La théorie de la généralisabilité appliquée à un instrument de mesure des attitudes face à l´apprentissage d´une langue étrangère[1]

Gianreto Pini & Dagmar Hexel

Dans de nombreux domaines en sciences humaines et sociales, chercheurs et praticiens sont amenés à récolter et à traiter des données quantitatives. La collecte de ces données (au moyen de tests, d´épreuves, de questionnaires, de grilles d´observation ou d´analyse) exige la mise en place d´un dispositif approprié, dont le but est d´assurer la fiabilité des informations que l´on se propose d´obtenir.

Élargissant la notion classique de fidélité, la théorie de la généralisabilité permet de répondre à des questions du type: les mesures dont on dispose sont-elles suffisamment précises ? sont-elles «indépendantes» des instruments utilisés pour les recueillir ? sont-elles généralisables à l´ensemble des conditions d´observation possibles, dont chaque dispositif ne peut considérer qu´un échantillon particulier ?

La théorie de la généralisabilité n´est pas une méthode d´analyse au sens habituel du terme; son objectif est plutôt de vérifier si les données possèdent les caractéristiques métriques nécessaires pour qu´une utilisation ultérieure puisse être envisagée avec une sécurité suffisante, tant sur le plan de l´action (décisions consécutives à une démarche d´évaluation par exemple) qu´en ce qui concerne différents traitements statistiques (analyse factorielle, analyse de la variance, etc.).

Le but du présent article est d´illustrer l´intérêt et les apports de la généralisabilité lorsqu´on souhaite étudier les qualités techniques d´un instrument de recherche. A cet effet nous montrerons différentes applications de la méthode aux données fournies par une échelle que nous avons élaborée, et qui a pour but d´appréhender l´attitude des élèves de l´enseignement secondaire I face à l´apprentissage de l´allemand.

---

[1] Article publié dans *Éducation et recherche*, 20, 2/1998, 289-302 (actuellement *Revue suisse des sciences de l'éducation)*. Les données correspondant à celles analysées dans cet article peuvent être obtenues auprès du groupe Édumétrie, p.a. Daniel Bain, Moulin-Roget 49, CH-1237 Avully, e-mail : daniel.bain@bluewin.ch, ou téléchargées sur le site Web du groupe (en préparation) à la page « Exercices ».

Étude des qualités métriques d´un instrument de recherche

Dans le cadre méthodologique que les chercheurs en sciences humaines et sociales ont hérité des travaux de la psychométrie classique, les instruments d´observation et de mesure étaient généralement conçus pour étudier différentes caractéristiques d´un ensemble d´individus, comme leur niveau d´aptitude et de connaissance dans le domaine cognitif ou leur niveau d´attitude, d´intérêt, de motivation dans le domaine que l´on qualifie plutôt de socio-affectif.

Dans cette perspective, l´attention portée aux qualités techniques d´un instrument, et notamment à sa *fidélité*, avait principalement pour but de déterminer son aptitude à fournir une mesure «précise» (fiable, fidèle) des caractéristiques individuelles considérées. Plusieurs techniques étaient alors mises en oeuvre, qui n´avaient pas a priori de lien direct les unes avec les autres, et dont chacune permettait d´étudier un aspect particulier de l´instrument: la stabilité des résultats obtenus à des moments différents, l´équivalence de diverses versions dites parallèles ou l´homogénéité du répertoire d´items qui le composent (consistance interne). Toutes ces techniques reposaient, directement ou indirectement, sur l´application de méthodes corrélationnelles; de ce fait, elles prenaient en considération uniquement la situation des individus les uns par rapport aux autres (leurs positions «relatives» sur l´échelle de mesure), et non la valeur effective de leur score ou de leur résultat (positions «absolues»).

Dans le cas d´instruments dont tous les items devaient permettre d´appréhender une même caractéristique non directement accessible à l´observation (aptitude, attitude, motivation, intérêt, etc.), la procédure la plus courante était celle qui consistait à étudier leur consistance interne. On calculait alors le coefficient alpha (α) de Cronbach dont on peut montrer qu´il constitue une sorte de moyenne des différents coefficients de corrélation obtenus par la méthode *split-half* à partir de toutes les partitions possibles en deux groupes de l´ensemble des items. Cette approche permet de vérifier dans quelle mesure les différents items conduisent à un classement stable (donc fiable) des individus; plus exactement, elle permet de déterminer si la position des individus au sein de la distribution varie plus ou moins en fonction du répertoire d´items (conçu comme un échantillon aléatoire de tous les items théoriquement possibles) utilisé pour mesurer la caractéristique en question.

Ce type d´analyse met donc en jeu deux sortes d´éléments, qui jouent des rôles clairement dissymétriques: d'une part les individus, qui constituent les entités faisant

l´objet de la mesure, d'autre part les items, qui constituent l´outil à partir duquel la mesure est effectuée.

