

## 運動能力 테스트의 信賴度 檢査

金 鍾 澤

身體活動이 탐구의 대상이 되는 체육학에서 새로운 이론의 정립이나 어떤 문제의 해결은 文獻考察이나 調査·實驗에 의한 資料分析을 통하여 이루어진다. 오늘날에는 체육학의 중요 영역인 운동생리학, 스포츠심리학, 그리고 스포츠사회학 등에서 조사나 실험연구가 주를 이루면서 연구자에게 전공지식 못지 않게 자료를 수집하고, 분석·해석하는 능력이 요구된다. 이는 主觀的인 의사결정보다는 실증적 자료의 統計的 分析에 의한 의사결정이 보다 합리적이고 과학적이라는 實證的 思考의 논리에서 이해된다.

체육학에서 어떤 개인이나 현상의 特性이나 能力의 검사는 실기검사와 필기검사로 구분되며, 일반적으로 運動能力(motor ability)의 측정·평가는 실기검사로 이루어진다. 정확한 자료를 구하기 위해서는 측정대상의 명확한 操作的 定義(operational definition)은 물론 표준화된 측정도구나 과정이 우선되어야 한다. 그리고 측정치의 정확성 혹은 검사도구의 가치는 일반적으로 테스트의 信賴度와 妥當度 정도로서 평가된다. 그러므로 체육측정·평가나 자료분석을 통한 연구에서 신뢰도와 타당도는 중요한 논제가 된다.

本稿에서는 운동능력 테스트의 신뢰도 검사의 의미와 신뢰도 추정시 고려되어야 하는 測定理論 그리고 신뢰도 推定의 統計的 方法에 대한 이론적 고찰을 하였다. 이는 단편적이거나 검사도구 혹은 측정치의 정확한 평가의 요령과 체육측정의 의미를 재고하기 위함이다.

### 信賴度の 意味

테스트의 信賴度(reliability)란 검사도구가 얼마나 정확하고 믿을만하게 측정하였나 하는 정도를 말한다. 그러므로 동일한 테스트로 동일한 대상을 측정하였을 때 측정치의 차이가 적을수록 테스트의 신뢰도는 높다. 이런 점에서 신뢰도란 테스트의 確實性(dependability), 一貫性(consistency), 安定性(stability) 등으로 설명된다(Anastasi, 1982; Thorndike, 1982). 또한 통계적인 관점에서 테스트의 신뢰도는 동일 피험자에 대한 반복 측정시 검사도구, 피험자 또는 검사실시 과정에서 오는 검사점수가 변동하는 원인이 무엇이며 분산적 오차의 정도를 나타낸다. 이러한 신뢰도의 구체적인 의미는 신뢰도 관련 측정이론과 신뢰도 추정 방법에서 설명된다.

정확한 자료의 수집을 위하여 높은 신뢰도와 타당도를 갖는 테스트의 선택이 필수적이다.

여기에서 신뢰도의 중요성을 찾을 수 있으며 보다 구체적인 신뢰도의 의미는 다음과 같다.

첫째, 신뢰도가 妥當度(validity)에 영향을 미친다. 높은 신뢰도를 갖는 테스트가 언제나 높은 타당도를 갖는다고 할 수는 없으나 어떤 테스트가 높은 타당도를 갖기 위해서는 높은 신뢰도를 가져야 한다(Baumgartner and Jackson, 1982). 그러므로 낮은 신뢰도는 테스트의 타당도, 즉 테스트의 適切性이나 判別性을 감소시키는 요인이다.

둘째 신뢰도가 相關係數에 영향을 미친다. 두 변인간의 상관의 정도를 나타내는 상관계수의 해석사 각 변인의 신뢰도가 고려되어야 한다. 이는 어떤 두 변인이 높은 상관을 갖는다 해도 각 변인의 측정이 신뢰롭지 못했을 때 두 변인간의 상관에 의문을 갖게 되는 것이며, 나아가서는 한 변인의 점수로 다른 변인의 점수를 예측했을 때의 예측력도 감소되기 때문이다. 보다 정확한 상관의 해석은 각 변인의 신뢰도를 고려한 상관계수의 교정을 위한 公式(correction for attenuation)에 의해서 가능하다(Ghiselli, Campbell and Zedeck, 1981).

