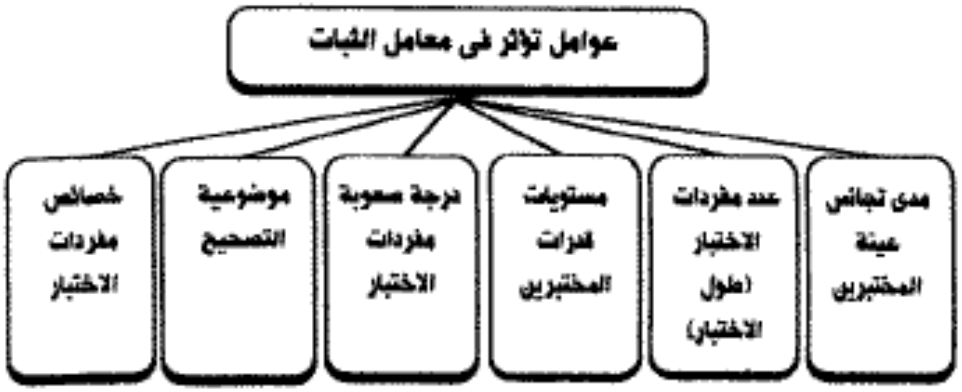

شكل (٤ ـ ١٠) يوضح بعض العوامل المؤثرة فى قيم معامل الثبات

وسوف توضح فيما يلى العوامل الموضحة فى شكل (٤ ـ ١٠)

مدى تجانس عينة المختبرين :

تعتمد القيمة التقديرية لمعامل الثبات اعتماداً كبيراً على مدى الفروق بين الأفراد المختبرين، فكلما زادت هذه الفروق ازداد تباين الدرجات الحقيقية للأفراد وبالتالى تزداد قيمة معامل الثبات. فالثبات كما أوضحنا يشير إلى اتساق قياس الفروق الفردية الحقيقية، فكلما زادت هذه الفروق يسهل الحصول على قياسات متسقة إذا تكررت عملية القياس على مجموعة الأفراد، فإذا كانت مجموعة الأفراد متجانسة فى القدرة أو السمة التى يقيسها الاختبار فإن تباين الدرجات الحقيقية يقل وبالتالى تنخفض قيمة معامل الثبات أو ربما تصبح مساوية صفراً . ولهذه الخاصية تطبيقاتها الميدانية عند استخدام درجات الاختبارات والمقاييس فى اتخاذ قرارات تتعلق بالطلاب أو العاملين الذين تكون قدراتهم متجانسة. فعندئذ ينبغى مراعاة انتقاء اختبارات اعتمد تقدير ثبات درجاتها على عينة من الأفراد مستوى قدراتها فى نطاق مستوى المجموعة التى نهتم بدراستها.

فإذا كانت قيمة معامل الثبات المذكورة فى دليل الاختبار تم تقديرها من مجموعة من تلاميذ الصفوف الثلاثة الأخيرة من مرحلة التعليم الأساسى، ولكننا طبقنا الاختبار بعد ذلك على تلاميذ صف واحد فقط، فإننا نتوقع أن تقل القيمة التقديرية لمعامل ثبات الدرجات عما هو مذكور فى دليل الاختبار نظراً لأن تلاميذ هذا الصف أكثر تجانساً من تلاميذ الصفوف الثلاثة.

عدد مفردات الاختبار:

سبق أن أوضحنا عند مناقشتنا معامل الاتساق الداخلي أنه كلما زاد عدد مفردات الاختبار أي طول الاختبار كلما زادت قيمة معامل ثبات درجاته. ويرجع ذلك إلى أن زيادة عدد المفردات يسمح للأخطاء العشوائية الموجبة والسالبة أن تتلاشى بعضها بعضاً مما يجعل الدرجة الملاحظة للفرد في الاختبار تقترب من درجته الحقيقية. وعند مناقشتنا صيغة سبيرمان وبراون Spearman - Brown Formula أوضحنا أنه يمكن تقدير القيمة المتوقعة لمعامل ثبات درجات الاختبار إذا أطلنا الاختبار أو جعلناه أقل طولاً. ولعل جدول (٤ ـ ١) يوضح كيف أن زيادة عدد مفردات الاختبار تؤدي إلى زيادة القيمة المتوقعة لمعامل الثبات.