Dans le même cadre méthodologique, on pourrait envisager de renverser le problème, en considérant les items comme «objets» et les individus comme «instruments» de la mesure. Ce serait le cas par exemple si, à propos d´un test d´aptitudes ou de connaissances, on souhaitait classer les items en fonction de leur degré de difficulté. Le problème est alors de savoir si on peut distinguer (différencier) de manière fiable les items faciles des items difficiles, indépendamment des sujets auxquels le test est administré. Un tel changement de perspective n´a pourtant jamais été envisagé par la psychométrie traditionnelle. C´est la théorie de la généralisabilité, notamment grâce aux travaux de J. Cardinet et de Y. Tourneur (1985), qui a montré la pertinence et la richesse de cette approche, et qui a conçu les développements statistiques nécessaires pour pouvoir traiter adéquatement les problèmes auxquels elle ouvre la voie.

Notre intention n´est pas de donner ici une définition détaillée de la théorie de la généralisabilité, ni d´énumérer les différentes questions auxquelles elle permet de répondre ou l´éventail des domaines d´application possibles (Cardinet & Tourneur, 1985; Bain & Pini, 1996). Disons simplement que cette méthode élargit très nettement la perspective psychométrique classique en intégrant ses différentes approches (stabilité, formes parallèles, consistance interne) dans un même cadre général, au sein duquel elles n´apparaissent plus comme un recueil de techniques hétérogènes et disparates mais où elles sont unifiées dans une vision cohérente des problèmes relatifs à la fidélité de la mesure. L´élargissement que la généralisabilité apporte aux méthodes traditionnelles comporte différents aspects. Nous nous contenterons d´en signaler deux, que nous illustrerons ultérieurement à l´aide d´exemples concrets d´analyse.

D´abord, sur un plan que l´on qualifiera d´essentiellement technique, la théorie de la généralisabilité introduit *un double regard* sur la qualité des mesures fournies par un instrument. Le premier reprend dans son principe (mais avec des moyens plus puissants et plus féconds) la logique de la démarche traditionnelle, car il s´intéresse toujours à la stabilité des positions individuelles au sein de plusieurs distributions (par exemple, les positions d´un ensemble d´élèves en fonction des résultats obtenus aux différents items d´un test de connaissances). Le deuxième, en grande partie nouveau, concerne en revanche la précision des mesures en tant que telles: c´est-à-

dire la stabilité des scores eux-mêmes lorsqu´on passe d´un item à l´autre à l´intérieur d´un même instrument. La théorie de la généralisabilité conduit donc au calcul de deux indices et non plus d´un seul, qu´on appelle précisément des *coefficients de généralisabilité*, l´un *relatif* et l´autre *absolu*. Ainsi, un instrument peut présenter des caractéristiques métriques excellentes quant à son aptitude à classer de manière fiable un ensemble d´individus (c´est ce que le coefficient relatif permet de déterminer), mais des caractéristiques beaucoup moins satisfaisantes quant à la possibilité de fournir une estimation exacte (précise, fiable) du niveau effectif atteint par ces mêmes individus (coefficient absolu).

Sur un plan plus conceptuel, la théorie de la généralisabilité enrichit considérablement le point de vue instauré par l´approche classique, en introduisant un principe apparemment banal (mais souvent ignoré jusque-là), selon lequel les objets de la mesure peuvent être des entités autres que les individus, comme par exemple les items qui figurent dans une épreuve, les objectifs d´un programme de formation, différents domaines de compétences, les étapes d´un processus d´apprentissage, des méthodes pédagogiques. Dans cette perspective, on peut donc étudier l´aptitude d´un dispositif à différencier de manière fiable une série d´items en fonction de leur degré de difficulté, divers types de compétences en fonction du niveau de maîtrise atteint par les élèves, différentes étapes d´un parcours de formation en fonction du progrès enregistré lorsqu´on passe de l´une à l´autre, deux ou plusieurs approches pédagogiques en fonction de leur efficacité.

Dans la recherche en éducation (comme dans la recherche en sciences humaines de manière générale), cette nouvelle manière de concevoir les choses est évidemment très pertinente, car l´intérêt du chercheur ne concerne pas toujours les individus eux-mêmes, mais peut porter sur d´autres aspects qui caractérisent son champ d´investigation. Il est dès lors nécessaire qu´il puisse déterminer si les moyens mis en oeuvre présentent les propriétés techniques requises pour réaliser de manière fiable les observations ou les mesures qui l´intéressent. A cet égard il est opportun de signaler que, comme l´analyse le montre assez souvent, un instrument permettant d´obtenir une mesure satisfaisante des compétences individuelles ne fournit pas nécessairement une mesure fiable du niveau de difficulté de différents items. De même, un dispositif qui donne une estimation précise du progrès réalisé par une population d´élèves entre le début et la fin de l´année scolaire n´est pas forcément de nature à permettre une comparaison fidèle des résultats obtenus avec deux méthodes pédagogiques différentes.

