

태권도 품새 경기의 주관적 평가결과의 오차원 분석: 일반화가능도 이론 적용

조은형¹

¹성균관대학교 체력과학연구소

접수 2016년 1월 4일, 수정 2016년 1월 21일, 게재확정 2016년 3월 7일

요약

본 연구는 G-Theory를 적용하여 태권도 품새 평가항목에 대한 채점자 간 평가점수의 신뢰도를 추정하기 위해 G-연구로 경기일 수, 채점자 수를 다중오차원으로 선정하고 이들 변인간의 상호작용으로 인한 오차변량의 상대적 크기에 의해 오차원을 분석하고, G-연구의 결과를 토대로 D-연구를 수행하여 최적의 측정조건을 결정하는 데 목적이 있다. 결과를 요약하면 다음과 같다. G-Theory를 적용하여 태권도 품새 평가항목 중 정확성에 대해 분산성분 추정치를 추정한 결과, 오차의 영향력은 채점자국면에서 가장 크게 나타났으며, 각 설계에 따른 상호효과 (피험자 내), 피험자 간 순서였으며, 표현성 평가항목에 대한 분산성분 추정치 오차의 영향력은 상호효과 (피험자 내)국면에서 가장 크게 나타났으며, 각 설계에 따른 피험자 간, 채점자 국면 순으로 나타났다. 마지막으로 D-연구를 통하여 일반화가능도계수를 추정한 결과, 채점자 수에 따른 최적수준의 측정조건은 정확성 평가항목에서 8명의 채점자일 때 안정적인 신뢰도를 얻을 수 있으며, 표현성 평가항목에서는 7명의 채점자일 때 안정적인 신뢰도를 얻었다.

주요용어: 다중오차원, 분산성분, 신뢰도, 일반화가능도 이론.

1. 머리말

자료의 신뢰도를 추정하기 위해 연구자들은 여러 가지 방법을 통해 신뢰도 계수를 산출한다. 신뢰도를 나타내는 대표적인 방법은 α 계수, 급내 상관 (intraclass correlation), 그리고 일반화가능도 이론 (generalizability theory)이다. 세 방법 모두 고전검사이론 (classical test theory)에 기초한 추정 방법으로서 공분산의 모형 (α 계수), 변량분석 (ANOVA)의 모형 (급내 상관, 일반화가능도 이론)을 적용하므로 많은 공통점이 있지만, 모두가 서로 다른 가정들에 근거하여 각각의 신뢰도 값을 산출한다 (Allen과 Yen, 1979). 이러한 여러 가지 방법에 의해 신뢰도 계수의 산출방법의 발전은 측정이론 (measurement theory)에 의해 이루어져 왔다 (Park과 Pi, 2013; Park, 2014). 측정 (검사)이론의 발전과정은 크게 두 가지 세대로 나누어 설명할 수 있다 (Kim, 1989). 첫 번째는 관찰점수는 진점수 (true score)와 오차점수 (error score)의 합으로 이루어진다는 Spearman의 고전검사이론과 두 번째로 고전검사이론의 문제점을 해결하고자 제안된 이론이 무선표집모형 (random sampling model)에 기초한 일반화가능도 이론과 척도모형인 잠재적 특성모형 (latent trait model)에 기초한 문항반응이론 (item response test)이다.

¹ (440-746) 경기도 수원시 장안구 천천동 300, 성균관대학교 체력과학연구소, 선임연구원.
E-mail: taedori1974@hanmail.net

고전검사이론에서 신뢰도 (reliability)는 관찰 점수분산 (observed variance)에 대한 진점수 분산 (true variance)의 비율의 수리적 개념으로 설명할 수 있으며, 오차점수의 분산을 하나의 요인으로 취급하고 있다. 여기서 말하는 오차는 단지 시간의 흐름에 의한 구조적 관점에 따른 오차를 추정하고 있는 것이 특징 (Hagtvedt, 1997)으로 실제적으로 측정의 목적, 측정대상의 전집(universe score), 그리고 측정과정에 따라 진점수 (true score)가 변화된다는 것을 지나치고 있다. 이러한 것은 결국 오차의 원인과 비중의 규명을 가로막아 체계적 신뢰도, 측정절차 및 정확도의 발전을 저해하는 요인이 된다. 이러한 고전검사이론의 단점을 보완하기 위해 Cronbach 등 (1972)은 측정이론의 2세대에 해당하는 일반화가능도 이론을 정립하였으며, 특히 행동관찰, 임상평가, 수행평가와 같이 도구를 사용한 측정의 신뢰도를 추정하기 위한 측정이론의 한 분야로서 1972년 Cronbach 등에 의해 이론적 모형이 형성되었고, Brennan (1983), Shavelson 등 (1981)에 의해 이론적 체계가 갖추어지면서 그 적용 범위가 확대되었다.

이러한 일반화가능도 이론은 측정상황에서 발생할 수 있는 다양한 오차 분산을 분화되지 않는 하나의 단일 분산으로 규정한 고전검사이론의 한계를 극복하여, 모형의 설계에 따라 한 번의 측정으로 다중 오차요인의 크기를 규명하고, 오차 분할에 따른 측정 조건의 조정이 가능하다는 장점을 지니고 있다. 이는 관찰된 행위 하나하나가 행동 전집 (universe)을 대표하는 일부분이라고 전제하고, 전집을 구성하는 국면 (facet) 또는 조건들을 바탕으로 표집된 행위가 얼마나 일반화 되어질 수 있는가를 추정하는 것이다 (Kim, 1989). 이러한 일반화가능도 이론을 이해하기 위한 주요 개념들은 일반화전집, 전집점수, 일반화연구, 결정연구, 국면, 분산성분, 상대오차분산, 절대오차분산, 일반화가능도 계수, 의존도 계수 등의 개념으로 정리될 수 있다 (Kim과 Kim, 2001). 일반화가능도 이론은 크게 측정점수에 대한 다중 오차요인을 규명하고 그 상대적 크기를 추정하는 일반화가능도 연구 (G-연구)와, G-연구에서 얻어진 다중 오차요인에 대한 정보를 토대로 각각의 오차분산을 최소화하는 최적의 측정 조건을 찾아내는 결정연구 (D-연구)로 나누어진다 (MacMillan, 2000).

