Il n'y a chaque fois que deux possibilités:

+ appartient au stade ou

— n'appartient pas au stade.

Ainsi il est évident qu'un enfant appartenant au stade A (maximum) appartient automatiquement aux stades B et C, puisque si le sujet sait faire la construction de A il saura toujours faire celles de B et C. Il n'y a donc plus une préférence ici, mais des nécessités (inclusions).

Dans notre exemple schématique et théorique il y a trois solutions distinctes possibles, si l'on admet que le stade C inférieur est atteint par tous les sujets examinés. On aura donc:

sujet au stade A:

	A	B	C
A		—	—
B	+		—
C	+	+	
S_1	2	1	0

sujet au stade B:

	A	B	C
A		+	+
B	—		—
C	—	+	
S_2	0	2	1

sujet au stade C:

	A	B	C
A		+	+
B	—		+
C	—	—	
S_3	0	1	2

Il suffit de ne considérer que ce qui est sous la diagonale. Ainsi un sujet au stade B est nécessairement au stade C: ce qui est au-dessus de la diagonale est alors simplement l'inverse de ce qui est sous la diagonale et ceci symétriquement. La diagonale elle-même reste vide, comme dans les exemples classiques de comparaisons par paires, car on ne considère que les nécessités pour les stades inférieurs et non pas les égalités. Il est vrai qu'un sujet au stade C est certainement au stade C, mais cette évidence n'est pas inscrite dans la diagonale.

La théorie des échelles hiérarchiques de L. Guttman nous permettra d'approfondir les questions abordées[2].

[1] Piaget Jean et Inhelder Bärbel: Le développement des quantités chez l'enfant. Neuchâtel/Paris 1941.

[2] Nous devons des précisions sur la méthode des comparaisons par paires essentiellement à L. L. Thurstone: Psychological Analysis. Amer. J. Psychol. 38/1927.

3. **Echelles hiérarchiques** (par ordre d'inclusion)

Louis Guttman a développé une analyse hiérarchique des parties de tests pouvant aboutir à des échelles[1]. Serge Moscovici donne un résumé des idées principales[2]. Quant à nous, notre but est la transformation de ces échelles en échelles de développement.

On a l'habitude, en psychologie appliquée, d'utiliser des tests composés d'un nombre suffisant d'items. Supposons un test de 20 items. On peut se poser la question théorique suivante: quelle est la probabilité de fournir un résultat 10, si l'on attribue 1 point par réponse juste et sans limitation du temps. Or, le résultat 10 peut se produire de

$$C_{20}^{(10)} = \frac{20 \cdot 19 \cdot 18 \cdot 17 \cdot 16 \cdot 15 \cdot 14 \cdot 13 \cdot 12 \cdot 11}{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6 \cdot 7 \cdot 8 \cdot 9 \cdot 10} = 184756$$

manières différentes. En pratique, tous les sujets avec 10 points ne se distinguent pas et cela malgré la nature différente des réponses fournies. Une telle méthode présuppose une équivalence psychologique rigoureuse des items, ce qui est très difficile à atteindre, même en approximation seulement. Si tel était le cas, cela voudrait dire que toutes les fréquences pour les différentes manières d'atteindre un résultat de 10 seraient identiques, ce qui n'est pas le cas, en général[3]. Sur la base de tels résultats hautement discutables on a fait des raisonnements inductifs grâce à des distributions caractéristiques de la variable.

L. Guttman a les mérites d'avoir élaboré «une théorie mathématique capable de traiter correctement les attributs, en respectant les exigences d'une quantification qui leur soit propre[4]. L'échelle de Guttman évite précisément la dispersion exagérée des fréquences de la variable[5]. Dans un cas limite, les sujets ayant obtenu le même résultat, l'ont atteint grâce aux mêmes questions. Autrement dit: il faudrait atteindre le cas idéal dans lequel une réponse juste à une question (item) implique des réponses justes pour toutes les questions précédentes. Or, ceci convient très bien à toute étude génétique, dès qu'on considère tout item comme un stade dans l'ordre de l'ensemble des stades.