Ce dispositif sera d´ailleurs souvent plus complexe que ceux auxquels nous venons implicitement de faire allusion. En effet, lorsqu´on s´intéresse par exemple à l´évolution des compétences entre deux moments différents (ou à l´évolution de toute autre caractéristique de même nature), le dispositif instrumental ne se réduira pas nécessairement à un ensemble plus ou moins vaste d´items, mais pourra aussi prendre en considération d´autres facteurs (en langage technique d´autres *facettes*) qui définissent les conditions dans lesquelles la mesure a lieu. On pourra donc y inclure non seulement des individus, mais également les classes et/ou les établissements auxquels ils appartiennent, des domaines de compétences, des filières d´étude.

Les avantages d´une telle approche sont considérables, car il devient possible de compléter l´appréciation globale sur les caractéristiques métriques du dispositif par un ensemble d´autres informations utiles. En particulier, on pourra identifier les facteurs qui influencent négativement la qualité de la mesure, et apprécier l´importance des perturbations introduites par différentes composantes du dispositif. Ainsi, l´étude des qualités techniques de l´instrument trouve tout naturellement un prolongement intéressant dans une démarche dite d´*optimisation*, qui permet de déterminer la nature et l´ampleur des modifications nécessaires pour que l´on puisse obtenir des mesures de meilleure qualité.

Dans les pages qui suivent, nous nous proposons d´illustrer quelques-uns des aspects qui viennent d´être évoqués, et nous le ferons en montrant différentes applications de la théorie de la généralisabilité pour étudier les qualités métriques d´une échelle d´attitude. Notre présentation comprendra:

- une description de l´instrument de recherche et du type de mesure qu´il doit permettre de réaliser;
- une brève comparaison entre les résultats fournis par la méthode classique (α de Cronbach) et par la théorie de la généralisabilité, en insistant tout particulièrement sur l´interprétation des coefficients relatif et absolu;
- une illustration des possibilités d´optimisation suggérées par les résultats de l´analyse;
- quelques indications sur la qualité des mesures fournies par l´instrument lorsqu´il est appliqué à différents objets d´étude.

Brève description de l´instrument

L´instrument dont nous avons analysé les qualités techniques est une échelle d´attitude de type Likert, conçue pour étudier l´attitude des élèves du Cycle d´orientation genevois face à la langue allemande. Cet instrument comprend deux sous-échelles qui devraient permettre de cerner l´attitude des élèves dans deux domaines différents: d´une part, l´attitude face à leurs expériences d´apprentissage de la langue (première sous-échelle) et, d´autre part, l´attitude face à la langue allemande en tant que telle, à son utilité et aux liens qu´elle devrait permettre d´établir avec la réalité et la culture germanophones (deuxième sous-échelle).

La première sous-échelle considère quatre aspects caractéristiques de l´expérience d´apprentissage des élèves, que nous avons définis de la manière suivante: (1) Réussite et progrès en allemand; difficultés d´apprentissage; (2) Plaisir et intérêt pour les leçons d´allemand; (3) Participation et investissement personnels; (4) Perception de l´enseignant et du contexte d´apprentissage.

En revanche, la deuxième sous-échelle a été élaborée en considérant les trois thèmes que voici: (5) Utilité de l´allemand; (6) L´allemand comme «vecteur culturel»; (7) Appréciation des caractéristiques de la langue.

L´instrument comporte en tout 28 items, quatre pour chacun(e) des thèmes (des catégories) qui précèdent, répartis aléatoirement au sein de la sous-échelle correspondante.

Chaque item énonce une opinion traduisant une attitude positive ou négative, favorable ou défavorable face aux différents aspects qui viennent d´être énumérés. Pour chaque énoncé, l´élève exprime son degré d´adhésion ou de rejet au moyen d´une échelle qui comporte quatre modalités de réponse: tout à fait ou assez d´accord; plutôt ou tout à fait en désaccord avec l´opinion correspondante.

Voici, à titre d´illustration, deux items (l´un exprimant une opinion «positive» et l´autre une opinion «négative») qui figurent dans l´instrument:

«En classe d´allemand je fais volontiers les exercices que l´enseignant nous donne.» (Première sous-échelle, catégorie 3)

«L´allemand est une langue désagréable à entendre.» (Deuxième sous-échelle, catégorie 7)

Par un procédé de codage prévoyant l´attribution aux différents items d´une «note» comprise entre 1 et 4 selon le sens (positif ou négatif) de l´opinion et la réponse du sujet, nous avons établi un score d´attitude pour chaque individu, situé sur une échelle allant d´un minimum de 28 à un maximum de 112 points. Ainsi, un score supérieur à 70 (point central de l´échelle) est censé traduire une attitude globale positive, d´autant plus nette que la valeur numérique du résultat est élevée. En revanche, les scores situés sur la partie inférieure de l´échelle révèlent une attitude plutôt négative, d´autant plus marquée que la valeur numérique est faible.