G-연구와 D-연구 두 가지를 구분하여 진행하는 이유는 연구자의 목적에 따라 분산성분 크기의 절대적 비교연구도 가능하고 같은 G-연구 결과로 상이한 D-연구 설계로 연결될 수 있기 때문이다. D-연구 설계는 G-연구 설계와 동일한 모형으로 일반화전집을 추정한다. 기호 상의 변화는 자료 분석의 기준이 더 이상 개인의 관찰점수가 아니라 일반화 전집을 포함하는 일련의 국면에 대한 관찰 평균점수이기 때문에 대문자로 표현한다. 위에서 나타난 G-연구 결과는 D-연구 설계로 연결됨으로써 결정과정에 참여한다. 연구자는 각 분산성분의 상대적 크기에 따라 국면의 차원을 조정한다. 물론, 상대적 결정인가 절대적 결정인가에 따라 오차분산에 포함되는 분산성분과 그에 따른 일반화 가능도 계수 산출은 차이가 있다. 상대적 결정은 관찰대상 간의 차이 (변동) 파악에 초점을 맞추는 반면, 절대적 결정은 관찰자 대상 행위자체의 수준을 주시한다. 따라서 절대적 결정에 포함되는 오차분산성분은 전집분산을 제외한 각 국면의 차이와 상호작용에서 나타나는 불일치를 모두 포함하지만, 상대적 결정에서 오차분산에 포함되는 성분은 관찰대상과 상호작용하는 분산성분들만을 포함한다.

G-연구는 허용가능한 관찰전집을 추정하기 위하여 측정오차에 영향을 주는 다중오차요인을 동시에 분석하고 분산분석절차를 이용하여 설계모형에 따라 각 효과의 분산성분을 추정하는 과정이다. G-연구는 변량분석 결과에서 얻어진 평균제곱 (mean square)으로 관찰 점수 분산을 구성하는 독립적인 분산성분 (variance component)을 추정하는 과정으로 채점결과에 나타나 있는 각각의 오차원 (error source)을 규명함과 동시에 각각의 오차원의 상대적 크기를 추정하는 과정으로 정리된다. 예를 들어, 상이한 측정조건에서 검사 또는 평가를 받은 피험자들의 검사결과를 변량분석방법을 적용하여 분산성분과 상호작용성분을 추정한 후, 일반화가능도 계수 (G-계수)를 산출하는 것이다. G-계수는 전집점수 변량에 대한 관찰점수 변량의 비율로 산출된다. 검사는 하나의 G-계수를 가지는 것이 아니라 G-연구 설계에 포함되는 국면에 따라 여러 개의 계수를 갖는다. 결국 G-연구는 피험자의 전집점수를 가장 정확하게 추정하는 방법에 대한 정보를 제공하게 된다. 즉 G-연구에서는 측정에 영향을 주는 오차요인의 분산성

분과 그 크기를 추정하는 것이 목적이다.

D-연구는 연구자가 측정결과를 일반화하고자하는 일반화 전집을 규정하고, 측정의 변동에 영향을 주는 관측조건을 조정하거나 결정하는 과정이다. 이는 G-연구에서 산출된 오차요인의 분산성분의 절대적 (absolute decision) 또는 상대적 (relative decision) 크기에 따라 측정의 조건을 조절함으로써 오차분산을 최소화하는 일반화 계수 혹은 의존도 계수 (dependability)를 동시에 추정할 수 있게 해준다. D-연구에서는 G-연구 결과 산출된 각 점수의 분산성분 추정치를 토대로 측정대상과 일반화 전집의 정의에 따라 전집 점수 분산, 오차 점수 분산 및 일반화가능도 계수의 정보를 제공해 준다. 따라서 D-연구에서는 G-연구에서 얻어진 결과를 토대로 각각의 오차변량을 최소화하여 가장 최적화 된 검사조건을 찾아내는 과정으로 정리할 수 있다. 다시 말해 일반화가능도 이론은 적정수준의 검사횟수, 검사자수, 검사 실시 일수 등 최적의 검사조건을 규명해주는 이론이라고 할 수 있다. 즉, 일반화가능도 이론은 고전검사이론에 변량분석 (ANOVA)의 체계를 적용하여 측정상황에서 발생할 수 있는 다중오차요인 (multiple sources of error variance)을 동시에 분석하고, 측정점수에 대한 오차요인의 상대적 영향력을 산출하여 일반화가능도 계수와 함께 안정적인 점수를 얻기 위한 측정조건을 제시함으로써 신뢰도 추정과정을 한 단계 향상시킨 측정이론이다 (Kim 과 Kim, 2001). 이는 진점수가 하나라는 고전검사이론의 한계에서 벗어나 전집 (universe)의 정의에 따라 측정상황에서 나타날 수 있는 다양한 오차요인과 그 영향력을 설명하여 최적의 측정조건을 규명해준다고 할 수 있다.

결국, 일반화가능도 이론은 사용되는 것에 따라 분석, 관찰 상태, 측정과제의 특이성을 강조하는데 도움이 된다고 제시하고 있으며 (Hagtvet, 1997), 신뢰도이론의 확장으로 각각의 주요 국면과 관련된 변량만 뿐만 아니라 설계 내에서의 상호효과적인 변량의 관계를 결정하는 것도 허용한다고 하였다. 따라서 일반화가능도 이론의 연구로부터 얻게 되는 충분한 정보는 연구자에게 현장 연구의 신뢰성 예상치를 제공할 뿐만 아니라 미래의 연구를 위해 신뢰할 수 있고 효율적인 측정 절차를 설계하는 것에 도움이 될 수 있다 (Sun 등, 1997). 그동안 일반화가능도 이론은 교육·심리측정분야를 중심으로 활발하게 연구되어왔으며, 미국 체육학 분야에서도 Safrit 등 (1976)에 의해 소개된 이후 국내에서는 운동수행검사의 신뢰도 평가, 체육교수 관찰도구 개발, 피하지방검사와 농구기능검사 신뢰도 추정 (Cho, 1991), 초등학교 기계체조 운동의 실기평가 (Ko, 1997), 일반화가능도 이론을 적용한 주관적 배구기능검사의 신뢰도 추정 (Kim과 Heo, 2002), 국가대표볼링선수 선발방법의 신뢰도추정: 일반화가능도 이론 적용 (Lee, 2005), 볼링국가대표선수 선발전의 오차요인 탐색과 신뢰도 평가 (Cho, 2006), 일반화가능도 이론을 적용한 태권도 심판관정의 신뢰도 분석 (Jung, 2009) 등에 적용되어 왔다. 이렇듯 스포츠과학 및 체육학 연구에서 다양한 측정 및 검사에 따른 다차원적 오차원 (specify multiple error source)들을 가지고 있기 때문에 일반화가능도 이론을 적용한 연구결과의 중요성은 더욱 크다고 볼 수 있다 (Cho, 1991).