마지막으로, 신뢰도는 통계적 가설검증에서의 檢證力(power)에 영향을 미친다. 통계적 가설검증에 의한 연구에서 통계적 유의수준이나 피험자수 등의 조건이 같다면 신뢰도가 낮은 테스트에 의한 자료가 사용되었을 때 연구결과의 타당성을 나타내는 검증력은 감소된다(Keppel, 1982; Rothstein, 1985).

그러므로 검사도구나 검사점수(측정치)의 정확성이나 안정성을 나타내는 신뢰도는 체육 측정, 평가, 혹은 연구에서 그 중요성이 인식된다.

### 通賴度 推定을 위한 測定理論

信賴度の 구체적인 의미와 추정방법을 알기 위해서는 다음의 두가지 측정이론의 이해가 선행되어야 한다.

첫번째 이론적 모델은 Theory of true and error scores(classical theory)이다. 이 이론은 1900년 초기 도입된 이론으로 테스트의 점수가 眞點數(true score)와 誤差點數(error score)의 합으로 정의되며, 이론상으로 가능한 진점수는 무수한 반복측정의 평균값이다. 이 이론에 근거한 신뢰도는, 측정대상인 한 개인의 특성 또는 능력은 동일테스트로 측정할 경우 어떤 경우에도 일정하며, 각 측정 시행간 오차점수는 서로 獨立的(independent)이고 無選的(random)이라는 통계적 가정을 만족시킬 때 전체점수 分散에 대한 眞점수 分散의 比로서 정의한다(Allen and Yen, 1979; Ghiselli, Campbell and Zedeck, 1981).

두번째 이론은 Domain sampling theory이다. 이 이론은 1950년 중반부터 사용되어 왔으며, 어떤 特性(trait)이란 공통의 성질을 갖는 수많은 속성들의 조합으로 이루어진 어떤 Domain 혹은 Universe로 정의되며, 어떤 Domain을 측정할 때는 하나의 테스트만이 아닌 여러 테스트의 조합으로 가능하다고 믿는다. 그러므로 이 논리에서의 진점수는 이론적으로

가정되는 어떤 Domain의 모든 요인들의 점수의 합이며 이 진점수 계산은 Domain을 구성하는 모든 요인들에서 무작위로 표집된 요인들의 값을 수없이 반복해서 구했을 때의 평균값으로 간주한다. 어떤 Domain에 대한 구체적인 설명과 이에 따른 구성요인들의 정의가 불투명하기 쉬운 단점이 있으나 Nunnally(1979)가 설명하는 Factorial domain sampling과 Binominal모델을 참조하면 이 이론에 근거한 신뢰도 추정의 논리가 분명해진다.

추상적인 개념의 측정이 추가되는 심리측정시 자주 사용되는 再檢法(test-retest), 半分法(split-half), 그리고 同型法(parallel test)에 의한 신뢰도 추정시는 Classical theory에 근거를 두나 K-R20(Kuder-Richardson 20), K-R21 혹은 Cronbach's coefficient alpha 등에 의한 問項內的合致度(inter-item consistency)은 Domain sampling theory가 적합하다(Ghiselli, Campbell and Zedeck, 1981; Thorndike, 1982). 그리고 추상적으로 규명되는 특성이나 능력의 측정인 필기테스트가 아닌 실기테스트로서의 운동능력 테스트는 Classical theory에 근거한 신뢰도 추정이 타당하나 絶對評價와 관련된 테스트의 신뢰도 추정시는 Domain sampling theory에 근거한 신뢰도 추정이 보다 합리적이다(Safrit, 1981).

### 운동능력 테스트의 신뢰도推定

本稿에서 운동능력 테스트의 신뢰도 추정 방법의 고찰은 추상적인 개념의 측정을 위한 필기테스트를 제외하고 실기테스트에 국한 하였으며 측정단위도 최소한 等間尺度(interval scale)인 連續資料(continuous data)로 제한하였다. 운동능력 테스트의 신뢰도 추정시의 이론적 근거는 앞에서 설명 하였듯이 Classical theory이다. 그러므로 신뢰계수는 진점수에 의해 설명되어 지는 분산을 전체분산으로 나눈 값이다(Safrit, 1981; Baumgartner and Jackson, 1982).