Le même procédé peut naturellement aussi être appliqué à chacune des deux sous-échelles. On obtient alors deux scores distincts, qui concernent l'un l´attitude des élèves face à leurs expériences d´apprentissage de l´allemand, l´autre l´attitude face à la langue et aux aspects qui lui sont associés. Les deux sous-échelles donnent ainsi lieu à des résultats compris respectivement entre 16 et 64 et entre 12 et 48 points.

Cet instrument a été administré deux fois à un échantillon d´environ 400 élèves du Cycle d´orientation: au début et à la fin de la 7e année (premier degré de l´enseignement secondaire I). Les élèves ont été recrutés dans deux des 17 établissements du canton et appartenaient soit à la filière pré-gymnasiale (orientation conduisant vers des études longues) soit à la filière non gymnasiale (orientation conduisant plutôt vers une formation de type professionnel).

Le but de la recherche était d´étudier *l´évolution des attitudes* entre les deux moments considérés. Sans entrer dans les détails (car ce n´est pas l´objectif de notre présentation) signalons que, globalement, nous avons constaté une baisse du niveau moyen d´attitude d´environ 9 points (67.7 et 58.7). Il semblerait donc qu´au terme de l´année scolaire, l´attitude des élèves soit sensiblement moins positive qu´elle ne l´était au début (différence statistiquement très significative).

Le coefficient ∝ de Cronbach et les coefficients de généralisabilité

Nous avons déjà rappelé que, parmi les méthodes développées dans le cadre de la psychométrie classique, le coefficient α de Cronbach constitue probablement l´indice le plus approprié pour ce genre d´instrument. Évaluant l´homogénéité du répertoire d´items auquel nous avons eu recours, il indique en effet si l´échelle est en mesure de produire un classement suffisamment stable des sujets malgré les fluctuations accidentelles engendrées par le choix aléatoire des items eux-mêmes.

Selon des critères qui, comme tous les critères statistiques de décision, reposent sur des considérations dictées par l´usage et l´expérience plus que sur des arguments théoriques véritablement rigoureux, on considère généralement que l´homogénéité d´un instrument est acceptable si la valeur du coefficient α (dont la marge de variation se situe entre 0 et 1) est égale ou supérieure à 0.80.

Appliquée aux résultats fournis par notre échelle lors de la première passation (début de l´année scolaire), cette méthode fournit un coefficient tout à fait satisfaisant, avec $\alpha = 0.889$. Ajoutons simplement que si l´analyse est faite en considérant séparément les deux sous-échelles, on obtient également des résultats supérieurs à la limite conventionnelle que nous venons d´évoquer (respectivement $\alpha = 0.825$ et $\alpha = 0.838$). Sur la base de ces résultats on peut donc considérer que l´instrument produit un classement fiable des sujets. De plus, cette conclusion s´applique non seulement à l´ensemble de l´échelle, mais également à chacune des deux sous-échelles qui composent l´instrument.

Comme nous l´avons déjà signalé, les même données peuvent aussi être analysées en appliquant la théorie de la généralisabilité. A ce propos, un premier point qu´il convient de préciser (et que nous nous limitons à évoquer dans le cadre de cette présentation) concerne la conception même de l´étude et la manière de définir le dispositif de mesure.

Le calcul du coefficient α est basé sur une démarche faisant intervenir deux sortes d´éléments seulement: les sujets d´une part et les items d´autre part, qui constituent respectivement les objets et l´instrument (le dispositif) de la mesure. La théorie de la généralisabilité conserve cette distinction fondamentale entre objets et dispositif instrumental, mais permet de prendre explicitement en considération toutes sortes d´éléments qui (outre les items) définissent le contexte de la mesure, et peuvent par conséquent affecter la précision (la stabilité, la fiabilité) des résultats.

Ainsi, dans l´exemple qui nous intéresse, on pourra tenir compte de deux aspects qui sont généralement ignorés par la théorie classique de la mesure:

- d´une part, la structure de l´instrument lui-même, c´est-à-dire la manière dont les items ont été choisis: chaque item appartient à une catégorie particulière; chaque catégorie appartient à l´une des deux sous-échelles (à l´un des deux domaines d´attitude);
- d´autre part, la structure de l´échantillon d´individus retenus pour cette étude, c´est-à-dire la façon dont les élèves ont été sélectionnés: chaque élève appartient à la fois à une filière et à un établissement donnés.

Sur la base de ces éléments, on peut établir un *plan* dit *d´observation*, dont le but est de décrire les relations entre les différentes composantes (facettes) qui caractérisent le contexte de la mesure. Précisons également que, dans cette phase de la démarche, l´utilisateur doit se conformer à un certain nombre de contraintes imposées par le modèle d´analyse. Pour cette raison, nous avons dû supprimer une catégorie d´items appartenant à la première sous-échelle (en l´occurrence celle qui concerne la perception de la réussite et du progrès en allemand). Par ailleurs, l´étude a été réalisée en considérant un échantillon aléatoire de 64 élèves sur les 400 qui avaient participé à la recherche.