한편 태권도는 우리나라의 국기이며 전 세계 사람들이 공유하고 있는 한국의 'Culture Sport Brand'로 발전한 종목이다. 태권도 인구는 206개국 회원국에서 약 8천만 명 이상으로 추산하고 있으며 (WTF, 2014), 약 900만 명의 유단자가 수련 (Kukkiwon, 2015)하는 종목이다. 이러한 태권도의 기술체계는 겨루기, 품새, 격파, 시범으로 구성되어 있으며, 겨루기 종목은 세계태권도선수권대회에서 14년간 약 75개의 메달을 획득하여 세계최고의 경기력을 보유한 국가로 대한민국의 태권도가 국제대회에서 최고의 경기력을 수행할 수 있었던 것은 많은 수련인구와 두터운 선수층, 그리고 선진화된 기술체계를 바탕으로 성장해 왔다 (Choi, 2013). 그러나 태권도의 품새는 겨루기와 달리 태권도 기술이 단계별로 분류된 일련의 연결동작을 정확하고 능숙하게 표현하는 것에 평가의 기준을 두고 있다 (Cho, 2014). 태권도를 관련 주제로 수행된 국내 연구물은 약 450편 (KERIS, 2015)에 이르고 있으며, 주로 체력, 심리, 경기분석 등의 다양한 분야에서 태권도를 과학적으로 한 단계 도약시킬 수 있는 기초자료를 제공하는데 많은 노력을 해왔다 (Choi, 2013). 따라서 이 연구는 태권도 품새 경기현장에서 선수들에 대한 채점자의 평가 원자료 일부를 무선표집하여 일반화가능도 이론의 접근을 통해 태권도 품새 평가

항목에 대한 채점자간 평가점수의 신뢰도를 추정하는데 목적이다. 구체적으로 심사자간 평가점수에 영향을 미칠 것으로 판단되는 경기일 수, 심사자 수를 다중오차원으로 선정하고 이들 변인간의 상호작용으로 인한 오차변량의 상대적 크기에 의해 오차원의 영향력을 비교하고, 이를 기반으로 일반화를 위한 전집과 국면의 수에 따라 G계수를 산출하고 G-연구의 결과를 토대로 D-연구를 수행하여 최적의 측정조건을 결정하는데 목적이 있다.

2. 연구방법

2.1. 연구자료수집

본 연구의 자료수집은 피험자인 태권도 품새 선수와 채점자인 태권도 품새 심판으로 구분하여 2015년 4월 11일~12일 전라남도 화순군 하니움 센터에서 개최된 Y대학교 총장기 태권도 품새 대회에 참가한 고등부 남자 ($n=12$), 여자 ($n=8$) 총 ($n=20$)을 원자료 (raw data)로 수집하였으며, 평균나이는 17.9세이다. 선수 20명은 심판 ($n=5$)에 의해 평가된 자료이다. 원자료의 수집은 대한태권도협회 (KTA) 품새 분과 위원 및 기록 분과의 협조 하에 이루어졌으며, 심판들은 현재 대한태권도협회 품새 심판으로 활동 중인 자로서 심사자 모두 3급 이상의 심판자격을 소유하고 있으며 각각 5년 이상 심판경력이 있는 자들로 구성되었다. 대회에 참가한 선수 및 심판들의 일반적 특성은 Table 2.1, Table 2.2와 같다 (부록: 원자료 첨부).

Table 2.1 General characteristics of participants

Total	male	female	age, M (SD)
20	12 (60%)	8 (40%)	17.9 (.759)

Table 2.2 General characteristics of raters

Measure	Rater A	Rater B	Rater C	Rater D	Rater E	M (SD)
Age	59	54	50	55	51	53.8 (.3.56)
Teaching career	30	22	10	25	20	21.4 (7.40)
Judge career	30	8	5	7	6	11.2 (10.5)
Region	Gyeonggi-do	Seoul	Seoul	Gyeonggi-do	Seoul	-

2.2. 연구 설계 및 측정모형

본 연구는 태권도 품새 선수를 대상으로 평가항목의 신뢰도를 추정하기 위해 일반화가능도이론을 적용하여 분석하였다. 품새의 평가항목은 정확성/숙련성 (표현성)으로 구성되어 있으며, 정확성은 국기원이 규정한 태권도 기본동작과 각 품새 동작의 수행기준에 맞는 동작 시행의 여부를 평가하는 것으로 기본동작이나 각 품새별 세부동작 및 균형에 실수가 있을 시 0.1에서 0.3 감점을 부여하여 3점 만점으로 채점한다. 또한 숙련성 (표현성)은 각 품새가 갖는 고유의 의미와 기법의 특징과 동작의 연결 특성을 표현하는 능력을 평가하는 것으로 7점 만점을 기준으로 0.1에서 0.3점 감점하여 평가한다. 본 연구에서 사용한 채점점수의 구성표는 Table 2.3과 같다.

Table 2.3 Check list

	Day				
	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5
subject1	-	-	-	-	-
subject2	-	-	-	-	-
⋮	⋮	⋮	⋮	⋮	⋮
subject19	-	-	-	-	-
subject20	-	-	-	-	-

2.2.1. G-연구: p × d × r 설계

본 연구에서는 채점자 (rater, r) 각자가 모든 피험자 (person, p)의 품새 시행에 대한 수행동작을 채점하였다. 따라서 피험자 (p)와 2개의 측정조건이 서로 교차되어 있는 2국면 완전교차모형 (two-facet completely crossed design)이며, 이 때 측정조건 (condition of measurement)인 채점자 (r), 경기일 (d)은 전집 (universe)이 된다. 본 연구 설계에서 피험자와 채점자, 경기일은 무한대의 모집단 ($N_p \rightarrow \infty$)과 전집 ($N_r \rightarrow \infty, N_d \rightarrow \infty$)으로부터 무선 추출되었음을 가정하였다. 따라서 본 설계는 임의효과 모형 (random effect model)이며 측정항목에 대한 채점자와 관찰시기 또는 경기 일이 달라짐에 관계없이 동일한 관찰자가 피험자를 관찰하는 균형설계 (balanced design)이다. 본 연구의 분산성분을 추정하기 위한 분산성분을 추정하기 위한 선형방정식 모형은 Figure 2.1과 같다.

