Movement & Sport Sciences – Science & Motricité 83, 1–10 (2014)

DOI: 10.1051/sm/2012042

La fiabilité de la mesure en sciences du sport : une approche statistique générale de l'erreur de mesure

Stéphane Champely, Yohann Blache, Mickaël Campo, Samuel Rota, Karine Monteil et Isabelle Rogowski

Université Lyon 1, EA 647, Centre de Recherche et d'Innovation sur le Sport (CRIS), 27-29 boulevard du 11 Novembre 1918, 69622 Villeurbanne Cedex, France

Reçu le 27 Octobre 2011 – Accepté le 19 Décembre 2012

Résumé. Une méthode statistique générale, y compris graphique, de l'étude de l'erreur de mesure (fiabilité) pour des données quantitatives est proposée. Les trois modèles probabilistes de Shrout et Fleiss (1979) sont repris et illustrés à l'aide d'exemples réels (biomécanique et psychologie du sport). À partir des paramètres de ces modèles sont définis divers indicateurs de l'erreur : coefficient de corrélation intra-classe, erreur standard de mesure, coefficient de variation, limites de concordance. Les trois modèles probabilistes étant des cas particuliers du modèle linéaire à effets mixtes, leurs paramètres peuvent être estimés par la méthode du maximum de vraisemblance restreint. Les avantages en sont la cohérence théorique et informatique, la possibilité de calculer des intervalles de confiance et de gérer les données manquantes.

Mots clés : Coefficient de corrélation intra-classe (ICC), erreur standard de mesure (SEM), coefficient de variation (CV), modèle linéaire à effets mixtes

Abstract. Measurement reliability in sports sciences: a general statistical approach of measurement error.

A general statistical approach to study measurement error (reliability) of quantitative data is proposed, including graphical analyses. The three probability models by Shrout and Fleiss (1979) are used and illustrated with sports data (psychology, biomechanics). By using the models parameters, several measurement indexes of error can be defined: intraclass correlation coefficient, standard error of measurement, coefficient of variation, limits of agreement. The three probability models are special cases of the linear mixed-effects model whose parameters can be estimated through restricted maximum likelihood. The main advantages of this approach are its theoretical and computing coherency, that datasets with missing values could nevertheless be analysed and confidence intervals computed.

Key words: Intraclass correlation coefficient (ICC), standard error of measurement (SEM), coefficient of variation (CV), linear mixed-effects model

1 Introduction

La qualité d'une étude scientifique dépend évidemment de ses données, et donc de celle des mesures effectuées. Si ces dernières sont toujours des approximations, certaines sont plus acceptables que d'autres. La *fiabilité* de la mesure est la constance avec laquelle cette mesure évalue « quelque chose » (Hopkins, 2000). La fiabilité est caractérisée par l'*erreur de mesure*, définie dans Bland et Altman (1996, p. 41) comme « la variation entre des mesures de la même quantité sur le même sujet »... dans des conditions identiques.

L'erreur de mesure peut être parfois le cœur de l'étude, pour énoncer par exemple des normes de cette erreur à respecter (contrôle qualité). Dans le cadre des recherches en sciences du sport, si ces erreurs de mesure ont un impact direct sur l'étude des changements au cours de mesures répétées, elles sont également une préoccupation pour quiconque s'intéresse à une mesure unique sur le sujet (Hopkins, 2000). En effet, l'erreur de mesure

« s'ajoute » aux variations aléatoires de l'expérience, diminue la puissance statistique et complique ainsi la détection d'une différence. Quantifier l'erreur de mesure est le premier pas avant d'entreprendre des actions correctrices pour la réduire et par conséquent améliorer le protocole d'étude. La caractérisation de l'erreur de mesure permet également pour des applications dans le cadre décisionnaire de l'entraînement sportif d'identifier une évolution individuelle dès lors qu'elle peut être distinguée d'un simple bruit et traduire une variation significative. Comme le soulignent Atkinson et Nevill (1998), il est nécessaire d'utiliser des outils suffisamment sensibles afin de détecter des changements subtils de performances, mais pouvant cependant « faire la différence » chez des sportifs de haut niveau.

Quel qu'en soit l'objectif, pour mener à bien une étude complète de fiabilité ou pour une simple vérification dans le cadre d'une recherche, il faut répéter la mesure un nombre *raisonnable* de fois sur un nombre *raisonnable* de sujets (Hopkins, 2000), ceci afin d'atteindre une puissance... raisonnable. Pour ce faire, le dispositif expérimental le plus simple est le *test-retest* : la mesure est répétée deux fois chez les mêmes sujets[1]. Plus généralement, le choix du dispositif (en mesures répétées) dépend des sources de variation possibles que sont l'instrument de mesure, les opérateurs et la variabilité inévitable du sujet.

Il convient de distinguer cette situation d'étude de fiabilité d'une seule méthode de mesure de celle de la comparaison de méthodes où sont testées sur les mêmes sujets une méthode standard mais complexe et une méthode nouvelle, souvent plus économique, et où l'on cherche à savoir si elles sont en concordance (*agreement*). Même si les plans expérimentaux et parfois les techniques statistiques employées sont proches, les objectifs en sont bien différents (Newell, Aitchinson, & Grant, 2010).

En ce qui concerne l'analyse statistique de l'erreur de mesure, comme le remarquent Atkinson et Nevill (1998), l'accord est en revanche moins net sur ce qu'il convient de faire. D'une part, il existe plusieurs indicateurs de fiabilité, d'autre part, le calcul même de ces indicateurs peut différer, enfin l'interprétation des résultats elle-même reste confuse (Weir, 2005). On peut caricaturer la littérature sur le sujet en la réduisant à deux extrémités suivant que des modèles probabilistes pour les données sont ou non employés. D'une part, certains articles à visée pédagogique, tendent à éluder de tels modèles et présentent directement une collection de formules de calcul de différents indicateurs de fiabilité. Ces derniers y apparaissent un peu juxtaposés, les liens entre eux restant peu explicités, et surtout, à la surprise des lecteurs, les méthodes de calcul peuvent varier d'une présentation à l'autre. D'autre part, des articles publiés dans des revues plus théoriques, reposent sur des modèles probabilistes mais soit se limitent à des variations autour d'un indicateur de fiabilité spécifique (Shrout & Fleiss, 1979), soit se concentrent sur un modèle probabiliste spécifique (Quan & Shih, 1996), soit, datant de quelques années, ne pouvaient alors recourir à des techniques modernes d'estimation ou des procédures informatiques sophistiquées.