Pour pouvoir réaliser l´analyse proprement dite, deux autres étapes sont nécessaires: l´élaboration d´un *plan d´estimation* (indiquant, pour chaque facette, si l´échantillonnage de ses modalités a été fait dans un univers aléatoire infini, aléatoire fini ou fixé) et d'un *plan de mesure* (distinguant les objets que l´on souhaite différencier des éléments du dispositif utilisé à cet effet).

A partir de ces différents plans la méthode fournit un ensemble de résultats, dont les plus caractéristiques et les plus facilement interprétables sont les *deux coefficients de généralisabilité* (*relatif* et *absolu*), que l´on désigne habituellement avec le symbole ρ^2 (rhô carré). Comme pour le coefficient α, on considère en général que l´instrument présente des caractéristiques satisfaisantes si la valeur de ces indices atteint ou dépasse la limite de 0.80 (80 % de la variance totale concerne les écarts «vrais» entre les éléments que l´on cherche à différencier).

Une première étude de généralisabilité a été faite à partir d´une situation analogue à la précédente (résultats obtenus lors de la première passation de l´échelle), avec les élèves comme objets de la mesure. On obtient alors les coefficients suivants:

$\rho^2_{rel.} = 0.892$ $\rho^2_{abs.} = 0.876$

On remarque tout d´abord que la valeur du coefficient relatif est très proche de celle du coefficient α calculé précédemment (0.889). Ce résultat n´est pas fortuit, car les deux indices présentent un certain nombre d´analogies et conduisent souvent à une conclusion de même nature. On peut d´ailleurs montrer que, dans certains cas particuliers (dispositifs comportant deux facettes seulement), les deux coefficients sont identiques.

Dans le langage propre à la généralisabilité, cette conclusion sera formulée en disant que le dispositif est apte à *différencier les sujets de manière fiable*: c´est-à-dire apte à les situer de façon suffisamment précise sur l´échelle définie par notre instrument de recherche. Une fois encore, la précision dont il est question ici concerne non pas les résultats obtenus par les élèves (la valeur effective des scores individuels), mais les positions relatives que ces élèves occupent les uns par rapport aux autres.

Ce que la généralisabilité apporte de nouveau est en revanche le coefficient absolu, qui nous renseigne sur la précision des résultats obtenus par les élèves. Comme on peut le constater, la valeur de cet indice dépasse largement la limite de 0.80. Ce résultat permet donc d´affirmer que, globalement, le dispositif instrumental fournit une mesure précise (fiable) du niveau d´attitude des élèves. Contrairement à la situation qui vient d´être décrite, la fiabilité de la mesure concerne ici la valeur effective (ou absolue) des scores eux-mêmes: pour chaque élève, le résultat que l´instrument lui attribue, compris entre 24 à 96 points.

Dans le cas d´une échelle d´attitude, ce type de précision n´est pas nécessairement exigé, car, dans la plupart des cas, les scores individuels ne sont pas utilisés en tant que tels. Dans d´autres situations en revanche, l´aptitude de l´instrument à fournir une estimation précise des résultats est nettement plus importante, comme par exemple lorsqu´on a recours à des tests ou à des épreuves de connaissance qui doivent permettre de fonder certaines décisions pédagogiques (comparaison des résultats individuels à un seuil de maîtrise ou à un critère de réussite dans des situations de notation, de certification, d'orientation ou de remédiation).

Peut-on estimer de manière fiable l'évolution du niveau d'attitude ?

Après cette première série d´analyses (dont l´objectif principal était d´établir une sorte de comparaison entre le coefficient α de Cronbach et les coefficients de généralisabilité), considérons maintenant une situation différente, qui permet d´illustrer un deuxième apport décisif de cette méthode statistique.

Rappelons tout d´abord que l´objectif de notre recherche était d´étudier l´évolution du niveau d´attitude entre le début et la fin de l´année scolaire. Sur le plan descriptif, ce type d´information est obtenu en calculant et en comparant les moyennes des deux moments considérés. Cependant, avant de procéder à des

analyses plus approfondies, il convient de vérifier si l´instrument est apte à *différencier* ces moyennes de manière fiable, c´est-à-dire apte à fournir une mesure précise de l´attitude des élèves à chacun des deux moments et/ou de son évolution entre le début et la fin de l´année scolaire.

Contrairement à la situation précédente, le problème n´est plus de savoir si l´instrument fournit une estimation fiable des résultats individuels, mais plutôt de vérifier s´il est apte à estimer de manière satisfaisante le niveau global d´attitude pour les deux moments considérés. De ce fait, ce ne sont plus les élèves qui constituent les objets de la mesure, mais les moments eux-mêmes.