Il faudrait cependant qu'on observe quelques conditions propres à cette méthode. Ainsi les stades doivent appartenir à un continuum psychologique unique, par exemple à l'étude de la notion de conservation chez l'enfant. Il faudrait enfin pouvoir préciser un point neutre (zéro) qui délimite le positif du négatif. En effet, il faudrait pouvoir dire avec précision à partir de quel comportement on peut parler d'une acquisition de la notion de la conservation.

[1] Studies in Social Psychology in World War II. Vol. IV. Measurement and prediction. Princeton 1950.
[2] Moscovici Serge: L'analyse hiérarchique. L'Année psychologique 1954. Paris 1954.
[3] Faverge J.-M.: Les appréciations professionnelles et la recherche des critères. Dans: Ombredane/Faverge: L'analyse du travail. Paris 1955.
[4] Moscovici S.: op. cit. p. 85.
[5] Studies in Social Psychology. op. cit. Chap. II et III.

La quantification, lors de la description d'une variable qualitative, se fait donc en deux étapes: tout d'abord on constate la présence ou l'absence de la qualité, puis on dénombre des individus suivant la présence et l'absence de cette qualité.

Si L. Guttman a étudié une succession de questions hiérarchisées portant sur une attitude des sujets, nous considérons une succession de stades psychologiques. Cela signifie que le sujet se trouvant au stade A (maximum) est nécessairement aux stades B et C inférieurs; exprimé logiquement nous avons donc

$$A < B < C$$

où $<$ est le symbole des inclusions: un sujet en A est inclus automatiquement en B, un sujet en B est inclus automatiquement en C, etc.

Guttman signale plusieurs modèles (pour la plupart du type parallélogrammique) d'échelles hiérarchisées que nous donnons ici en nous basant sur l'idée des stades.

Modèle I

Rang	Points	Fait partie de A	B	C	Ne fait pas partie de A	B	C
1	3	+	+	+			
2	2		+	+	+		
3	1			+	+	+	
4	0				+	+	+

C'est l'exemple ressemblant à la méthode expliquée au paragraphe précédent et décrit à l'instant.

Modèle II

Un psychologue pourrait cependant s'intéresser plutôt aux stades intermédiaires à deux bornes fixes. Cela revient à dire que le passage d'un stade à un autre ne serait pas lent, mais brusque. On serait ainsi amené à distinguer les quatre critères suivants:

D, stade faisant suite à A
E, stade entre A et B
F, stade entre B et C
G, stade précédant C.

Voici les situations possibles:

Rang	Fait partie du stade intermédiaire D	E	F	G	Ne fait pas partie du stade intermédiaire D	E	F	G
1	+					+	+	+
2		+			+		+	+
3			+		+	+		+
4				+	+	+	+	

On constate que dans ce cas particulier on ne peut classer les sujets que grâce à leur situation dans un stade. Il n'est pas question ici d'attribuer 1 point par stade, car on ne trouverait que le total 1 par rang et toute différenciation deviendrait difficile. Le rang est un fait implicite de la succession des stades. On ne tient compte que de la situation actuelle et observée des sujets et non pas des stades parcourus précédemment.

Modèle III

Souvent, en psychologie, il est difficile de préciser les stades dans lesquels les sujets se trouvent. Tel est surtout le cas si les passages d'un stade à un autre sont lents. On retrouve alors le modèle I en maintenant les stades D, E, F et G, mais en indiquant deux stades voisins. Tel ou tel sujet fait le passage du stade E au stade D, etc. Les stades se recoupent mutuellement.

Rang	Passages				Stades non impliqués			
	A	B	C	D	A	B	C	D
1	+	+					+	+
2		+	+		+			+
3			+	+	+	+		

En réalité, les situations ne sont pas si nettes. Mais avant de discuter comment on procéderait en pratique, il nous faut préciser quelques caractéristiques des échelles de Guttman.