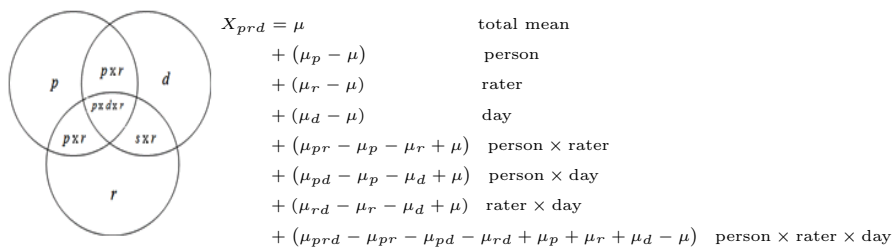


Figure 2.1 Experimental designs and estimates of variance component for two-facet designs in g-study (p × d × r)

2.2.2. D-연구: p × D × R 설계

일반화가능도 이론의 D-연구는 G-연구를 통해 추정된 분산성분을 활용한 의사결정 과정이다. 본 연구에서는 G-연구와 D-연구를 동일하게 설계하였으며, 모두 무선효과 모형을 가정하였다. 또한 본 연구를 통해 일반화하고자 하는 측정조건, 즉 일반화 전집은 채점자 국면, 경기일 국면으로 정의하였다. 본 연구에서 피험자의 분산 (σ_p^2)은 모든 오차국면의 효과를 포함한 즉, 전집 점수의 분산을 나타낸다. 즉, X의 분산을 $\sigma^2(X)$ 라고 할 때, 표본 평균 \bar{X} 의 분산 $\sigma_{\bar{X}}^2$ 는 $\sigma^2(X)$ 를 표본의 크기 n으로 나눈 값과 같다. 이를 응용하면 G-연구결과를 토대로 일반화 전집을 만족하는 조건의 평균 점수에 관심을 갖는 D-연구의 분산성분은 G-연구 분산성분추정치에 해당 효과의 표본크기로 나눈 값이 된다. D-연구에서는 G-연구를 통해 추정된 전집점수와 각 오차국면의 분산성분을 바탕으로 일반화 전집을 만족하는 조건의 평균점수를 추정한다. D-연구의 분산성분은 Figure 2.2, 일반화가능도 계수 산출 공식 (calculation of g-coefficient)은 식 (2.1)과 같다. $\sigma^2(p) + \sigma^2(\delta)$ 부분이 관찰점수가 되는데 이때 오차원으로 상대오차 $\sigma^2(\delta)$ 을 사용한다. 이 의미는 상대적 결정을 위해서 측정대상과 상호작용하는 국면의 분산만을 포함하는 오차인 상대오차만을 오차원으로 고려하여 신뢰도를 추정하겠다는 것으로 이때의 계수를 일반화가능도 계수라고 한다.

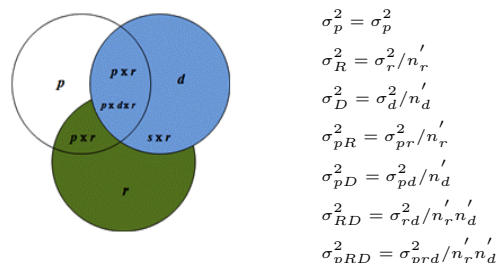


Figure 2.2 Experimental designs and estimates of variance component for two-facet designs in d-study (p × D × R)

$$E(\rho^2) = \frac{\sigma^2(p)}{\sigma^2(p) + \sigma^2(\delta)}$$

(2.1)

$E(\rho^2)$: G - coefficient, $\sigma^2(p)$: variance of universe score
 $\sigma^2(\delta)$: variance of reative error

2.3. 자료 처리

일반화가능도 이론을 적용한 주관적 태권도 품새 평가 점수의 신뢰도를 추정하기 위해 G-연구와 D-연구 설계로 분석을 하였으며, 측정설계모형 (measurement design model: $p \times d \times r$)으로 피험자 (person) by 경기일 수 (days) by 채점자 (raters)를 국면 (facet)으로 설정한 교차설계 (crossed design)에 의한 분산성분 추정치를 통해 각 오차원의 상대적 크기 (%)를 산출하였으며, D-연구의 결과는 상대적 결정 (relative decision)과 절대적 결정 (absolute decision) 중 상대적 결정에 의한 G-계수로 결과를 해석하였다. 이때 적정수준의 일반화가능도 계수는 Mehrens과 Lehmann (1973)이 제안한.80으로 설정하였다. 본 연구 설계를 분석하기 위해 사용한 Statistical Package는 Crick과 Brennan (2001)이 개발한 GENOVA a FORTRAN program (GENeralized analysis Of VAriance) 3.1 Ver program을 이용하여 분석을 하였다.

3. 결과 및 논의

이 연구에서는 일반화가능도 이론을 적용하여 태권도 품새 선수와 채점자들을 대상으로 한 ‘정확성 및 표현성’ 평가항목에 대한 신뢰도를 추정하였다. 연구 설계에 따라 ‘정확성 및 표현성’의 분산성분 추정치의 상대적 크기를 분석하는 G-연구와 이를 바탕으로 최적의 측정조건을 추정하는 D-연구를 진행하였다.

3.1. ‘정확성’ G-연구 결과 : $p \times d \times r$ 설계

태권도 품새 경기 채점 항목인 정확성과 표현성 2개 평가항목의 경기일 수와 채점자 수에 따른 정확성 평가항목에 G-연구 결과 나타난 전집점수의 분산과 각 오차원의 분산의 상대적 크기는 Table 3.1과 같다. 전집 분산인 피험자 분산 (s)은 측정결과에서 나타나는 개인 간 차이를 의미하는 것으로 정확성 평가항목에 대한 분산성분 추정치에서는 전체 분산성분에서 피험자간 분산성분은 0.0156316로 전체 분산성분 중 13.6%를 차지하고 있는 것으로 나타나 전체 분산성분 중 상대적으로 작은 것으로 나타났다. 이는 태권도 품새 동작의 정확성에 대한 기술적 부분이 전체 관측점수에 영향을 미치는 개인차가 크지 않다는 것으로 의미한다. 그러나, 선수 개인간 능력의 차이에 따라 전체 관측점수가 달라질 수 있다는 것을 시사할 수도 있다. 전집 분산과 잔차 분산을 제외한 오차 분산 중 가장 큰 비율을 차지하는 분산은 채점자간 분산으로 분산성분은 0.0695816로 전체 분산성분 중 60.8%로 상대적으로 높은 것으로 나타났다. 피험자x채점자간 상호작용 분산은 0.0292684로 전체 분산성분 중 전체분산에 25.6%로 나타났다.