Sur le plan méthodologique ce changement de perspective comporte des implications fondamentales. En effet, les qualités techniques d´un instrument ne sont plus évaluées en considérant uniquement son aptitude à différencier des individus, mais elles le sont désormais de cas en cas, en fonction des éléments auxquels on s´intéresse et dont on souhaite obtenir une mesure précise (des moments par exemple, mais aussi des objectifs, des items, etc.). Il n'est d´ailleurs pas inutile d´insister sur un point que nous avons déjà évoqué et qui sera brièvement illustré par la suite, selon lequel les utilisations d´un instrument ne sont pas interchangeables à souhait: d´où, précisément, la nécessité d´évaluer ses propriétés et ses qualités métriques en tenant compte de l´*utilisation exacte* que l´on souhaite en faire dans chaque cas particulier.

Après avoir défini convenablement les différents plans exigés par le modèle d´analyse (d´observation, d´estimation et de mesure), le logiciel calcule les coefficients de généralisabilité qui concernent la différenciation des deux moments:

$\rho^2_{rel.} = 0.979$ $\rho^2_{abs.} = 0.702$

Comme on s´en aperçoit aisément, le coefficient relatif est particulièrement élevé. On peut donc conclure que le dispositif permet de situer avec précision les deux moyennes l´une par rapport à l´autre. Puisque, dans ce cas, deux mesures seulement sont effectuées (une pour chaque moment), la conclusion de l´analyse peut être formulée très simplement de la manière suivante: le dispositif instrumental fournit une estimation tout à fait fiable de l´*écart* entre les résultats moyens obtenus au début et à la fin de l´année scolaire.

Rappelons à ce propos que nous avions observé une diminution du niveau moyen d´attitude d´environ 9 points entre les deux moments. Or, les résultats de l´analyse indiquent que *cet écart* est évalué de manière précise, l´estimation n´étant que très

légèrement affectée par les fluctuations imputables au choix aléatoire des établissements, des élèves et des items. On dira donc que la différence observée ne dépend pas (ou seulement de façon négligeable) des répertoires d´établissements, d´élèves et d´items retenus dans le cadre de cette étude. Selon les principes classiques du raisonnement inférentiel (dont la forme est ici plus proche de la démarche d´estimation que de celle qui est à la base des tests d´hypothèse usuels), on peut donc conférer à cette différence un caractère (une validité, une portée) *général(e)*.

Le coefficient absolu est en revanche beaucoup moins satisfaisant, sa valeur étant relativement éloignée de la limite généralement requise. Ce résultat ne permet donc pas de considérer comme fiable l´estimation du *niveau* d´attitude pour chaque moment, car la «marge d'incertitude» qui lui est associée (définie habituellement par l'étendue de l'intervalle de confiance autour de la moyenne) apparaît ici comme étant trop élevée: pour un niveau de confiance $1-\alpha = .95$ elle est en effet égale à $m \pm 5.7$. Par conséquent, *on ne généralisera pas* les résultats obtenus avec notre échelle (67.7 et 58.7) aux «univers» d´établissements, d´élèves et d´items correspondants.

Au-delà de ce résultat relativement global, la théorie de la généralisabilité fournit un certain nombre d´autres renseignements qui permettent de mieux cerner les facteurs responsables du phénomène. On peut en effet constater que la source principale d´«erreur» (variations des résultats engendrées par les procédures d´échantillonnage) réside dans l´écart relativement important entre les moyennes des deux établissements tous moments confondus (respectivement 60.6 et 65.8). Ce constat semble indiquer qu´il existe une différence non négligeable du niveau d´attitude lorsqu´on passe d´un établissement à l´autre. On peut donc en déduire que l´estimation des moyennes relatives aux deux moments risque de varier excessivement selon que l´étude est faite en considérant tel ou tel autre échantillon. C´est précisément en raison d´une telle *instabilité* que le résultat obtenu en considérant *deux établissements seulement* ne peut pas être généralisé.

Dans le cas qui nous intéresse, cette «faiblesse» du dispositif n´est pas véritablement gênante, car le but de la recherche était surtout d´estimer de manière fiable l´*évolution* (c´est-à-dire l´écart positif ou négatif) entre les moyennes des deux moments, et non la valeur absolue de ces moyennes. On peut toutefois imaginer des situations où les conséquences d´une telle limitation seraient beaucoup plus problématiques (conformité à un seuil ou à un critère par exemple). La question est

alors de savoir si le dispositif doit être écarté purement et simplement, ou *si l´on peut envisager des ajustements* susceptibles d´améliorer la qualité des mesures qu´il produit. Cet aspect du problème présente un intérêt évident lorsque (comme cela devrait être fait au moins dans certains cas) la première application de l'instrument n'a d'autres objectifs que d'étudier ses caractéristiques techniques et d'y apporter les modifications éventuellement nécessaires avant la mise au point de la version définitive.