مستويات قدرات المختبرين:

يؤثر مستوى قدرات المختبرين أيضاً في قيم معامل الثبات الذي نحصل عليه، فتباين درجات الخطأ يزيد بالنسبة للمجموعة منخفضة الدرجات حيث تلعب عوامل التخمين والصدفة دوراً أكبر في درجاتهم.

لذلك لا يجوز استخدام درجات مجموعة من مستوى قدرة معين للتنبؤ بثبات درجات الاختبار إذا طبق على مجموعة من مستوى آخر. فمثلاً إذا أردنا استخدام اختبار معين لانتقاء الأفراد في إحدى المهن الدنيا، أو أفراد مبتدئة في إحدى المهن يجب ألا يتأثر القرار بالبيانات المتعلقة بالثبات المستمدة من أفراد مهنة عليا أو أفراد من ذوي مهارات متقدمة. فالاختبار عندئذ تتباين درجة دقته في قياس هذه المستويات المختلفة. وهذا يتطلب مراجعة دليل الاختبار بعناية لتعرف مدى تأثر درجات الاختبار بتباين مستوى متوسط قدرات المختبرين.

درجة صعوبة مفردات الاختبار:

تؤثر درجة صعوبة مفردات الاختبار في قيم معامل الثبات، فإذا كانت المفردات غاية في السهولة أو الصعوبة، فإنه لا نستطيع باستخدامها قياس الفروق الفردية. ففي الحالة الأولى يستطيع كل فرد أن يجيب إجابة صحيحة عن جميع المفردات، والعكس في الحالة الثانية، وبذلك يكون توزيع الدرجات منتظماً في الحالتين. فدرجة صعوبة الاختبار ككل هي متوسط توزيع درجات الاختبار مقسوماً على عدد مفرداته، فكلما

اقتربت قيمة هذا المتوسط من عدد مفردات الاختبار كان الاختبار أكثر سهولة، وإذا تساوى المتوسط مع عدد المفردات كان الاختبار غاية فى السهولة، وعندئذ يستطيع كل فرد أن يجيب إجابة صحيحة عن جميع مفرداته ويصبح تباين الدرجات صفراً، وبالتالى تكون قيمة معامل الثبات صفراً أيضاً. وبالطبع تختلف الاختبارات فى مدى سهولة أو صعوبة مفرداتها، فالاختبارات مرجعية الجماعة أو المعيار Norm - Referenced Tests التى تصمم لقياس الفروق الفردية تكون مفرداتها عادة متوسطة الصعوبة لإبراز هذه الفروق، أما الاختبارات مرجعية المحك Criterion - Referenced Tests التى تصمم للأغراض التعليمية مثل الاختبارات التحصيلية الصفية، فإنها تهدف لقياس تمكن جميع الطلاب من مجال دراسى معين. لذلك فإن مفرداتها ربما تتميز بالسهولة النسبية لكى تقلل من عوامل التخمين الذى يؤدى إلى الأخطاء العشوائية فى الدرجات التى تسهم بدورها فى خفض قيمة معامل الثبات. ولذلك فإن تقدير ثبات درجات هذا النوع من الاختبارات باستخدام الطرق السابقة التى تعتمد على معامل ارتباط بيرسون لا تكون مناسبة، وسوف نشير إلى طرق أخرى تناسب الاختبارات مرجعية المحك فى الفصل الثامن.

موضوعية التصحيح:

سبق أن ناقشنا مفهوم الموضوعية Objectivity ،وتبين لنا أنه كلما تأثر تصحيح الاختبار أو تقدير درجاته بعوامل ذاتية أو عوامل التحيز انخفضت قيمة معامل ثبات الدرجات .