Pour la suite des discussions, restons fixés au premier modèle que Guttman a utilisé pour procéder à l'investigation des attitudes. Voici un exemple pour illustrer la méthode: on pose trois questions à la population examinée:

A. Etes-vous citoyen de Genève?
B. Etes-vous citoyen suisse?
C. Etes-vous Européen?

Or, vouloir répondre par l'affirmative à la question B signifie qu'il faut répondre par l'affirmative à la question C; vouloir répondre par l'affirmative à la question A suppose des réponses positives aux questions B et C. Ainsi on oberservera un pourcentage croissant des réponses positives de A à B, puis à C.

Cet exemple, ainsi que ceux discutés précédemment sont des exemples exclusivement dichotomiques: la construction était juste ou fausse ou le sujet fait partie d'un stade ou non. Or, dans ce cas on observe un phénomène curieux: en calculant l'association q entre les critères (questions) A et B, par exemple, on observe un casier = 0 dans le tableau à double entrée. Supposons l'exemple suivant: pour les enfants de 10 ans on remarque les fréquences suivantes (voir modèle I):

90% sont au stade B
60% sont au stade A

Il n'est, en principe, pas possible d'obtenir un pourcentage inférieur en B, puisqu'il y aura moins de sujets qui ont déjà atteint le stade A. Le tableau à double entrée peut s'établir:

		B +	B —	
A	+	60	0	60
A	—	30	10	40
		90	10	100

Dans notre exemple on trouve une association de 0,41 sans correction concernant l'homogénéité[1]. Elle est de 0,37 pour l'association corrigée, de telle façon que pour N = 100 sujets on trouve $\chi^2 = 13{,}9$ pour un degré de liberté 1. Cette valeur est très significative, on peut donc dire que les différences observées se distinguent significativement, ce qui est certainement dû avant tout au casier = 0.

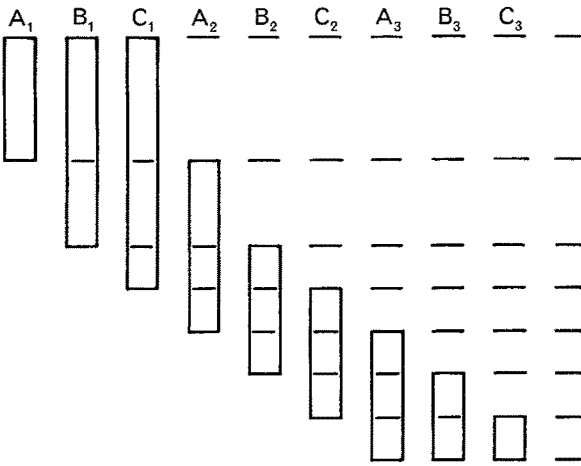
Supposons maintenant la situation suivante: on pourrait atteindre le stade suivant de trois manières différentes sur le plan psychologique que nous distinguerons par les indices 1, 2 et 3. Un échantillon de 100 sujets fournit les résultats suivants:

manières psychologiques	stades A	B	C
1	30	50	60
2	40	30	20
3	30	20	10

Il n'y a pas chevauchement pour l'instant entre 1, 2 et 3, de telle façon que le total pour chaque stade est de 100.

Dès qu'on établit un graphique correspondant, on rejoint les théories des scalogrammes, mais le cadre de cette étude ne permet pas de décrire les détails. En ordonnant convenablement les réponses, on peut trouver la classification univoque que voici:

[1] Fischer Hardi: op. cit. pp. 81–83.



Solutions	%	Rangs
111	30	1
112	20	2
122	10	3
222	10	4
223	10	5
233	10	6
333	10	7

Il y a en effet une correspondance bi-univoque et on peut répartir les rangs en partant de l'hypothèse que la solution 1 est toujours psychologiquement supérieure à la solution 2 et que la solution 2 est supérieure à la solution 3.