Pour répondre à cette question, reprenons l´exemple que nous venons d´évoquer concernant la variabilité des résultats entre établissements. A ce propos, la théorie de l´estimation statistique fournit un moyen simple (bien que parfois coûteux sur le plan pratique) pour «stabiliser» l´estimation d´un paramètre relatif à une population qui présente une variabilité importante. En effet, puisque la précision de l´estimation est proportionnelle à $\sqrt{n}$, elle sera d´autant plus satisfaisante que l´effectif de l´échantillon est élevé. Or, un des avantages de la généralisabilité réside dans le fait qu´elle permet, à travers une simulation effectuée sur la base des données disponibles, d´estimer le coefficient que l´on devrait obtenir en modifiant le nombre d´éléments d´un ou de plusieurs échantillon(s). Grâce à cette démarche, qui constitue l´une des procédures d´*optimisation* possibles, la méthode fournit des indications précieuses sur la nature et sur l´ampleur des modifications nécessaires pour que le dispositif fournisse des mesures présentant une plus grande fiabilité.

Appliquée aux données de notre exemple, cette démarche montre qu´en considérant quatre établissements plutôt que deux, on obtiendrait un coefficient absolu de 0.803. Puisque cette procédure conduit à doubler le nombre total d´observations, il n´est pas sans intérêt de constater qu´on aboutirait à un résultat au moins acceptable (0.78) en doublant le nombre d´établissements (4) mais en réduisant de moitié le nombre d´élèves sélectionnés dans chacun d´entre eux (16 et non plus 32). On vérifie donc sans peine que la qualité des résultats ne dépend pas seulement du nombre total d´observations, mais également (et parfois surtout) des critères selon lesquels ces observations sont échantillonnées.

La fiabilité de l'instrument: une notion relative

Le dernier aspect que nous aborderons ici permet d´illustrer un point déjà évoqué précédemment, selon lequel les qualités d´un dispositif instrumental ne sont pas données une fois pour toutes, mais dépendent du type de mesure que l´on se propose d´obtenir.

Comme nous venons de le montrer, notre instrument fournit des mesures relatives d´excellente qualité, mais des mesures absolues nettement moins satisfaisantes, lorsqu´on souhaite appréhender le niveau d´attitude des élèves au début et à la fin de l´année scolaire. Imaginons maintenant que ce même instrument soit utilisé pour estimer le niveau d´attitude des élèves:

a) dans chacun des deux domaines (attitude face aux expériences d´apprentissage de l´allemand et face à la langue en tant que telle);

b) dans chacune des deux filières (pré-gymnasiale et non gymnasiale).

Voici les résultats fournis par le logiciel d´analyse:

a) Concernant les deux domaines : $\rho^2_{rel.} = 0.905$ $\rho^2_{abs.} = 0.809$

b) Concernant les deux filières : $\rho^2_{rel.} = 0.509$ $\rho^2_{abs.} = 0.226$

On peut ainsi constater que si le dispositif différencie de manière satisfaisante les domaines d'attitude aussi bien sur le plan relatif qu´absolu (moyennes de 35.0 et 28.2 respectivement pour la première et la deuxième sous-échelle), il en va tout autrement pour les deux filières, car les fluctuations aléatoires de l´échantillonnage sont beaucoup trop importantes par rapport à la différence observée entre les moyennes des deux groupes. Dans ce deuxième cas, l´instrument se révèle donc inapte à fournir une estimation fiable des résultats, et cette conclusion s´applique tout autant à l´évaluation de l´écart séparant les deux filières (différence de 3.4 points selon les résultats de notre étude) qu´à celle du niveau moyen d´attitude pour les élèves de chaque orientation (respectivement 64.9 et 61.5). Sur la base de ces résultats on peut donc conclure qu´un instrument de recherche n´est pas un outil «bon à tout faire». Il est en effet frappant de constater à quel point ses qualités techniques peuvent se modifier en fonction des objets soumis à la mesure et des conditions dans lesquelles cette mesure est réalisée.

Références bibliographiques utiles

Bain, D. & Pini, G. (1996). *Pour évaluer vos évaluations - La généralisabilité: mode d'emploi.* Genève: Centre de recherches psychopédagogiques du Cycle d'orientation.

Brennan, R.L. (1992). Elements of generalizability theory (2nd ed.). Iowa City, IA: ACT.

Cardinet, J. (1998). Von der klassischen Testtheorie zur Generalisierbarkeitstheorie: der Beitrag der Varianzanalyse. *Éducation et recherche, 20*, 2/1998, 271-288.

Cardinet, J. & Tourneur Y. (1985). *Assurer la mesure.* Berne: Peter Lang.

Cardinet, J., Tourneur, Y. & Allal, L. (1976). The symmetry of generalizability theory: applications to educational measurement. *Journal of Educational Measurement*, 13 (2), 119 - 135.

Cronbach, L.J., Rajaratnam, N. & Gleser, G.C. (1963). Theory of generalizability: a liberalization of reliability theory. *British Journal of Mathematical and Statistical Psychology*, 16 (2), 137 - 163.