فتصحيح الاختبارات التى تشتمل على مفردات مثل الاختيار من متعدد، أو الصواب أو الخطأ، أو الإكمال يكون عادة موضوعياً سواء أجرى يدوياً أو آلياً. ولكن المشكلة تبدو واضحة فى تقدير درجة اختبارات المقال وبعض مقاييس الأداء والشخصية، حيث يتضمن التصحيح أحكاماً فردية تتعلق بنوعية الاستجابات، وهذا بدوره يؤثر تأثيراً بالغاً فى ثبات التقديرات. وقد سبق أن أوضحنا فى هذا الفصل إمكانية التغلب على هذه المشكلة عند مناقشتنا ثبات تقديرات المحكمين.

خصائص مفردات الاختبار:

تؤثر مفردات الاختبار في ثبات درجات الاختبار ككل، فخلو المفردات من الخطأ يعتمد على كيفية بناء هذه المفردات، فبعض المفردات ربما تشتمل على مؤشرات لإجابة مفردات أخرى في الاختبار مما يساعد بعض الأفراد على التخمين بدرجة جيدة مما يعمل على خفض قيمة معامل الثبات.

وكذلك المفردات الغامضة ، أو غير محددة الهدف ، أو التي تكون صياغتها أو تعليمات إجابتها غير دقيقة، أو غاية في الصعوبة، تؤثر تأثيراً بالغًا في ثبات درجات الاختبار.

ومن هذا يتضح أن قيم معامل الثبات ليست قيماً مطلقه وإنما تعد قيماً تقديرية تؤثر فيها عوامل متعددة ينبغي مراعاتها في تصميم أدوات القياس وعند انتقاء هذه الأدوات واستخدامها وتفسير نتائجها. فالفروض التي تستند إليها طرق تقدير الثبات يصعب تحققها جميعًا في الواقع الميداني، مما يتطلب الحيطة في تفسير قيم معامل الثبات. إذ لا يوجد أسلوب إحصائي يُغني عن الحكم المنطقي والفكر المستنير في تفسير هذه القيم وبخاصة في القياس التربوي والنفسي.

4. Facteurs affectant l'estimation de la fidélité des résultats

Les facteurs affectant l'estimation de la fidélité des résultats à un test proviennent de deux sources principales :

- les limites inhérentes au calcul de la corrélation linéaire au moyen du *r* de Pearson ;
- les conditions empiriques de l'administration du test, telles que la longueur du test et la limite de temps imposée.

Parce que, dans la pratique, l'estimation de la fidélité procède par un calcul de corrélation, les valeurs de fidélité dépendent du modèle de la corrélation linéaire de Pearson et des postulats de ce type de calcul statistique (voir Annexe 1). Les limites statistiques du *r* de Pearson s'étendent donc au coefficient de fidélité. Voici un bref rappel de ces limites dont il faut tenir compte dans toute interprétation d'un coefficient de fidélité.

4.1 La difficulté d'un test

Celle-ci affectera le calcul de la fidélité parce qu'un test trop facile ou trop difficile entraînera une certaine asymétrie des résultats : asymétrie positive dans le cas d'un test trop difficile, asymétrie négative dans le cas d'un test trop facile. Or, la corrélation *r* de Pearson ne peut atteindre sa valeur maximum de 1 que lorsque les distributions des deux variables en corrélation sont symétriques ou possèdent le même type d'asymétrie.

Prenons le cas du calcul d'un coefficient de stabilité au moyen de la corrélation test-retest. Dans la situation où les scores se distribuent de manière symétrique lors d'une première administration, puis de manière asymétrique lors d'une seconde administration, la valeur maximale du coefficient de corrélation entre les scores au test et au retest ne pourra atteindre la valeur maximum de + 1.

Il est donc important de prendre en considération les facteurs affectant la fidélité. Dans ce dernier cas, il est tout aussi important – sinon plus – de savoir que la distribution des scores a changé que de savoir que la valeur de stabilité est faible. En effet, le changement de distribution peut expliquer pourquoi la fidélité est faible. Un test devenu trop facile au moment du retest peut expliquer que la distribution des résultats, symétrique au moment du test, soit devenue asymétrique négative au moment du retest. La contamination des résultats ou l'apprentissage peuvent expliquer ce genre de phénomène.

4.2 L'étendue des différences individuelles

La variance totale d'un test est une condition nécessaire, mais non suffisante à la fidélité des résultats. C'est ce que nous avons vu en traitant de la variance du score total à un test. Toute réduction de l'étendue des scores individuels entraîne une sous-estimation de la corrélation entre deux variables (voir Annexe 1).