On peut enfin se demander si le choix des solutions est toujours le même ou s'il y a association entre le choix successif des diverses solutions dans un sens fixe ou si c'est simplement dû au hasard. Entre les stades C et A on obtient le tableau suivant:

		stade C 1	stade C 2	stade C 3	
stade A	1	30	0	0	30
	2	30	10	0	40
	3	0	20	10	30
		60	30	10	100

Le $\chi^2 = 73{,}66$ qu'on obtient est très significatif pour le degré de liberté 4. Le coefficient de contingence fournit un résultat $C = 0{,}65$, c'est-à-dire une association très marquée. On peut donc dire que les sujets ont plutôt tendance à utiliser le même schéma de solution. Mais on ne prévoit pas encore, à partir d'un rang 5, par exemple, à quels stades correspondent les types de solutions 1, 2 et 3. En tenant compte de cette correspondance aux stades, on obtient:

solutions et stades			rangs
$C_1B_1A_1$	ou	$A_1B_1C_1$	1
$C_1B_1A_2$	ou	$A_2B_1C_1$	2
$C_1B_2A_2$	ou	$A_2B_2C_1$	3
$C_2B_2A_2$	ou	$A_2B_2C_2$	4
$C_2B_2A_3$	ou	$A_3B_2C_2$	5
$C_2B_3A_3$	ou	$A_3B_3C_2$	6
$C_3B_3A_3$	ou	$A_3B_3C_3$	7

Dès lors, il y aura pour chaque rang une correspondance bien précise entre stades (A, B, C) et types de solution. Un rang désigne non seulement le niveau du sujet examiné, mais nous renseigne tout de suite sur les types de solutions qu'il a choisis pour se situer dans tel ou tel rang. Voici un exemple: le rang 5 signifie que le sujet se situe au stade C grâce à la solution 2, au stade B grâce à la solution 2 et au stade A grâce à la solution 3. Or, ce schéma présuppose, sous cette forme, que chaque sujet atteint le stade A maximum. Or, tel n'est pas toujours le cas. Il suffit de préciser les solutions comme suit: 3 = aucune solution. Il nous reste deux critères: 2 et 1 avec la précision que du point de vue psychologique la solution 2 vaut moins que la solution 1.

Répétons donc les significations des rangs:

rang 7: aucune solution pour les trois stades;
rang 6: stade C est atteint par la solution 2. Les autres stades ne sont pas atteints;
rang 5: stades C et B sont atteints par la solution 2. Le stade A n'est pas atteint:
rang 4: tous les trois stades sont atteints par la solution 2;
rang 3: le stade C est atteint par la solution 1. Les stades B et A sont atteints par la solution 2;
rang 2: les stades C et B sont atteints par la solution 1. Le stade A est atteint par la solution 1;
rang 1: les trois stades sont atteints par la solution 1.

On remarque que la représentation graphique des scalogrammes donne une sorte de parallélogramme, ce qui ne serait pas le cas s'il n'y avait pas, dès le début, cette condition importante qu'un sujet d'un stade supérieur se situe automatiquement dans les stades précédents.

Le scalogramme permet d'ordonner les sujets sur un continuum (vertical), car chaque dichotomie (ou trisection) produit une coupure (des coupures) adéquate(s) sur ce continuum. Inversement, toute caractéristique du développement est fonction simple du rang le long du continuum[1].

[1] Moscovici S.: op. cit. p. 97.

En réalité les situations idéales que nous avons décrites ne se produisent pas ou se produisent rarement. Il se pourrait qu'un sujet «saute» du stade C au stade A, sans passer par le stade B. On ne pourrait donc pas, sans autre, prétendre que le rang 3 dans notre échelle précédente signifie «le stade C est atteint par la solution 1, les stades B et A sont atteints par la solution 2».