L'objectif de cet article est dans un premier temps de partir du très simple modèle probabiliste dit du vrai score (*true score theory*), en montrant que ce modèle peut être facilement étendu pour inclure les trois situations classiques de collecte de données décrites dans Shrout et Fleiss (1979). Dans un second temps, les principaux indicateurs de fiabilité : coefficient de corrélation intra-classe (ICC, *intraclass correlation coefficient*), erreur standard de mesure (SEM, *standard error of measurement*), limites de concordance (LOA, *limits of agreement*) et coefficient de variation (CV), seront construits comme fonctions des paramètres de ces modèles. Dans un troisième temps, l'emploi d'une méthode générale pour l'estimation de ces paramètres, permet également d'estimer les indicateurs de fiabilité qui en dépendent.

Partir d'un modèle probabiliste présente de grands avantages : 1) une unité conceptuelle plus grande, les relations entre les différents indicateurs et les méthodes de calcul qui en découlent se dessinant plus clairement, 2) il est possible de calculer des intervalles de confiance des indicateurs comme conseillé dans Morrow et Jackson (1993), 3) la présence de données manquantes peut être gérée et 4) des situations de collecte de données plus complexes sont envisageables.

Ces propositions seront appliquées à trois jeux de données représentatifs des sciences du sport qui illustrent en particulier les variations possibles du modèle probabiliste sous-jacent. Toute entreprise de modélisation doit également vérifier si les données sont en adéquation avec le modèle probabiliste postulé. Cette adéquation est souvent testée par des méthodes graphiques, comme le font notamment Altman et Bland (1983) dans le cas simple du test-retest. Leurs propositions seront étendues au cas général car les représentations visuelles, souvent révélatrices, restent peu connues et donc peu utilisées dans le domaine de l'analyse statistique de l'erreur de mesure.

2 Modèles probabilistes pour quantifier l'erreur de mesure

2.1 Le modèle du vrai score

Le dispositif expérimental classique utilisé pour quantifier l'erreur de mesure consiste à répéter la même mesure y_{ij} sur un certain nombre de sujets i ($i = 1,\ldots, N$) avec

[1] Dans le cadre de cet article, les études de fiabilité, courantes en psychologie et en marketing, portant sur des items différents basés généralement sur des échelles de Likert et qui servent à mesurer la même variable latente, ne seront pas considérées. On trouvera un rapide exposé des méthodes spécifiques : *formes alternatives*, *split-halves*, *alpha de Cronbach* voire *measurement model* dans Bollen (1989).

n_i mesures par sujet. Le nombre de mesures n'est pas toujours constant, et ces dernières ne sont pas forcément comparables d'un sujet à l'autre.

En reprenant la définition de l'introduction, « [l'erreur de mesure est] *la variation entre des mesures de la même quantité sur le même sujet* », le modèle du vrai score postule que la mesure obtenue, seule observable, est fonction additive d'un vrai score du sujet (V) et d'une erreur (E) :

$$y_{ij} = v_i + e_{ij}$$

noté modèle (1,1) chez Shrout et Fleiss (1979), où v_i suit une loi normale d'espérance μ et d'écart-type σ_v alors que e_{ij} suit indépendamment une loi normale d'espérance 0 et d'écart-type σ.

Les vrais scores des sujets ne nous intéressent pas « individuellement » en l'espèce, ils sont en fait interchangeables, il s'agit juste d'en tenir compte pour définir correctement l'erreur de mesure : σ. C'est pourquoi ils sont modélisés globalement par une loi normale dépendant de deux paramètres μ et σ_v. On parle d'un *effet aléatoire*.

Soulignons que l'on suppose ici que l'erreur est caractéristique de la mesure, et donc ne dépend pas de l'individu (homoscédasticité versus hétéroscédasticité), ce qui peut pourtant arriver au moins sous deux formes : une stratification des individus, certains (par exemple les plus entraînés) ont une erreur différente, ou bien l'erreur dépend du vrai score de l'individu, phénomène (que nous observerons plus avant) qui peut s'obtenir par une fonction multiplicative du score et de l'erreur. La normalité du vrai score et de l'erreur reste également une hypothèse.

2.2 Prise en compte d'un effet condition

La définition complète de l'introduction était : « [l'erreur de mesure est] *la variation entre des mesures de la même quantité sur le même sujet… dans des conditions identiques* ». Parfois, les mesures prises sur un sujet i sont comparables à celles prises pour un autre sujet k : y_{ij} est alors comparable à y_{kj}, car il s'agit du même observateur j (de la même date j… de façon générale de la même condition expérimentale j) évaluant les sujets i et k. Dans ce cas, les conditions ne sont pas exactement identiques, il est possible que des variations systématiques de ces mesures se produisent, ce qu'on appelle un biais (B).

La difficulté est qu'on peut selon les cas envisager cet effet condition de deux façons différentes, comme un effet aléatoire ou comme un effet fixe. Si les conditions sont interchangeables (des observateurs en particulier) ou qu'elles ne nous intéressent pas individuellement, on souhaite alors simplement en tenir compte pour estimer correctement l'erreur de mesure, et éventuellement avoir une idée de la variation des biais dans une « population » de conditions. Il s'agit alors de la version aléatoire qui donne :

$$y_{ij} = v_i + b_j + e_{ij}$$

noté modèle (2,1) chez Shrout et Fleiss[2] (1979), avec b_j qui suit une loi normale d'espérance 0 et d'écart-type σ_b, modélisant une « population » de conditions indépendamment de v_i et e_{ij}.

Au contraire, si les conditions ne sont pas interchangeables, nous devons alors privilégier une version avec un effet condition fixe. C'est le cas de mesures effectuées à différentes dates, où l'intérêt porte le plus souvent sur la comparaison de dates successives, mais de façon générale de conditions expérimentales intrinsèquement intéressantes. Le modèle devient alors

$$y_{ij} = v_i + \beta_j + e_{ij}$$

noté (3,1) chez Shrout et Fleiss (1979), les β_j étant des paramètres fixes correspondant à l'effet des conditions, les deux autres éléments ne sont pas modifiés.

3 Les indicateurs de l'erreur de mesure

Classiquement, on distingue les indicateurs *absolus* qui décrivent comment les mesures prises sur un même sujet varient (éventuellement en %) et les indicateurs *relatifs* qui permettent d'observer si les sujets étudiés peuvent être bien différenciés (entre eux).