Lors de l'étude de la fidélité d'un instrument de mesure, plusieurs situations peuvent se produire contribuant à réduire les différences individuelles et, par conséquent, nos chances d'obtenir une estimation correcte de la fidélité. C'est le cas, notamment, des situations suivantes :

1. L'étude-pilote porte sur un échantillon qui possède une variance moindre que la population générale. C'est le cas d'un test dont les résultats ne sont recueillis que dans des écoles provenant de milieux favorisés. On peut suspecter que la variance des résultats ainsi recueillis est moindre que celle qui aurait été obtenue au moyen d'un échantillon représentatif.
2. Un test a été mis à l'essai sur une population scolaire à plusieurs niveaux, plus étendue que le seul niveau dans lequel le test doit être employé. Il faut être prudent dans l'appréciation de la fidélité rapportée dans de telles conditions.

Les résultats peuvent donner lieu à une variance des scores qui soit artificiellement grande lorsque les répondants sont de plusieurs niveaux scolaires. Par contre, cette variance risque d'être réduite, et la fidélité de même, si l'on emploie le test à un seul niveau scolaire.

Magnusson (1967) a mis au point une formule permettant de corriger l'estimation de la fidélité lorsque nous avons de bonnes raisons de croire que notre échantillon de sujets est homogène et contribue ainsi à sous-estimer la variance totale des scores observés au test. Cette formule de correction est donnée par l'équation suivante :

$$r_{UU'} = 1 - \frac{s_X^2(1 - r_{XX'})}{s_U^2} \tag{3.45}$$

Dans cette équation, $r_{UU'}$ est la fidélité estimée pour le nouvel échantillon U, s_X^2 est la variance de l'échantillon pour lequel nous avons déjà calculé la fidélité, s_U^2 est la variance du nouvel échantillon et $r_{XX'}$ est la fidélité estimée à partir de l'échantillon de départ X.

Cette correction de Magnusson postule que l'erreur aléatoire est la même dans les deux groupes et que la différence dans les variances des scores observés est imputable à des différences dans les variances des scores vrais dans les deux groupes. C'est pourquoi, lors de l'utilisation de normes, il est important de s'assurer que notre échantillon provient de la même population qui a servi au calcul des valeurs de la fidélité des résultats, sinon il sera plus prudent de réaliser une étude-pilote sur la fidélité des résultats obtenus avec l'échantillon concerné.

4.3 Limite de temps

Lorsqu'un test est chronométré, plusieurs élèves n'arrivent pas à répondre à toutes les questions dans le temps imparti. Les questions omises se trouvent généralement à la fin du test et celles-ci sont généralement cotées 0. Cette procédure a pour effet de créer une inflation artificielle de la corrélation entre les derniers items, ce qui aura pour effet de faire paraître ces items plus homogènes qu'ils ne le sont en réalité. Cette homogénéité ne sera pas due au fait que les items mesurent la même chose, mais plutôt au fait qu'ils ont été omis par les sujets parce qu'ils se trouvaient en fin de test.

Il faut donc être très prudent lorsque l'on administre un test chronométré et que l'on souhaite déterminer la fidélité des résultats. L'estimation de la fidélité risque d'être faussée par la corrélation artificielle entre les items dans le cas des méthodes de bissection ou encore de cohérence interne (α de Cronbach). Dans des conditions identiques, par contre, les résultats obtenus par la méthode test-retest ne sont pas affectés.

4.4 La longueur du test

Plus un test comprend un grand nombre d'items correspondant à ce que nous souhaitons mesurer, plus cette mesure devrait être précise. En effet, la somme des erreurs aléatoires de mesure devrait tendre vers zéro lorsqu'un grand nombre d'items est utilisé. C'est le principe de la théorie de l'échantillonnage : plus un échantillon est grand, plus l'estimation des caractéristiques de la population dont il est tiré tend à être précise.