Table 3.1 Estimated variance component for ‘Accuracy’ G-study $p \times d \times r$

Source of Variance	df	SS	MS	$\hat{\sigma}^2$	% VAR
p	19	2.0411	0.10743	0.0156316	13.6
r	4	5.6836	1.42090	0.0695816	60.8
$p \times r$	76	2.2244	0.02927	0.0292684	25.6
Total	99	9.9491		0.1144816	100.0

3.1.1. '표현성' G-연구 결과: $p \times d \times r$ 설계

태권도 품새 경기의 표현성 평가항목에 대한 경기일 수와 채점자수에 따른 G-연구결과 나타난 전집점수의 분산과 각 오차원 분산의 상대적 크기는 Table 3.2와 같다. 전집 분산인 피험자 분산 (s)은 측정결과에서 나타나는 개인 간 차이를 의미하는 것으로 표현성 평가항목에 대한 분산성분 추정치에서 전체 분산성분에서 피험자간 분산성분은 0.0162947로 전체 분산성분 중 34.0%를 차지하고 있는 것으로 나타나 전체 분산성분 중 상대적으로 높은 것으로 나타났다. 이는 태권도 품새 동작의 표현성에 대한 태권도 품새 기술의 표현 부분이 전체 관측점수에 영향을 미치는 개인차가 상대적으로 크다는 것을 의미한다. 또한 전집 분산과 잔차 분산을 제외한 오차 분산 중 가장 큰 비율을 차지하는 분산은 피험자 \times 채점자간 상호작용 분산 0.0245053으로 전체 분산성분 중 60%로 상대적으로 매우 높은 비율을 차지하고 있는 것으로 나타났다. 반면에 채점자간 분산 성분은 0.0으로 전체 분산성분 중 0%를 차지하고 있는 것으로 나타나 채점자간에 의한 오차 영향력이 없는 것으로 나타났다.

Table 3.2 Estimated variance component for 'Expressivity' G-study $p \times d \times r$

Source of Variance	df	SS	MS	$\hat{\sigma}^2$	% VAR
p	19	2.01360	0.10598	0.0162947	34.0
r	4	0.07360	0.01840	0.0	0.0
$p \times r$	76	1.86240	0.02451	0.0245053	60.0
Total	99	3.94960		0.0408	100.0

위 G-연구에서 나타난 결과에 대한 논의는 다음과 같다. 본 연구에서 적용한 일반화가능도 이론은 고전검사 이론의 전집수 논리에 이론적 근거를 두고 있다. 즉, 측정을 통해 나타난 관측점수는 피험자의 실제 능력외의 다양한 오차원을 포함하고 있음을 가정하고, 이러한 오차원의 크기를 변량 (분산)분석법을 통해 규명하는 분석방법이다. 이에 본 연구에서 전체 분산성분 추정치 중 정확성 평가항목에 대한 태권도 품새 대회에 출전한 선수들의 개인차가 차지하는 비중이 상대적으로 작다는 것을 의미하는 반면, 채점자들이 차지하는 비중이 상대적으로 크다는 것으로 선수들의 경기력에 절대적 영향을 미치는 것으로 해석할 수 있다. 그러나 표현성 평가항목에 대해서 채점자들이 차지하는 비중이 낮은 것으로 나타나 경기력에 영향을 미치지 않은 것으로 나타났다. 이러한 결과가 나타난 것은 본 연구에서 수집한 태권도 품새 대회의 채점자들이 본 대회에서 선수들의 경기를 보다 엄격하고 일관성 있는 평가에 임하여 채점자들의 채점점수들의 차이가 나타나지 않는 결과로 판단해 볼 수 있을 것이다. 일반화가능도이론 (g-theory)을 이용하여 태권도 겨루기 경기의 심판판정의 신뢰도를 분석한 연구 (Jung, 2009)들에서도 이 연구의 결과에서 나타난 것과 같이 태권도 선수의 개인차에 의한 분산성분 추정치보다 심판경력에 의한 분산성분 추정치가 다른 오차원보다 상대적으로 높은 것으로 나타나는 것으로 평가되고 있으며, 심판 수의 분산성분 추정치도 오차원으로 영향을 미치는 것으로 제시하였다. 이러한 결과는 본 연구의 결과를 지지하고 있는 것으로 볼 수 있다. 그러나 본 연구결과에서 표현성 평가항목에 대한 결과에서 채점자 즉, 심판의 분산성분 추정치가 0.0 (0%)로 나타나 채점자가 오차원 (error source)으로 작용하지 않아 판정의 결과에 영향을 미치지 않은 것으로 나타났다.

이러한 결과는 체조나, 무용, 다이빙과 같이 선수들의 행위나 수행된 동작에 의해 경기력을 평가하는 종목들의 특성이 객관적 평가기준이 각 종목마다 있음에도 불구하고 전문성이라는 명분하에 심사자들의 주관적 판정에 의해 경기력을 평가하여 순위나 등위를 매기는 특성을 가지고 있기 때문에 심사자들이 평가에 대해 절대적 영향을 가지고 있다. 이러한 절대적 영향을 주는 것으로 심판의 경력, 심판의 수, 배정된 심판의 국가 및 출신에 따라 판정에 많은 영향을 미치는 것으로 다양한 선행연구에서 나타난 결과 (Kim과 Heo, 2002)들을 비추어 볼 때 상반된 결과가 나타났다. 이러한 상반된 결과는 운동기술을 평가하기 위한 준거지향검사 도구의 문제로 지적되어 오던 채점자의 자의성 (arbitrariness)과 분류오

류 (misclassification)가 태권도 품새 기술의 평가항목인 ‘표현성’의 측정결과에 내재하고 있는 이유로도 볼 수 있다는 판단과 지극히 채점자 즉, 전문가의 직관 및 주관적 판단에 의존할 수 밖에 없는 평가항목이기 때문에 보다 전문적인 교육과 강도 높은 심판 훈련을 받은 채점자들에 의한 결과가 결국, 준거 기준에 대한 교육이 선행되면서 ‘표현성’ 평가항목에 대한 채점결과의 신뢰도 높아질 수 있기 때문에 나타난 결과로도 판단을 할 수 있을 것으로 사료된다. 따라서 이러한 심판관정과 관련한 연구를 수행한 다양한 연구들에서 심판의 오차원의 영향력을 줄이고 판정의 객관성 (objectivity)을 높이기 위해서는 꾸준하고 지속적인 심판 판정의 훈련을 통한 교육이 중요하다고 지속적으로 주장을 해왔다 (Kang과 Ann, 1999; Ahn, 1999; Kim과 Kang, 2003; Cho 등, 2005; Looney, 1997; Zhu 등, 1998).