3.1 Les indicateurs absolus

À partir des modèles probabilistes précédents, et plus précisément de leurs paramètres, il est possible de définir plusieurs indicateurs de l'erreur de mesure. L'erreur standard de mesure (*SEM*), parfois appelée erreur typique, est simplement

$$SEM = \sigma, \tag{1}$$

quel que soit le modèle probabiliste sous-jacent. Elle est exprimée dans l'unité de mesure et est directement comparable entre études. Elle réclame néanmoins, pour décider de sa qualité dans le cadre d'une seule étude, d'avoir préalablement fixé des objectifs analytiques (*analytic goals*).

Lorsqu'on s'intéresse aux variations des mesures prises sur le même individu (définition de l'erreur de mesure), il est également possible de raisonner en termes de différences entre deux mesures effectuées sur le même sujet. S'il n'y a pas de biais de mesure, son espérance est nulle et son écart-type est

$$\sigma_D = \sqrt{2}\sigma. \tag{2}$$

Cet écart-type est employé dans la méthode des limites de concordance de Bland et Altman (1986), et de la

[2] On remarquera que les modèles employés par Shrout et Fleiss (1979) contiennent de plus une interaction. Elle est ici inutile car dans les plans d'expériences envisagés il est impossible de l'estimer et elle reste indiscernable de l'erreur de mesure.

différence minimale (*minimal difference*), en le multipliant par 1,96.

Afin de rendre le *SEM* indépendant des unités de mesures, il est rapporté à la moyenne des mesures dans le coefficient de variation :

$$CV = \frac{\sigma}{\mu} \tag{3}$$

pour les modèles (1,1) et (2,1). En ce qui concerne le modèle (3,1), il existe une ambiguïté car on peut le rapporter à différentes espérances $(\mu + \beta_j)$ selon la condition expérimentale considérée. S'il existe une condition standard, il semble judicieux de la choisir comme référence.

3.2 Les indicateurs relatifs

L'indicateur relatif classique est le coefficient de corrélation intra-classe. Il se définit comme la corrélation entre deux mesures effectuées sur le même sujet, corrélation qui s'avère ne pas dépendre des deux mesures choisies sur les trois modèles envisagés. En revanche, l'expression de ce coefficient dépend du modèle :

$$ICC(1,1) = ICC(3,1) = \frac{\sigma_v^2}{\sigma_v^2 + \sigma^2} \tag{4}$$

et

$$ICC(2,1) = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_b^2 + \sigma^2}. \tag{5}$$

Bien que ces coefficients soient compris entre 0 et 1, il est difficile de dire à partir de quel seuil ils sont intéressants, et de les comparer entre études. En effet, par définition, ces quantités dépendent de l'hétérogénéité de la population étudiée par le truchement de σ_v^2. Si une population suffisamment diversifiée est testée, cas le plus fréquent dans les études de fiabilité comme le remarquent Atkinson & Nevill (1998), l'*ICC* est immanquablement proche de 1 et manque de surcroît de sensibilité par rapport au *SEM* (qui peut doubler alors que l'*ICC* en correspondance change assez peu). La comparaison, en se basant uniquement sur le coefficient intra-classe, de deux études avec des populations très différentes (performances de l'élite très resserrées et performances d'amateurs très étalées) a par conséquent peu de sens.

Le coefficient de corrélation linéaire r est parfois employé pour quantifier la fiabilité. Il souffre, comme l'ont démontré Altman et Bland (1983), de la même dépendance à l'hétérogénéité des sujets, mais plus encore, du fait d'être un coefficient d'association plus que de fiabilité. En effet, on peut bien multiplier l'une des deux mesures, ou lui ajouter une quantité quelconque, la valeur de r restera la même, tandis que les deux mesures n'auront alors plus grand-chose à voir ! S'il peut devenir sensible au biais, dans la version du coefficient de concordance proposée par Lin, Hedayat, Sinha et Yang (2002), cela ne règle pas pour autant le fait qu'il reste limité au cas de deux mesures, bien identifiées, ce qui n'est pas forcément le cas notamment pour le modèle (1,1).

4 Estimation des paramètres et des indicateurs

Les modèles (1,1), (2,1) et (3,1) sont en fait des instances du modèle linéaire à effets mixtes (Pinheiro & Bates, 2000). Leurs paramètres (et par conséquent les indicateurs de l'erreur de mesure, qui sont fonctions de ces paramètres) peuvent être estimés par la méthode du maximum de vraisemblance ou celle du maximum de vraisemblance restreint (REML) que suggère Hopkins (2000). Cette dernière sera employée car elle permet de retrouver exactement les résultats classiques de Shrout et Fleiss (1979) dans le cas de plans d'expériences équilibrés. Surtout, cette méthode est applicable, même avec des données manquantes ou un plan d'expériences volontairement déséquilibré.

Si des méthodes asymptotiques existent pour définir les intervalles de confiance des paramètres, nous leur préférerons, sachant que les échantillons sont parfois de petite taille dans ces études, une méthode de rééchantillonnage : le *bootstrap paramétrique* (Efron & Tibshirani, 1993). Il s'agit d'utiliser l'estimation du modèle, obtenue à partir des données originales, pour simuler de nouvelles données y_{ij} selon un plan identique, et d'en estimer à nouveau les paramètres. Si le processus est répété, cela génère une distribution empirique des estimateurs des paramètres (ou de fonctions de ces paramètres *i.e.*, les indicateurs de fiabilité), et permet alors de déterminer des intervalles de confiance.

5 Applications

Les calculs ont été effectués à l'aide du logiciel libre de distribution R (R Development Core Team, 2008) et du package contributif lme4 (Bates, Maechler, & Bolker, 2011).

5.1 Étude tennis

L'objectif de ce travail est d'étudier une méthode de normalisation du signal électromyographique (EMG) pour les muscles du membre supérieur chez des joueurs de tennis. Les joueurs ($N = 18$) ont réalisé les exercices à intensité maximale en vue d'atteindre un niveau maximal d'activation EMG (en µV) des muscles étudiés. Chaque sujet a effectué les exercices lors d'une première session (t), répétée le lendemain ($t + 1$) et une deuxième fois une semaine plus tard ($t + 7$).