Cronbach, L.J., Gleser, G.C., Nanda, H. & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley.

Cronbach, L.J., Linn,R.L., Haertel, E.H. (1997). Generalizability analysis for performance of student achievement or school effectiveness. *Educational and Psychological Measurement*, 57 (3), 373 - 399.

Anwendung der Generalisierbarkeitstheorie
auf eine Einstellungsskala zum Fremdsprachenlernen

Zusammenfassung

In verschiedenen Bereichen der Geistes- und Sozialwissenschaften haben Forscher und Praktiker mit quantitativen Daten zu tun. Das Erheben dieser Daten (mittels Tests, Prüfungen, Fragebögen, Beobachtungsrastern) verlangt nach einem geeigneten Verfahren, das die Zuverlässigkeit der eingeholten Informationen bescheinigt.

Die Generalisierbarkeit erweitert den klassischen Begriff der Zuverlässigkeit und erlaubt eine Antwort auf folgende Fragen: Sind die erhobenen Daten präzise genug? Sind sie "unabhängig" von den eingesetzten Messinstrumenten? Sind sie auf die Gesamtheit der Beobachtungskonditionen generalisierbar, von denen die gegebene Untersuchungssituation nur eine Stichprobe darstellt?

Die Generalisierbarkeitstheorie ist keine Methode zur Datenanalyse im eigentlichen Sinne; ihr Ziel ist vielmehr zu überprüfen, ob die ein Instrument die notwendigen psychometrischen Eigenschaften besitzt, die mit ausreichender Sicherheit seinen weiteren Einsatz erlaubt, sowohl in der Praxis (zum Beispiel bei Ausleseentscheidungen) als auch im Hinblick auf statistische Analysen (Faktorenanalyse, Varianzanalyse, usw.).

Der vorliegende Artikel zeigt, von welchem Interesse die Generalisierbarkeitstheorie beim Bescheinigen von Qualitäten eines Messinstruments ist und welchen Beitrag sie dazu leisten kann. Wir zeigen die verschiedenen Anwendungsmöglichkeiten dieser Methode anhand von Daten, die mit einer Attitüdenskala zum Fremdsprachenerlernen bei Schülern der Sekundarstufe I erhoben wurden.

La teoria della generalizzabilità applicata ad uno strumento di misura dell'atteggiamento nei confronti dello studio di una lingua straniera

Riassunto

In molti settori delle scienze umane e sociali, le esigenze della ricerca o dell'azione quotidiana richiedono talvolta la raccolta e l'analisi di dati quantitativi. Queste informazioni possono essere ottenute impiegando strumenti di vario genere (test, prove, questionari, eccetera), che devono essere concepiti in modo da assicurare alle misure o alle osservazioni ottenute un grado di affidabilità e di precisione soddisfacente.

La teoria della generalizzabilità è un modello statistico che permette di verificare in che misura il dispositivo utilizzato rispetta questa esigenza fondamentale, determinando in particolare se le informazioni raccolte presentano une validità generale che trascenda le condizioni specifiche e contingenti nelle quali esse sono state prodotte.

La teoria della generalizzabilità non è dunque un metodo di analisi dei dati nel senso usuale del termine. Il suo obiettivo è piuttosto quello di determinare se i dati raccolti presentano un grado di precisione tale da garantire al loro impiego la sicurezza necessaria tanto sul piano dell'azione (decisioni consecutive a certe forme di valutazione per esempio) che per quanto riguarda il ricorso a diverse analisi statistiche.

Lo scopo dell'articolo è di mostrare i vantaggi ed i contributi di questo metodo per verificare le qualità techniche e metodologiche di uno strumento di ricerca. Questo aspetto del problema è illustrato considerando i resultati di un'indagine condotta per analizzare l'atteggiamento degli allievi riguardo allo studio del tedesco durante l'insegnamento secondario I.

Generalizability theory applied to an attitude scale concerning foreign language learning

Summary

In different fields of the humanities and social sciences researchers and practitioners are confronted with quantitative data. These data are collected by various instruments (tests, questionnaires, scales), which have to be conceived so as to ensure satisfactory precision and reliability of the measures or informations obtained.

Enlarging the classical concept of reliability, the theory of generalizability makes it possible to verify to what extent the device respects these basic requirements. It determines in particular whether the informations obtained have a general validity beyond the specific conditions under which they were produced.

The theory of generalizability is therefore not a method of data analysis in the proper sense. Its purpose is rather to ensure that data possess the required metrical characteristics allowing further utilisation with sufficient security, be it for decisions taken on the ground of these measures or for statistical treatment (factor analysis, analysis of variance, etc.).

The aim of the present article is to illustrate the interest and the contribution of generalizability theory for the verification of the technical qualities of a research instrument. We show different applications of this method in the case of a scale designed to measure attitudes toward German language learning of pupils in lower secondary schools.