3.2. ‘정확성’ D-연구 결과: $p \times D \times R$ 설계

D-연구에서는 앞서 G-연구를 통해 산출된 분산성분 추정치를 바탕으로 각 평가항목의 오차 국면의 차원을 조정함으로써 $p \times D \times R$ 설계의 D-연구를 수행하였다. 한편, 본 연구에서는 D-연구를 통해 산출되는 의존도 계수를 일반적으로 수용되는 신뢰도 기준 (>.80)과 비교적 양호한 수준의 신뢰도 기준 (>.70)을 고려하여 해석하였다 (Nunally, 1978; Kim, 1989; Shavelson 등, 1993; Kim, 1993; Webb 등, 2000). D-연구결과 산출된 의존도 계수를 산출하였으며 Table 3.3과 같다. 본 연구에서 태권도 품새 ‘정확성’ 평가항목의 경우 기본조건 (NRA=1), 즉 일반화전집은 허용 가능한 관찰전집의 모든 조건을 포함할 수도 있고, 허용 가능한 관찰 전집의 일부만을 포함할 수 있는 조건일 때, 채점자가 1명일 때 결정연구의 의존도 계수가.34814로 상대적으로 낮은 계수를 얻었다. 이러한 결과는 D-연구에서는 측정 대상에 대한 의사결정은 측정조건의 평균점수 (XpIR)에 기초하기 때문에 D-연구의 분산 성분은 평균점수 (XpIR)의 분산이라는 점이 G-연구의 분산과는 다른 특징을 가진다. 따라서 태권도 품새의 ‘정확성’ 평가항목에 대한 적정수준의 채점자 수는 적정수준의 신뢰도 기준 (>.7)을 고려했을 때, 신뢰도를 향상시킬 수 있는 채점자의 수는 5명이상인 것으로 나타났다.

Table 3.3 Estimated variance component for ‘Accuracy’ D-study $p \times D \times R$ design

Condition of Facets					
p	R	G-coefficient	p	R	G-coefficient
20	1	0.34814	20	5	0.72755
20	2	0.51648	20	6	0.76216
20	3	0.61571	20	7	0.78896
20	4	0.68115	20	8	0.81034

3.3. ‘표현성’ D-연구 결과: $p \times D \times R$ 설계

표현성 평가항목에 대한 G-연구 결과 산출된 분산성분 추정치를 이용하여 $p \times D \times R$ 설계의 D-연구를 수행하였다. D-연구 결과 산출된 의존도 계수를 Table 3.4에 제시하였다. Table 3.6에서 나타나듯이, 태권도 품새의 표현성 평가항목의 경우 기본조건 (NRA=1)에서 .39938로 상대적으로 낮은 의존도 계수를 확보하였다. 따라서 태권도 품새의 ‘표현성’ 평가항목에 대한 적정수준의 채점자 수는 신뢰도 기준 (>.7)을 고려했을 때, 신뢰도를 향상시킬 수 있는 채점자의 수는 4명이상인 것으로 나타났다. 이는 적정수준의 의존도 계수를 얻기 위한 채점자 수가 각각의 평가항목별 적정수준의 채점자검사자수 보다 많음을 나타내고 있다. 이는 여러 사람이 공동으로 채점하거나 평가하는 다면적 평가의 경우, 심사자나 평가자가 많으면 많을수록 좋으나 실제적으로 주관적 평가를 하는 종목에서 평균적으로 3명에서 5명 정도면 적정하다고 전문가들이 제언을 하였다 (Yang, 1999). 이에 본 연구결과에서도 선행연구 및 전문가들이 제안한 적정수준의 채점자 범위에서 결과가 나타남을 확인할 수 있었다.

Table 3.4 Estimated variance component for 'Expressivity' D-study $p \times D \times R$ design

Condition of Facets					
p	R	G-coefficient	p	R	G-coefficient
20	1	0.39938	20	5	0.76877
20	2	0.57080	20	6	0.79959
20	3	0.66609	20	7	0.82315
20	4	0.72676			

D-연구 (결정연구)는 G-연구를 통해 추정된 오차원의 분산성분추정치 (EVC)를 바탕으로 이러한 오차원의 분산성분을 최소화하여 적정수준의 신뢰도를 확보할 수 있는 오차 국면의 크기, 즉 최적의 측정조건을 제시하는 의사결정의 단계이다. 본 연구에서의 오차 국면의 크기는 채점자 5명, 시행일 1일 그리고 측정항목인 정확성 및 표현성 2개로 구성되어 있다. 연구결과의 일반화 가능성을 높이기 위하여 측정항목 국면을 일반화 전집에 포함하고 고정효과 모형 (fixed effect model)이 아닌 임의효과 모형 (random effect model)을 적용하여 분석한 결과 측정조건에 따라 최소 4명이상의 채점자에 의한 평가 결과의 신뢰도 (>.7)이상의 의존도 계수를 산출하였다. 따라서 신뢰도를 향상시키기 위한 접근에 대한 고민은 결국 채점자 측면에서 이루어져야 할 필요성을 제기할 수 있을 것으로 판단된다.