Nous nous limitons ici au traitement des niveaux d'activation maximal du grand dorsal. En l'espèce, les trois mesures (t, $t + 1$, $t + 7$) d'un individu à l'autre sont bien comparables, il y a donc nécessité de vérifier s'il y a un *effet condition expérimentale* (la session) en plus de l'*effet sujet*. Les trois dates ne sont pas interchangeables, et la comparaison détaillée entre les conditions-sessions est enrichissante, par exemple entre t et $t + 1$

pour détecter un effet de fatigue ou d'apprentissage à prendre en compte ultérieurement dans l'amélioration du protocole. Dans ce cas l'effet condition est donc *fixe*, et c'est le modèle (3,1) qu'il convient d'employer. Les estimations[3] sont alors : $\mu = 433,53$; $\beta_1 = 0$; $\beta_2 = -48,14$; $\beta_3 = 40,12$; $\sigma_v^2 = 20094$ et $\sigma^2 = 12425$.

On obtient ainsi (1) $SEM = \sqrt{12435} = 111,47\mu V$ et (3) $CV = 111,47/433,53 = 0,26$. L'erreur de mesure, rapportée à la moyenne générale semble raisonnable. Le coefficient de corrélation intra-classe (4) $ICC = 20094/(20094+12425) = 0,62$ est lui aussi encourageant. Cette dernière remarque peut être modérée par l'intervalle de confiance associé qui est de $0,30 < ICC < 0,80$. Sa grande largeur est le fait du faible nombre de sujets ($N = 18$) qui laisse place à une large incertitude. On trouve en outre : $85,3 < SEM < 138,5$ et $0,18 < CV < 0,34$ qui évoluent du simple au double.

La nullité des paramètres de la méthode (β_1, β_2 et β_3), traduisant l'intérêt de sa prise en compte, peut être éprouvée par un test au maximum de vraisemblance : $\chi^2(2) = 5,54$, $p = 0,06$. La significativité statistique est proche (vu le nombre moyennement élevé de sujets, la puissance est faible), ce biais est probablement à prendre en considération pour améliorer le protocole. Toutefois, il faut souligner comme Bland et Altman (1983), que l'étude de la significativité statistique du biais par un simple test t apparié dans le cas de deux mesures ou plus généralement comme ici avec un test de type F des différences entre les moyennes des méthodes, ne vaut pas à elle seule étude de l'erreur de mesure. En effet un test t non significatif ne signifie pas qu'il n'y ait pas d'erreur de mesure. Au contraire, plus celle-ci est importante (au dénominateur de la statistique $t$ou de la statistique F) et moins le test a de chance d'être significatif ! Inversement, la significativité statistique d'un biais peut indirectement être l'indication que l'erreur de mesure, du moins sa partie aléatoire, est faible. Enfin, même en se cantonnant à la partie systématique de l'erreur, c'est-à-dire au biais, il convient d'avoir toujours à l'esprit que la significativité statistique est fonction de la taille d'échantillon. Un biais même faible sera bien souvent détecté avec un large échantillon ($N = 418$ dans l'exemple qui suit sur le rugby) alors qu'un autre bien plus important ne le sera pas avec un petit nombre de sujets (comme en l'espèce pour le tennis avec $N = 18$).

La visualisation est essentielle afin de juger de l'adéquation des données au modèle probabiliste employé pour les analyser. Ceci a été clairement démontré dans le contexte des erreurs de mesure, mais limité au cas de deux mesures seulement par Bland et Altman (1983, 1986) en employant le graphique qui porte à présent leurs noms et

[3] Il est possible de définir selon les logiciels des contraintes différentes pour les paramètres β_j. Une technique classique est d'annuler l'un d'entre eux... de préférence celui de la méthode de contrôle, si elle existe, qui est ici la première date. Le paramètre μ représente alors directement la moyenne de cette valeur basale.

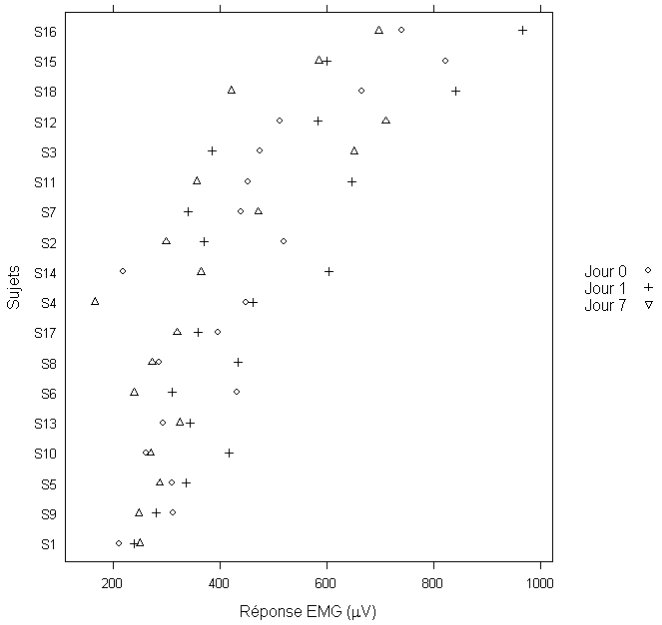

Fig. 1. Graphique en points pour les données du tennis (grand dorsal). En abscisses, la valeur de la réponse EMG, en ordonnées les sujets triés en fonction de leur réponse moyenne. Sur la même ligne sont superposées les réponses pour les trois jours considérés.

est une variation d'une proposition plus ancienne de John Tukey : le graphique des moyennes et différences.

En généralisant les graphiques conçus pour les données appariées, quatre approches principales sont envisageables que nous allons à présent illustrer sur les données de tennis.

La Figure 1 montre un graphique en points (*dotplot*, Cleveland, 1993) qui permet de visualiser clairement la large part de la variation inter-sujets par rapport à la variation intra-sujet. On peut plus difficilement détecter si une méthode présente une évolution d'une date à l'autre, il semble ici que la réponse du jour 7 (les triangles) soit souvent plus élevée que celle du jour 1 en particulier (les croix).

La Figure 2 représente le graphique en coordonnées parallèles (Wegman, 1990) qui étend le graphique dit avant-après ou graphique des profils (Fox, 2004). Contrairement à la procédure usuelle où les variables sont exprimées dans des unités différentes et doivent préalablement être standardisées d'une façon ou d'une autre, avec des mesures répétées, il faut prendre soin de conserver les unités originelles afin de pouvoir repérer les évolutions. Les trois mesures sont présentées à une même échelle sur autant de lignes et reliées par des traits. Ce graphique reste assez complexe à lire, surtout en noir et blanc, sauf en ce qui concerne le repérage de sujets extrêmes ou de regroupements de sujets.