운동능력 테스트의 신뢰도는 일반적으로 재검법과 반분법에 의한 두 테스트 시행간의 신뢰도계수(interclass correlation)와 분산분석에 의해 가능한 3번 이상의 시행내의 신뢰도계수(intraclass correlation)로 추정된다. 일반적으로 인성이나 지능지수와 같은 추상적인 변인의 측정시 동형검사의 상관에서 Classical theory의 설명이 가능하나 운동능력 테스트에서는 동일테스트에 의한 재검법에서의 Pearson의 積率相關(pearson r)으로 이해된다.

3번 이상의 반복측정시의 운동능력 테스트의 신뢰도 추정도 재검법에서의 신뢰도 추정과 동일한 개념으로 반분법에 의해 가능하다. 즉, 반복측정의 각 시행을 짝수와 홀수번째 시행으로 구분하여 모든 짝수번째 합이나 평균과 홀수번째 합이나 평균의 Pearson r을 구하여 여기에 Spearman-Brown Phrophecy공식을 이용하여 신뢰도계수를 재조정하는 방법이다. Baumgartner(1968)는 제자리 멀리뛰기 등 4종목의 운동능력 테스트의 신뢰도 추정시 Spearman-Brown Phrophecy공식과 반분법을 이용해서 구한 신뢰계수가 재검법에 의한 신

회계수보다 낮게 되는 것을 발견하고 이는 반복측정시 피로 등으로 인한 테스트 시행간에 점수차가 고려되기 때문이라고 하였다.

이상의 재검법이나 반분법에 의한 운동능력 테스트의 신뢰도 추정시 사용되는 Pearson r은 사용과 해석이 용이한 장점이 있으나 다음과 같은 단점을 내포한다. 첫째로, Pearson r은 서로 다른 두 변인간의 상관관계를 알아보는 統計的方法(bivariate statistic)이며 반복측정한 변인을 다루는 통계(univariate statistic)이므로 논리적으로 적합하지 않다. 둘째로, 3번 이상 반복측정시 Pearson r을 이용한 직접적인 테스트 신뢰도의 추정이 어렵고 가능한 모든 시행간 상관의 평균으로 계산하여야 하는 번거로움이 있다. 셋째로, 반복측정 시행간의 평균이나 분산의 변화를 알 수 없다. 마지막으로 測定誤差에 대한 설명이 불가능하다(safrit, 1981; Baumgartner and Jackson, 1982).

3번 이상의 반복측정시의 신뢰도는 반분법에 의한 간접적인 측정이 가능하나 위에서 열거한 것처럼 Pearson r의 사용시 문제점이 고려됨으로 이를 해결할 수 있고 직접적으로 시행내의 상관관계를 구할 수 있는 分散分析이 이용된다. 분산분석에 의한 신뢰도 추정은 체육학에서도 Brozek(1947)와 Feldt and McKee(1958) 등이 시도한 이래 자주 사용되지는 않았으나 근래에 주목을 받고 있는 것으로 사료된다. 이는 심리측정에서도 반분법이나 동형법에 의한 신뢰도 추정보다는 K-R20이나 chronbach's alpha 등에 의한 문항내적 신뢰도를 구하려는 경향에서도 찾아진다.

분산분석에 의한 신뢰도 추정은 동일테스트의 반복측정에 의한 것이므로 피험자 내의 설계(within subject model)가 추가된다. 또한 변인의 수와 특성에 따라 일원 혹은 이원분산분석 그리고 固定(fixed), 無選(random), 混合(mixed)모형 등으로 구분되며 같은 분산분석의 모형에서도 분산의 해석이 다를수도 있기 때문에 경우에 따라서 서로 다른 신뢰도 추정방법이 적용되기도 한다. 일례로 이원분산분석에서 신뢰도 추정시 시행간의 통계적인 유의한 차의 유무에 따라 4가지 서로 다른 신뢰도 추정방법이 있다(Liba, 1962; Stamm, 1976; Baumgartner and Jackson, 1982).