Voici comment on procède pratiquement. On examine, par exemple, 20 sujets en ce qui concerne leur appartenance ou leur non-appartenance aux stades A, B et C. On ordonne les sujets pour qu'ils s'approchent autant que possible du scalogramme mentionné. On désigne par x la «solution juste» et par 0 toute erreur dans cette échelle. Voici le tableau ainsi obtenu:

Appartenance aux stades			Non-appartenance aux stades			Numéro sujet	Rang
A	B	C	A	B	C		
			+	+	+	1	4
	+		+	0	+	2	4
		+	+	+		3	3
		+	+	+		4	3
		+	+	+		5	3
+		+	0	+		6	3
		+	+	+		7	3
		+	+	+		8	3
+		+	0	+		9	3
	+	+	+			10	2
	+	+	+			11	2
	+	+	+			12	2
	+	0	+		+	13	2
	+	+	+			14	2
	+	+	+			15	2
	+	+	+			16	2
+	+	+				17	1
+	0	+		+		18	1
+	+	+				19	1
+	+	+				20	1

Par une méthode simple, Suchman propose qu'on calcule un *coefficient de reproductibilité* CR qui se trouve par la formule:[1]

$$\mathrm{CR} = 1 - \frac{\text{nombre d'erreurs}}{\text{nombre de stades distinctifs} \times \text{nombre de sujets}}.$$

Ainsi, si l'on trouve 5 exceptions (erreurs) sur 20 sujets examinés pour les trois stades A, B et C, on aura:

$$\mathrm{CR} = 1 - \frac{5}{3 \cdot 20} = 1 - 0{,}08 \qquad \mathrm{CR} = 0{,}92.$$

[1] Suchman A. Edward: The scalogram board technique for scale analysis. In: Mesurement and prediction. p. 117.

La reproductibilité mesure le degré d'exactitude avec lequel les sujets suivent l'ordre de l'échelle. Autrement dit, elle indique la précision avec laquelle on peut conclure sur l'ensemble de l'appartenance aux stades précédents à partir d'un rang. On accepte l'existence d'une échelle hiérarchique (ou de développement parfaite) pour $CR > 0{,}90$.

Le coefficient CR est proportionnel au nombre d'erreurs et inversement proportionnel au nombre de stades et au nombre de sujets examinés. Notre coefficient permet d'accepter l'existence d'une échelle de développement véritable.

Guttman a cependant montré que même pour des constructions d'échelles qui n'obéissent pas aux conditions du coefficient de reproductibilité CR, représentant d'ailleurs une sorte de coefficient de fidélité, il sera encore possible de construire une échelle avec la propriété de «produire la même corrélation avec un critère extérieur comme la corrélation multiple des items individuels le fait avec ce même critère extérieur»[1]. Souvent il est cependant possible d'ordonner les items d'un test d'une manière cumulative en construisant des items de telle façon que l'erreur de reproductibilité soit rendue minimum. Traduite en termes génétiques, il faudrait trouver des caractéristiques cumulatives dans le développement de l'enfant et se baser seulement sur ces caractéristiques qui rendront un scalogramme possible et presque certain.

Signalons, pour finir, que l'inconvénient de cette méthode réside dans le fait que de telles échelles ne suffisent pas à la condition d'intervalles égaux. Il n'est même pas possible d'utiliser la transformation des rangs en échelle réduite selon C. L. Hull[2], car l'hypothèse d'une distribution normale des résultats de développement doit être rejetée avec une grande probabilité, car l'âge étant lui-même une variable du temps t, n'est pas, en général, distribué normalement et exigerait la transformation logarithmique $y = \log t$[3].

L'application du système des rangs n'est cependant pas décrite ici sous sa forme habituelle. Ce qui compte ici, c'est de donner par un chiffre ordinal le degré de développement.

[1] Stouffer Samuel A.: An overwiev of the contributions to scaling and scale theory. In Mesurement and prediction. p. 15.
[2] Hull C. L.: The Computation of Pearson's from Ranked Data. J. Appl. Psychol. 6/1922.
[3] Faverge J.-M.: Variables-critères, quantité du travail et qualité du travail. En: Ombredane/Faverge: L'analyse du travail. Paris 1955. pp. 64–65.