이러한 채점자, 심사자, 또는 평가자 등에 의해 평가결과의 신뢰도에 영향을 받는 다양한 선행연구로 체조심판 판정의 오차원 분석 (Ahn, 1997)연구, 주관적 배구기능검사의 신뢰도 추정 (Heo, 2003), 또는 농구기능검사의 신뢰도 추정 (Cho, 1991)등 국내 연구에서도 척도를 개발하고 G-Theory를 적용하여 신뢰도를 규명한 연구에서도 채점자의 신뢰도를 추정한 연구결과들을 제시하였다. 이러한 결과의 근본적인 원인은 일반화가능도 이론이 '행동측정의 신뢰성'에 대한 측정이론이기 때문인 것으로 설명할 수 있다. 다시 말해서 일반화가능도 이론은 피험자를 평가 혹은 관찰하는데 있어서 신뢰할 수 있는 결과를 얻으려면 얼마나 많은 평가자 혹은 관찰자가 필요한가를 결정짓기 위하여 적용하는 이론이라고 할 수 있다. 따라서 매년 피험자의 행동을 관찰하는 관찰자의 관찰행위 간의 상관정도가 일반화가능도를 높이는 데 중요한 역할을 하기 때문이다. 따라서 다양한 스포츠 종목들 중 심사자, 심판에 의해 경기력을 판정하는 종목들은 심판판정결과에 영향을 미치는 다양한 오차원의 그 상대적 크기를 추정하는 것과 결정연구 (D-study)를 통해 태권도 품새 경기현장에서 실시하고 있는 심판들의 판정결과가 어느 정도 신뢰성을 가지고 있는지를 확인하였으며, 적정수준의 신뢰도를 확보하기 위한 수정 및 보완의 이론적 근거를 제시하고 그 의미를 탐색하는 것은 의미 있는 연구라고 사료된다. 따라서 이러한 결과는 주관적 판정에 의해 선수들의 경기력을 평가하는 종목에서 심사자의 결과자료에 근거하여 심사자 선발 근거자료로서의 활용과 심사자 교육 및 훈련을 통해 심사자간 평가결과의 일관성을 유지할 수 있도록 교육 자료로서의 활용가능성과 객관적 근거를 제시할 수 있을 것으로 사료된다.

4. 결론

이 연구는 일반화가능도 이론을 적용하여 태권도 심판 즉, 채점자를 대상으로 태권도 품새 기술 평가 항목인 '정확성 및 표현성' 측정점수에 내재하고 있는 오차원의 상대적 크기를 규명하고, 더 나아가 적정 신뢰수준을 확보할 수 있는 최적의 측정조건을 제시하는데 그 목적이 있으며, 다음과 같은 결론을 얻었다. 일반화가능도이론 (G-theory)의 G-연구 결과, '정확성'에 대해 분산성분 추정치를 추정한 결과, 오차의 영향력은 채점자 (raters)국면에서 가장 크게 나타났으며, 각 설계에 따른 상호효과 (피험자 내), 피험자 간 순서로 나타나 채점자 요인이 측정 점수의 신뢰도에 부정적인 영향을 미치는 것으로 나타났다. 또한 '표현성'에 대해 분산성분 추정치를 추정한 결과, 오차의 영향력은 상호효과 (피험자 내)국면에서 가장 크게 나타났으며, 각 설계에 따른 피험자 간, 심사자 국면 순서로 나타났다. 이러한 결과는 주

관적 관정에 의해 선수들의 등위 및 순위가 결정되는 종목 특성에서 심판이 절대적 역할을 함에도 불구하고 채점자 오차원의 분산성분 추정치가 낮게 추정되었으며, 상호작용에 대한 분산성분추정치가 상대적으로 가장 높은 비율로 나타나, 본 연구에서 고려하지 못한 즉, 포함 되지 못한 오차 국면 이외의 알 수 없는 오차가 측정 점수에 상대적으로 영향을 미치는 것으로 의미한다. 따라서 이를 규명하기 위한 후속 연구의 필요성이 제기된다. 마지막으로 결정연구 (D연구)를 통하여 의존도 계수를 추정한 결과, 채점자 수에 따른 최적수준의 측정조건으로 일반적으로 수용되는 신뢰도 기준 (>.7)을 고려하여 분석한 결과, ‘정확성’ 평가항목에서 5명이상의 채점자일 때 안정적인 신뢰도를 얻을 수 있으며, ‘표현성’ 평가항목에서는 4명이상의 채점자일 때 안정적인 신뢰도를 얻었다 (Choi, 2014; Lee, 2010).

부록

Taekwondo Poomsae Judge's Score Data Set (Accuracy data set)						Taekwondo Poomsae Judge's Score Data Set (Expressivity data set)							
No	Name	Day 1					No	Name	Day 1				
		Rater1	Rater2	Rater3	Rater4	Rater5			Rater1	Rater2	Rater3	Rater4	Rater5
1	Lim○○	1.7	2.7	2.6	2.5	2.1	1	Lim○○	6.3	6.5	5.9	6.2	6
2	Maeng○○	1.7	2.7	2.4	2.2	2.3	2	Maeng○○	5.9	5.6	5.8	5.8	5.8
3	Kwon○○	2.0	2.6	2.2	2.3	1.7	3	Kwon○○	6.6	5.8	6.0	5.8	5.8
4	Ye○○	1.8	2.3	2.3	2.3	2.4	4	Ye○○	5.9	5.7	6.0	5.8	5.9
5	Jang○○	1.4	2.2	2.2	2.1	2.1	5	Jang○○	5.9	5.8	5.8	5.6	5.7
6	Shin○○	1.2	2.2	2.3	2.2	2.1	6	Shin○○	5.6	5.8	5.9	5.6	5.8
7	Park○○	1.8	2.3	2.4	2.4	2.0	7	Park○○	5.7	5.7	5.7	5.9	5.6
8	Jang○○	1.7	2.4	2.3	2.5	2.3	8	Jang○○	5.7	5.7	5.8	6.0	5.9
9	Lee○○	1.9	2.3	2.3	2.5	2.2	9	Lee○○	6.4	5.9	5.9	6.1	5.9
10	Jeon○○	1.3	2.1	1.7	2.1	2.2	10	Jeon○○	5.7	6.0	5.8	6.1	5.8
11	Noh○○	1.8	2.5	2.5	2.5	2.5	11	Noh○○	5.8	5.8	6.0	6.2	6.0
12	Park○○	1.8	2.3	1.9	2.2	2.2	12	Park○○	5.9	5.9	5.8	5.8	5.7
13	Kim○○	2.1	2.2	2.3	2.6	2.3	13	Kim○○	5.7	5.9	6.3	6.1	5.9
14	Park○○	1.6	2.4	2.4	2.4	2.1	14	Park○○	6.0	5.7	5.9	5.7	5.8
15	Kwak○○	2.1	2.4	2.5	2.5	2.4	15	Kwak○○	5.9	6.0	6.1	6.1	6.0
16	Cho○○	2.0	2.5	2.3	2.5	1.9	16	Cho○○	5.9	5.8	5.8	6.0	5.8
17	Bae○○	1.7	2.1	2.3	2.1	2.0	17	Bae○○	5.6	5.7	5.6	5.6	5.7
18	Park○○	1.4	2.3	2.1	2.1	2.3	18	Park○○	5.6	5.6	5.8	5.8	5.8
19	Kim○○	1.3	1.9	2.3	2.1	2.0	19	Kim○○	5.6	5.6	5.6	5.6	5.6
20	Lim○○	1.5	2.2	2.1	2.1	2.2	20	Lim○○	5.6	5.8	5.9	5.7	5.8