Le graphique des corrélations par paires (Fig. 3) montre à la même échelle les relations deux à deux entre les mesures (et n'est donc utilisable que si elles sont

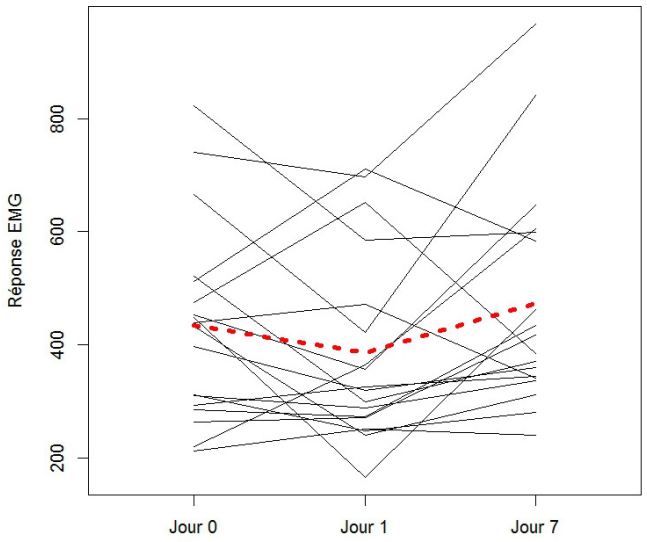

Fig. 2. Graphique en coordonnées parallèles pour le problème du tennis (grand dorsal). Les trois mesures EMG pour un même sujet sont reliées une ligne. La moyenne est indiquée en pointillés (rouges).

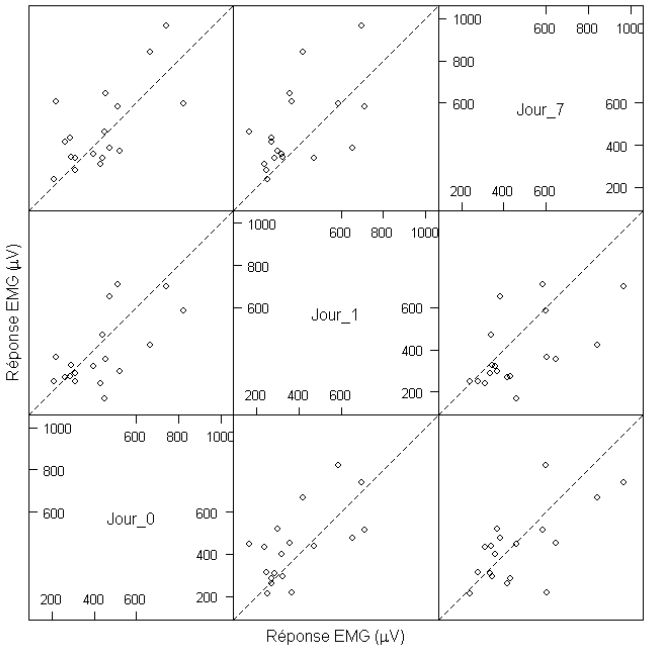

Fig. 3. Graphique des corrélations par paires entre les mesures du problème du tennis (grand dorsal). La première bissectrice est ajoutée car elle est la référence de l'égalité des deux mesures EMG.

comparables d'un sujet à l'autre). La première bissectrice donne un repère, elle indique l'égalité entre les deux mesures. Une déviation systématique, comme c'est le cas dans le croisement de la mesure du jour 1 et du jour 7 (quatre mesures sont plus élevées pour le jour 1 contre quatorze pour le jour 7), indique une évolution. Cependant, lorsque les mesures sont en bon accord, les points s'agglutinent sur la bissectrice et il est difficile d'en lire le détail.

Cet inconvénient n'existe pas dans le graphique des résidus (Fig. 4) qui étend celui de Tukey ou de Bland-Altman, et est un classique des diagnostics de modèles linéaires. L'idée est d'ajuster un modèle, ici simplement l'effet sujet dans une analyse de variance à un facteur fixe, et d'observer les résidus versus les prédictions (les moyennes par sujet donc). La figure 4 représente de façon séparée pour les trois conditions expérimentales (jours) ce graphe en ajoutant la ligne de référence qu'est la droite continue horizontale en zéro – moyenne générale des résidus – et une ligne en pointillés indiquant la moyenne des résidus pour la condition expérimentale étudiée. L'écart entre les deux lignes indique l'existence d'une évolution, comme c'est ici le cas pour le jour 1 et le jour 7.

Le graphique des résidus nous semble le plus efficace en général. Il est en effet utilisable avec un nombre de points assez grand ou assez petit. Nous le privilégierons par la suite.

5.2 Étude rugby

Au cours d'une étude cherchant à évaluer l'influence des émotions sur la performance sportive, quatre experts ont évalué la performance des participants en fonction de critères prédéfinis. Plus précisément, la qualité des actions ($N = 480$) au cours d'un match de rugby a été jugée à l'aide d'échelles visuelles analogiques de 10 centimètres, permettant ainsi d'obtenir, une évaluation pour chaque action des quatre experts allant de 0 à 10. Ces jugements étant comparables d'une action à l'autre puisque formulés par les mêmes personnes, nous pouvons donc essayer de voir s'il y a un biais entre les quatre mesures (*i.e.*, un phénomène reproductible d'une action à l'autre). À l'*effet sujet* (une action) s'ajoute donc un *effet condition expérimentale* (un juge). Cependant, comparer un juge à l'autre ne nous intéresse pas réellement, ils sont en quelque sorte interchangeables. Nous souhaitons davantage savoir si, dans leur ensemble, ces juges évaluent une même situation différemment[4] (pour en tenir compte dans l'estimation de l'erreur). Pour le dire autrement, nous voulons quantifier la disparité des notations dans une population de juges, les quatre juges en étant un simple échantillon, on dit en ce cas que l'effet condition est *aléatoire*.

[4] Un indicateur souvent employé pour mesurer la similarité de jugements est le coefficient Kappa. Toutefois, il est applicable lorsque les mesures sont de nature qualitative avec des variations possibles lorsqu'elles sont ordinales pour le cas très classique de jugements formulés sur la base d'échelles de Likert. Pour l'étude considérée sur le rugby, l'échelle analogique est continue de 0 à 10. Des méthodes réservées aux données numériques, telles que l'*ICC*, le *SEM*, sont donc plus adaptées.