Résumé

Les échelles de Binet-Simon sont basées sur une conception quantitative, tandis que les échelles génétiques (encore théoriques) se basent sur une conception qualitative. Ceci revient peut-être à dire que les échelles de Binet-Simon ne tiennent pas compte de stades de développement proprement dits, qui sont à leur base. On illustre cette différence par la notion d'erreur dans les théories quantitatives et qualitatives. On estime la *grandeur* de l'erreur dans le cas des variables quantitatives, tandis qu'on apprécie le *nombre* d'erreurs (fréquence de la qualité absente ou présente) dans le cas des variables qualitatives.

Le maniement des techniques d'une analyse hiérarchique est plus simple que les techniques habituelles de prédiction à partir de variables quantitatives, car celles-ci nécessitent l'emploi des méthodes statistiques (calcul de la moyenne, de la dispersion, des corrélations, de l'analyse factorielle). Si les instruments statistiques se simplifient pour l'analyse hiérarchique, le chercheur devrait néanmoins consacrer plus de temps à la construction de ses échelles, car il ne suffira plus d'utiliser des procédés statistiques bien établis. Nous sommes ainsi pleinement d'accord avec Serge Moscovici, qui écrit: «La construction d'une échelle devient de ce fait une opération plus délicate, non pas mathématiquement, mais expérimentalement.»

Ce ne sont pas les échelles de développement de Binet-Simon qui sont critiquées, car elles ont rendu de très grands services et enrendront peut-être encore pour bien longtemps. Mais il leur manque un rattachement étroit à une conception adéquate d'une théorie d'intelligence génétique. Il nous semble que les modèles décrits permettent aux psychologues et aux pédagogues de reprendre certaines questions concernant le développement mental des enfants et son évaluation qualitative et quantitative.

Zusammenfassung

Die Skalen nach Binet-Simon stützen sich auf quantitative Überlegungen, während die (noch theoretischen) Entwicklungsskalen sich auf qualitative Überlegungen stützen. Das heißt vielleicht, daß die Skalen nach Binet-Simon die eigentlichen genetischen Studien nicht berucksichtigen, die ihnen zu Grunde liegen. Man illustriert diesen Unterschied durch den Fehlerbegriff in den quantitativen und qualitativen Theorien. Man schatzt die *Größe* des Fehlers im Falle der quantitativen Variablen, während man die *Zahl* der Fehler (Häufigkeit der vorhandenen oder nicht vorhandenen Qualität) im Falle der qualitativen Variablen einschätzt.

Das Handhaben der Techniken der hierarchischen Analysen ist einfacher als die üblichen Techniken der Vorhersagen von quantitativen Variablen aus, denn diese erfordern den Gebrauch statistischer Methoden (Berechnung der Mittelwerte, der Streuung, von Korrelationen, der Faktorenanalyse). Wenn die statistischen Instrumente sich fur die hierarchische Analyse vereinfachen, so mußte der Forscher zumindest für die Konstruktion solcher Skalen mehr Zeit aufwenden, denn es wird nicht mehr genugen, gut etablierte statistische Vorgehensweisen zu verwenden. Wir sind so ganz mit Serge Moscovici einverstanden, der schreibt: «Die Konstruktion einer Skala wird deshalb eine delikatere Operation, nicht mathematisch, sondern experimentell.»

Nicht die Entwicklungsskalen von Binet-Simon sind kritisiert; sie haben sehr große Dienste geleistet und werden es vielleicht noch lange tun. Aber es fehlt ihnen eine enge Verbindung zu einer adäquaten Auffassung einer genetischen Intelligenztheorie. Es scheint uns, daß die beschriebenen Modelle es den Psychologen und Pädagogen erlauben, einige Fragen betreffend der geistigen Entwicklung der Kinder und ihrer qualitativen und quantitativen Abschätzung aufzugreifen.