Le rapport entre la longueur d'un test et la fidélité de ses résultats est exprimé par la formule de Spearman Brown (*Spearman Brown prophecy formula*). Elle nous indique à quel degré de précision l'on peut s'attendre de scores qui seraient calculés à partir d'un nombre accru d'items dans une proportion k (k pouvant être une fraction ou un entier). Voici un rappel de cette formule que nous avons déjà vue dans le cas de la méthode de bissection où $k = 2$ (formule 3.33) :

$$r_{xx'} = \frac{kr_{tt'}}{1 + (k - 1)r_{tt'}} \tag{3.46}$$

Dans l'équation précédente, $r_{xx'}$ représente la fidélité attendue du test modifié, $\rho_{tt'}$ représente la fidélité du test initial. Lorsque $k > 1$, nous calculons la fidélité pour un test allongé. Par exemple, si un test comporte 12 items et que l'on souhaite connaître la fidélité de ce test auquel nous avons ajouté 18 items parallèles, soit 30 items en tout, alors nous utilisons la formule (3.46) avec $k = 2,5$ ($2,5 \times 12 = 30$). Le même principe s'applique pour $k < 1$. Les valeurs de fidélité calculées le sont alors pour des tests plus courts.

La formule de Spearman-Brown nous permet de déterminer dans quelle proportion la longueur d'un test doit être augmentée pour atteindre un degré visé de fidélité. En modifiant l'équation précédente, l'on peut isoler k de la façon suivante :

$$k = \frac{r_{xx'}(1 - r_{tt'})}{r_{tt'}(1 - r_{xx'})} \tag{3.47}$$

Supposons que l'on veuille estimer dans quelle proportion un test de 30 items doit être prolongé pour que sa fidélité, actuellement de 0,75, soit portée à 0,85. En solutionnant l'équation (3.47) pour trouver k, on obtient :

$$k = \frac{0,85(1 - 0,75)}{0,75(1 - 0,85)} = 1.89$$

Une valeur $k = 1,89$ signifie que le nouveau test devra être 1,89 fois plus long que le test original. Il devra donc compter approximativement $1,89 \times 30$ items, soit 57 items. Il faudrait donc ajouter 27 items aux 30 items faisant déjà partie du test pour faire passer la fidélité du test de 0,75 à 0,85.

Il est important de se rappeler que la formule de Spearman Brown prend pour acquis que les items qui seront ajoutés (ou retranchés) sont parallèles aux items du test de départ, c'est-à-dire qu'ils sont de même contenu et de même degré de difficulté. En effet, la précision d'un test n'augmentera pas si l'on y ajoute des items de niveaux de difficulté fort différents ou de contenus variés, susceptibles de ne pas avoir une bonne corrélation avec les items faisant déjà partie du test.

La formule de Spearman Brown peut être très utile pour nous permettre de décider de la longueur qu'un test doit avoir pour posséder une précision acceptable. Cependant, cette méthode ne nous indique pas quelles sont les caractéristiques des items parallèles à ajouter, en termes de contenu et de format, afin d'accroître la fidélité des tests. Lorsque le contenu d'un test est défini de façon générale, comme c'est le cas de plusieurs épreuves sommatives en éducation et de certains tests psychométriques, le constructeur peut éprouver de la difficulté à définir les caractéristiques des items à ajouter pour qu'ils soient parallèles à ceux déjà construits. En éducation, par exemple, le concepteur pourra s'inspirer des objectifs pédagogiques pour ajouter

des items provenant des mêmes objectifs que le test initial. Plus les conditions ayant présidé à l'élaboration initiale du test sont claires, comme c'est le cas avec les techniques de spécification de domaine, plus il sera facile au concepteur de rédiger des items parallèles.

Le principal inconvénient de cette manière de procéder est d'employer une approche empirique pour créer des ensembles homogènes d'items. Il est possible que certains items possèdent des caractéristiques qui leur permettent de mesurer de façon plus précise les sujets d'un échantillon particulier. Il est plus facile d'améliorer la fidélité d'un test lorsque celui-ci a été construit selon des facettes ou une approche critériée (voir chapitre 1) et lorsque les caractéristiques de ces items sont bien connues. De plus, des tests construits selon de telles facettes se prêtent bien à une étude de généralisabilité (voir section 7 de ce chapitre).