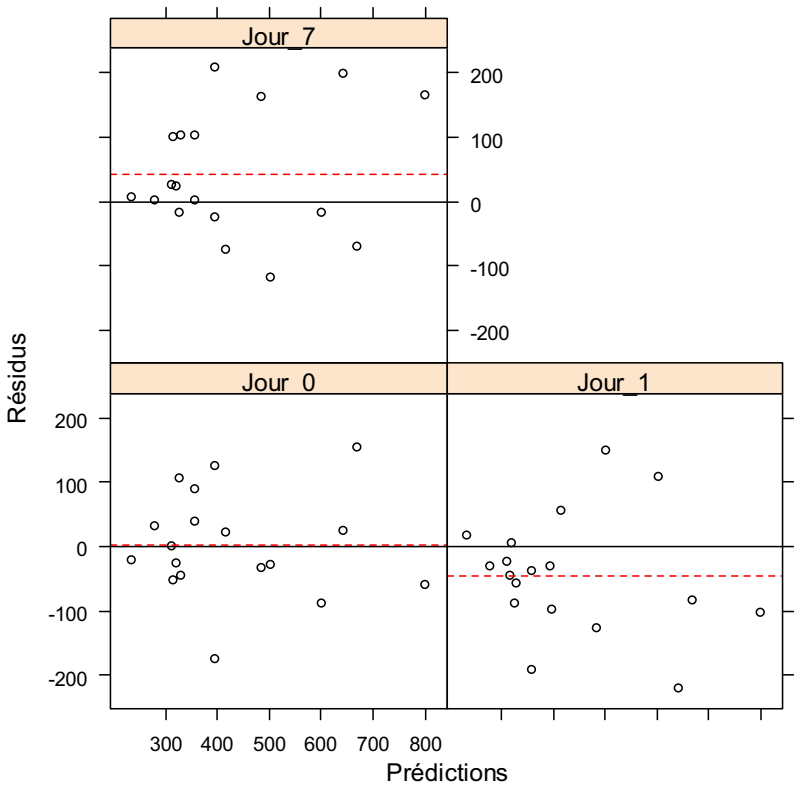

Fig. 4. Graphique pour le tennis (grand dorsal) des résidus en fonction de la valeur ajustée pour une ANOVA (EMG Sujet) autrement dit la moyenne du sujet correspondant. Représentation dans différents panneaux selon la condition expérimentale (jour). La ligne continue horizontale en zéro constitue la référence générale et pour chaque condition, la moyenne de ses résidus (indiquant son biais) est ajoutée comme une ligne en pointillés.

En employant par conséquent le modèle (2,1), les estimations REML sont : $\mu = 4{,}5481$, $\sigma_b^2 = 0{,}0072$, $\sigma_v^2 = 5{,}5583$ et $\sigma^2 = 0{,}6326$.

On obtient ainsi les indicateurs : $SEM = \sqrt{0{,}6326} = 0{,}7953$ cm (1) et $\sigma_D = \sqrt{2} \times 0{,}7953 = 1{,}1247$ cm (2). Sur une échelle de 10, les différences de score ne sont pas négligeables et réclament probablement d'utiliser la moyenne de plusieurs experts pour tempérer leur « subjectivité ». On a (3) $CV = 0{,}7953/4{,}5481 = 0{,}1749$ mais sur cet exemple, nous aurions aussi pu utiliser 5, le milieu naturel de l'échelle, comme dénominateur. L'erreur de mesure rapportée à la moyenne semble sur cet exemple plus modeste que dans l'étude précédente sur le tennis.

Quant à l'indicateur relatif (5) $ICC = 5{,}5583/(5{,}5583 + 0{,}0072 + 0{,}6326) = 0{,}8968$, ce qui est important et provient en partie du fait que les juges utilisent véritablement toute l'échelle de 0 à 10 pour estimer la qualité des actions. Vu le grand nombre d'actions observées ($N = 480$), les intervalles de confiance correspondants sont réduits : $0{,}88 < ICC < 0{,}91$; $0{,}77 < SEM < 0{,}82$; $0{,}16 < CV < 0{,}19$.

Un test au maximum de vraisemblance de l'intérêt de prendre en compte les paramètres de la méthode (le biais) donne $\chi^2\,(1) = 10{,}53$, $p = 0{,}002$. La significativité statistique est forte, grâce au nombre très important d'actions. Toutefois la valeur concrète de l'effet du biais appréhendé par $\sigma_b (= 0{,}085)$ montre sa moindre importance par rapport à $\sigma (= 0{,}795)$, le biais est environ dix fois moins grand que la variation aléatoire. Ceci est tout à fait clair sur la représentation graphique (Fig. 5). On peut également y apercevoir une chose plus fine concernant l'erreur : elle est plus grande au centre du graphique, lorsque le « vrai » score est vers 5, qu'aux extrémités 0 ou 10 ce qui semble normal puisqu'il y a alors « moins d'espace » pour exprimer des différences entre les juges.

5.3 Étude « squat jump »

L'objectif général de cette troisième étude est de déterminer une « meilleure » méthode d'optimisation mathématique pour évaluer les paramètres inertiels du tronc lors d'un mouvement dynamique. En ce qui concerne la mesure, les couples de forces à chaque articulation ainsi que le couple résiduel (en N.m) sont calculés. C'est ce dernier que nous étudierons ici. Ainsi, $N = 12$ sujets ont chacun réalisé dix *squat jumps* maximaux (saut vertical sans contremouvement avec une flexion initiale de genou de 90°). Des erreurs dans l'acquisition des données n'ont pas permis de garder tous les sauts. Ainsi le nombre de sauts par sujet est compris entre cinq et dix et on ne