體育學에서 특수한 경우의 분산분석에 의한 신뢰도 추정은 Kroll(1967) 그리고 Winer(1971) 등의 연구에서 찾아지며 한 테스트의 신뢰도 추정이 아닌 여러 종류의 운동능력 테스트의 신뢰도 추정도 Wood와 Safrit(1984)가 시도한 것처럼 正準分析(canonical analysis)에 의해서 가능하다.

이상으로 體育學에서 運動能力 테스트의 信賴도가 갖는 의미와 신뢰도 추정시 이해되어야 하는 측정이론 그리고 신뢰도 추정방법에 대하여 살펴보았다. 이는 체육현장에서 올바른 테스트 선택과 고안은 물론 평가와 연구시 올바른 자료수집과 분석을 위한 체육측정의 의미를 간접적으로나마 敷衍하고자 함이다.

## 參 考 文 獻

1. Allen, M.J, 1979, "Introduction to Measurement Theory". CA.: Woodsworth Inc.
2. Anastasi, A, 1982, "Psychological Testing". New York: Macmillan Publishing Co, Inc.
3. Baumgartner, T.A, 1968, "The Application of the Spearman-Borwn Prophecy Formula when Applied to Physical Performance Tests" Research Quarterly, 39:847-856.
4. Baumgartner, T.A, & Jackson, A.S, 1982, "Measurement for Evaluation in Physical Education." IA; Wm. C. Brown Co.
5. Ghiselli, E.E, Campbell, J.P. & Zedeck, S, 1981, "Measurement Theory for the Behavioral Science." CA: W.H. Freeman and Co.
6. Feldt, L.S, & McKee, M.E, 1958, "Education of the Reliability of skill Tests", Research Quarterly, 29:279-93.
7. Keppel, G, 1982, "Design and Analysis", NJ: Prentice-Hall, Inc.
8. Kroll, W, 1967, "A Note on the Coefficient of Intraclass Correlation as an Estimate of Reliability", Research Quarterly, 38:421-419.
9. Liba, M, 1962, "A Trend Test as a Preliminary to Reliability Estimation," Research Quarterly, 33:245-248.
10. Nunnally, J.C, 1978, "Psychometric Theory," New York: McGraw-Hill Book Co.
11. Rothstein, L.A, 1985, "Research Design and Statistics for Physical Education," N.J: Prentice-Hall, Inc.
12. Safrit, M.J, 1981, "Evaluation in Physical Education", NJ: Prentice-Hall, Inc.
13. Stamm, C.L, 1976, "An Alternative Method for Estimating Reliability," Journal of Physical Education and Recreation, 47:66-67.
14. Thorndike, R.L, 1982, "Applied Psychometric," MA: Houghton Mifflin Co.
15. Winer, B.T, 1971, "Statistical Principles in Experimental Design", New York: McGraw Hill Book Co.
16. Wood, J.M, & Safrit M.J, 1984, "A Model for Estimating the Reliability of Psychomotor Test Batteries", Research Quarterly for Exercise and Sport, Vol. 55. No. 1, pp. 53-63.

## The reliability of a motor ability test

Kim, Jong-Taek

The purpose of this paper was to scrutinize the meaning of the reliability of a motor ability test, especially a norm-referenced test, and the mathematical theories and statistical methods for estimating reliability.

Reliability is a important criterion of a measurement, a number obtained by a test, as known a reliable test measures whatever it measures consistently. Thus, reliability is defined as the dependability of a measure or the consistency of an individual in performing a test.

The theory of true and error scores(classical theory) and the domain sampling theory (generalizability) are the theoretical models to understand reliability. However the classical theory that assumes any measurement on a continuous scale contains an inherent component of test. a true and error score is used to assess the reliability of a motor ability In this concept, the reliability is the ratio of the true score variance to the obtained score variance.

There are many methods for estimating reliability based on the characteristics of a test such as stability reliability coefficient, pearson product momentum correlation, from the test-retest method and internal-consistency reliability, usually for a written test, from the split-half method. Intraclass correlation coefficient is the better method then interclass correlation coefficient when there is more than 2 scores per person. Also Spearman-Brown phrophecy formula is used when the legth of a test, the number of trials in a motor ability test, is increased. Many studies for estimating relialiity of a motor ability test by using various ANOVA models and the other statistical techniques like canonical analysis have been conducted in physical echecation and sports science.