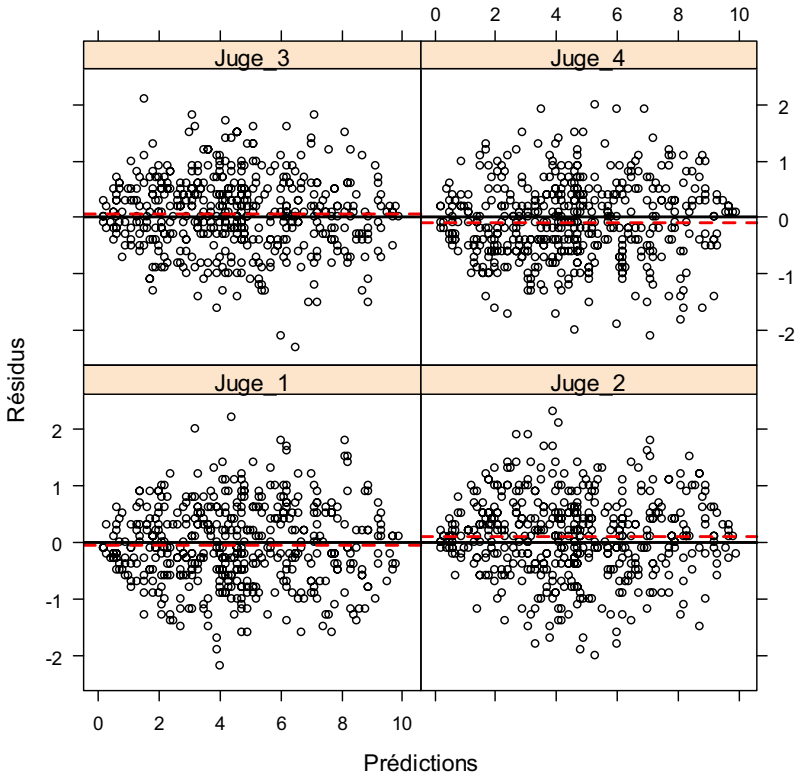

Fig. 5. Graphique des résidus pour le problème du rugby.

peut donc pas les faire correspondre exactement d'un sujet à l'autre, nous sommes ici sur le modèle plus simple du vrai score.

Les estimations REML des paramètres du modèle (1,1) sont : $\mu = 0{,}0042$, $\sigma_v^2 = 5{,}760 \times 10^{-6}$ et $\sigma^2 = 58{,}409 \times 10^{-6}$. On obtient ainsi (1) une erreur standard de mesure $SEM = \sqrt{58{,}409 \times 10^{-6}} = 0{,}0076$ N.m ou en ce qui concerne la différence entre deux mesures (2) $\sigma_D = \sqrt{2} \times 0{,}0076 = 0{,}0107$ N.m.

Sur cette seule méthode[5], il semble difficile de juger dans l'absolu de ces quantités, cette mesure n'étant pas, à notre connaissance, utilisée dans la littérature. Ainsi, les indicateurs relatifs seraient plus faciles à commenter. Toutefois la moyenne générale étant de 0,00418 on peut approximativement en déduire que pour un individu dont le score est « moyen », des valeurs négatives sont parfaitement envisageables (ce qui n'est pas possible avec ce type de mesure) ce qui tend à indiquer 1) une large variabilité et 2) un problème de symétrie des erreurs sur lequel nous reviendrons. La moyenne peut aussi être employée pour calculer le coefficient de variation $CV = 0{,}0076/0{,}0042 = 1{,}8261$ 3) qui est élevé. Ceci est confirmé par un coefficient de corrélation intra-classe de $ICC = 5{,}760 \times 10^{-6}/(5{,}760 \times 10^{-6} + 58{,}409 \times 10^{-6}) = 0{,}0898$ (4) qui peut certainement être considéré comme faible et indique en tout cas une bien plus grande homogénéité inter-sujets qu'intra-sujet.

Les intervalles de confiance correspondants sont : $0{,}0065 < SEM < 0{,}0088$; $1{,}19 < CV < 3{,}59$; $0{,}00 < ICC < 0{,}25$. On peut remarquer (plus clairement que dans les deux études précédentes) que les valeurs estimées sur le jeu de données ne sont pas forcément au centre de ces intervalles. Les distributions de ce type d'estimateurs sont en effet souvent non symétriques et plus proches de loi du χ^2 que d'une loi normale. On peut en outre noter la très forte variabilité du coefficient de variation essentiellement due au fait que la moyenne générale (au dénominateur du CV) soit proche de zéro.

Le graphique des résidus peut encore être employé dans le cas où il n'y a pas comparabilité des mesures, mais un seul panneau est alors utilisé. L'exemple des *squat jumps* soulève des problèmes extrêmement sérieux quant à l'adéquation du modèle employé sur les données. La Figure 6A est le graphique des résidus contre moyennes pour lequel on a sciemment choisi la même échelle en abscisses et ordonnées. La forme allongée montre bien que sur cet exemple, la variabilité intra-sujet est nettement plus grande que la variabilité inter-sujets. De plus, pour une moyenne donnée, on voit une dissymétrie des mesures, la plupart sont resserrées en deçà de zéro et quelques-unes sont bien éloignées en positif. Enfin, la dispersion est elle-même reliée à la moyenne (hétéroscédasticité) ce qui est encore plus net dans la Figure 6B dont l'objectif principal est de détecter de tels

[5] Le jeu de données originel comprend également un travail avec d'autres méthodes qui peuvent, elles, être directement comparées sur la base du *SEM*.

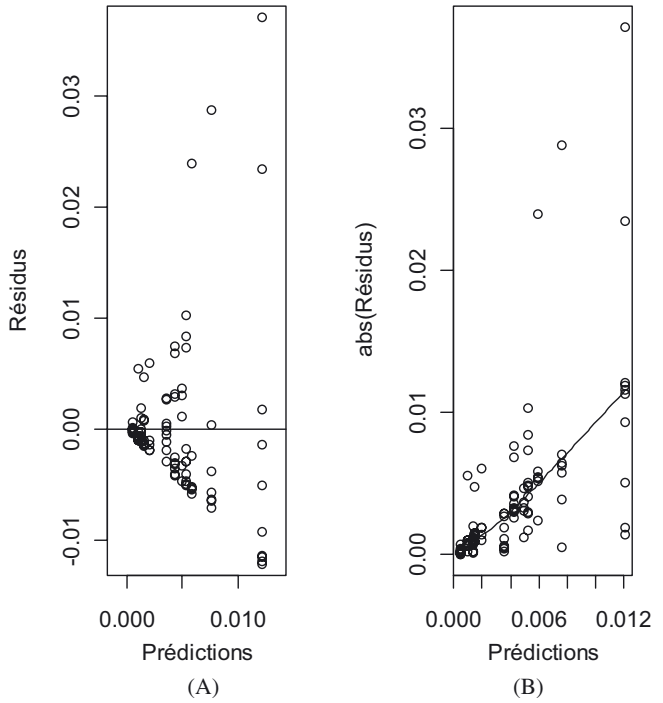

Fig. 6. (A) Graphique des résidus contre moyennes pour les données de squat jumps. La même échelle est employée pour les abscisses et les ordonnées. (B) La valeur absolue des résidus contre la moyenne, un lisseur est ajouté au graphique afin de représenter la relation entre la dispersion et la moyenne donc de détecter si les données sont hétéroscédastiques.

effets. Ce dernier graphique représente les valeurs absolues des résidus contre les moyennes et sert précisément à repérer l'hétéroscédasticité. Le lisseur ajouté au dessin indique pour chaque moyenne la dispersion correspondante des résidus, et il devient évident que les deux sont liées en l'espèce. Un modèle multiplicatif entre le vrai score et l'erreur semble plus adapté à la situation. Une transformation de type logarithmique de ces données semble donc une piste possible avant de pouvoir employer un modèle additif pour les analyser.

6 Conclusion

L'objectif de cet article était de proposer une démarche générale pour étudier l'erreur de mesure. Elle unifie les trois modèles classiques d'analyse de cette erreur de mesure, grâce au modèle linéaire à effets mixtes. Les indicateurs de l'erreur de mesure ont été clairement définis à partir des paramètres de ces modèles. La méthode d'estimation employée permet alors de résoudre le problème des plans d'expériences déséquilibrés, provenant le plus souvent de données manquantes. Des intervalles de confiance sont également calculables. Parmi les méthodes graphiques qui permettent de s'assurer de l'adéquation du modèle aux données, l'intérêt du graphique des résidus contre moyennes, et de ses variations, a été démontré.

On trouvera sur le site de l'éditeur en « matériel supplémentaire » les trois jeux de données utilisés dans cet article, et les commandes informatiques appliquées pour réaliser les calculs avec le logiciel libre de distribution R. Chacun pourra alors juger de l'intérêt de ces propositions et les utiliser. Il semble en effet important pour l'expérimentateur de s'inquiéter de la qualité de ses instruments de mesure et de tester leur fiabilité, particulièrement sur des populations sportives ayant rarement fait l'objet d'études spécifiques. Néanmoins, la philosophie des études d'erreurs de mesure s'avère différente de l'approche classique en statistique appliquée, souvent encore basée sur l'usage exclusif des tests de significativité. Nous avons en particulier vu que le test t apparié répond assez mal aux objectifs de ces études. Il s'agit plutôt 1) de quantifier l'erreur de mesure à l'aide d'indicateurs, si possible absolus mais une pluralité d'indicateurs appréhende certainement mieux le phénomène, et de les confronter à des objectifs analytiques concrets, 2) de fournir des intervalles de confiance de ces indicateurs pour juger du degré de certitude des résultats et 3) de veiller à séparer biais et variation aléatoire. En effet, en vue d'améliorer la qualité des protocoles, il convient d'une part de réduire le biais en contrôlant mieux les effets d'apprentissage, de fatigue, de réglage d'instruments, par exemple, et d'autre part, de diminuer la variation en utilisant à la place d'une unique et instable mesure la moyenne de plusieurs.

Au-delà des trois modèles présentés, le modèle linéaire à effets mixtes par la prise en compte de covariables inter et intra-sujet ouvre la possibilité d'analyser statistiquement des expérimentations plus complexes, par exemple celle de Eliasziw, Young, Woodbury & Fryday-Filed (1994) où plusieurs testeurs mesurent à plusieurs instants ce qui produit une structure intéressante sur les effets des conditions expérimentales. Une gestion plus fine des problèmes d'hétéroscédasticité semble également envisageable.

Enfin, si cet article donne des outils permettant de juger de la fiabilité d'une mesure, il convient de rappeler que ce n'est pas le seul critère de qualité à considérer et que les aspects complémentaires de validité[6], d'applicabilité[7] et de contextualité[8] participent tout autant au choix de mesures adaptées à la situation de recherche.

[6] Le fait que ce que l'on mesure en réalité soit effectivement ce que l'on souhaite mesurer.

[7] Prise en compte du temps, du coût et de la complexité de la mesure.

[8] Une même mesure peut s'avérer acceptable pour certains objectifs et pas pour d'autres.

Bibliographie

Altman, D.G., & Bland, J.M. (1983). Measurement in medicine: the analysis of method comparison studies. *The Statistician*, *32*, 307−317.

Atkinson, G., & Nevill, A.M. (1998). Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Medecine*, *26*, 217−238.

Bates, D., Maechler, M., & Bolker, B. (2011). lme4: linear mixed-effects models using S4 classes. R package version 0.999375-39, available on http://CRAN.R-project.org/package=lme4.

Bland, J.M., & Altman, D.G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, *1*, 307−310.

Bland, J.M., & Altman, D.G. (1996). Statistics notes : measurement error and correlation coefficient. *British Medical Journal*, *313*, 41−42.

Bollen, K.A. (1989). *Structural equations with latent variables*. New-York: Wiley.

Cleveland, W.S. (1993). *Visualizing data*. Summit, New-Jersey: Hobart Press.

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. New-York: Chapman & Hall.

Eliasziw, M., Young, S.L., Woodbury, M.G., & Fryday-Filed, K. (1994). Statistical methodology for the concurrent assessment of interrater and intrarater reliability: using goniometric measurements as an example. *Physical Therapy*, *74*, 777−788.

Fox, N.J. (2004). Speaking Stata: graphing agreement and disagreement. *The Stata Journal*, *4*, 329−349.

Hopkins, W.G. (2000). Measures of reliability in sports medicine and science. *Sports Medecine*, *30*, 1−15.

Lin, L., Hedayat, A.S., Sinha, B., & Yang, M. (2002). Statistical methods in assessing agreement: models, issues, and tools. *Journal of the American Statistical Association*, *97*, 257−270.

Morrow, J.R., & Jackson, A.W. (1993). How "significant" is your reliability? *Research Quarterly for Exercise and Sport*, *64*, 352−355.

Newell, J., Aitchinson, T., & Grant, S. (2010). *Statistics for sports and exercise science*. Harlow: Prentice Hall.

Pinheiro, J.C., & Bates, D.M. (2000). *Mixed-effects models in S and S-Plus*. New-York: Springer Verlag.

Quan, H., & Shih, W.J. (1996). Assessing reproductibility by the within-sublect coefficient of variation with random effects models. *Biometrics*, *52*, 1195−1203.

R Development Core Team (2008). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Available on http://www.R-project.org.

Shrout, P.E., & Fleiss, J.L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420−428.

Wegman, E.J. (1990). Hyperdimensional data analysis using parallel coordinates. *Journal of the American Statistical Association*, *85*, 664−675.

Weir, J.P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *Journal of Strength and Conditioning Research*, *19*, 231−240.