References

- Ahn, A. J. (1997). *Application of generalizability in analyzing sources of error variance of gymnastic judgement*, Ph. D. Thesis, Myungji University, Kyeonggido.
- Allen, M. J. and Yen, W. M. (1979). *Introduction to measurement theory*, 1st ed., Waveland Press, Illinois.
- Brennan, R. L. (1983). *Elements of generalizability theory*, ACT Publication, Iowa.
- Cho, E. H. (2014). *Development of computer-based assessment instrument and movement classification for poomsae skills in taekwondo*, Ph. D. Thesis, Sungkyunkwan University, Kyeonggido.
- Cho, E. H., Kwon, J. H., Cha, C. H. and Kim, K. S. (2005). The severity and consistency of the judge's decision in taekwondo. *Journal of Sport Leisure Studies*, **25**, 85-95.
- Cho, J. H. (1991). *Generalizability coefficient estimation of skinfold thickness test and basketball skill test*, Ph. D. Thesis, Korean National College of Physical Education, Seoul.
- Cho, J. H. (2006). An analysis of sources of error and optimal condition of generalizability coefficient in the Korean national bowling player selective match. *The Korean Journal of Physical Education*, **45**, 457-167.

- Choi, J. S. (2014). Projection analysis for two-way variance components. *Journal of the Korean Data & Information Science Society*, **25**, 547-554.
- Choi, Y. L. (2013). *The olympic experience and expertise of taekwondo national representative coach*, Ph. D. Thesis, Korea National Sport University, Seoul.
- Cronbach, L. J., Gleser, G. C., Nanda, H. and Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalization for scores and profiles*, Wiley, New York.
- Hagtvet, K. nut A. (1997). The function of indicators and errors in construct measures: An application of generalizability theory. *Journal of Vocational Education Research*, **22**, 247-266.
- Heo, C. K. (2003) *Generalizability coefficient estimation of subjectivity volleyball test*, Ph. D. Thesis, Kookmin University, Seoul.
- Jung, Y. H. (2009). *Analysis on reliability of referees judgment in Taekwondo competition applying generalizability theory*, Ph. D. Thesis, Kookmin University, Seoul.
- Kang, S. J. and Ann, A. J. (1999). Rater's objectivity in the assessment of gymnastic performance. *The Korean Journal of Physical Education*, **38**, 641-650.
- KERIS (2015). <http://www.riss.kr>.
- Kim, D. Y. and Heo, C. K. (2002). Generalizability coefficient estimation of subjectively volleyball test. *The Korean Journal of Measurement and Evaluation in Physical Education and Sport Science*, **4**, 15-28.
- Kim, M. Y. and Kang, S. J. (2003). Examining rater errors in performance assessment with a many-faceted rasch model. *The Korean Journal of Physical Education*, **42**, 1019-1027.
- Kim, S. S. (1989). *An analysis of sources of variation in teacher behaviors using generalizability theory*, Ph. D. Thesis, University of Virginia, Virginia.
- Kim, S. S. (1993). The generalizability of student ratings of instructors across sections. *The Korean Journal of Educational Research*, **31**, 23-40.
- Kim, Y. B. and Kim, S. S. (2001). *Generalizability theory*, Kyohaksa of Publisher, Seoul.
- Ko, B. G. (1997). A Generalizability of physical performance tests for apparatus exercises in elementary school. *Journal of the Research Institute of Physical Education*, **18**, 111-126.
- Kukkiwon (2015). *The countries geups & dans situation*, Retrieved March 5, 2015, from <http://www.kukkiwon.or.kr>.
- Lee, J. T. (2010). A Study on the performance pattern in the elite decathlon. *Journal of the Korean Data & Information Science Society*, **21**, 1071-1079.
- Lee, J. Y. (2005). *Estimating the reliability of the selection method of the national team bowlers; The application of generalizability theory*, Ph. D. Thesis, Korea National Sport University, Seoul.
- Looney, M. (1997). Objectivity measurement of figure skating performance. *Journal of outcome measurement*, **1**, 143-163.
- MacMillan, P. D. (2000). Classical, generalizability, and multifaceted rasch detection of interrater variability in large, sparse data sets. *The Journal of Experimental Education*, **68**, 167-190.
- Nunally, J. C. (1978). *Psychometric theory*, 2nd ed., McGraw-Hill, New York.
- Park, H. C. (2014). Comparison of confidence measures useful for classification model building. *Journal of the Korean Data & Information Science Society*, **25**, 365-371.
- Park, H. J. and Pi, S. Y. (2013). Study on the K-scale reflecting the confidence of survey responses. *Journal of the Korean Data & Information Science Society*, **24**, 41-51.
- Safrit, M. J., Atwater, A. E., Baumgarther, T. A and West, C. (1976). *Reliability theory*, American Alliance for Health, Physical Education and Recreation, Washington.
- Shavelson, R. J and Webb, N. M. (1981). Generalizability theory: 1973-1980. *British Journal of Mathematical and Statistical Psychology*, **34**, 133-166.

- Shavelson, R. J., Baxter. G. P. and Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement*, **30**, 215-232.
- Sun, A., Michael J. V. and Gao, X. (1997). Using generalizability theory to assess the reliability of student ratings of academic advising. *The Journal of Experimental Education*, **65**, 367-379.
- WTF (2014). <http://www.wtf.org>.
- Zhu, W., Ennis, C. D. and Chin, A. (1998). Many faceted rasch modeling expert judgement in test development. *Measurement in physical education and exercise science*, **2**, 21-39.

Analysis of error source in subjective evaluation results on Taekwondo Poomsae: Application of generalizability theory

Eun Hyung Cho¹

¹The Research Institution for Physical Fitness and Sport Science, Sungkyunkwan University

Received 4 January 2016, revised 21 January 2016, accepted 7 March 2016

Abstract

This study aims to apply the G-theory for estimation of reliability of evaluation scores between raters on Taekwondo Poomsae rating categories. Selecting a number of game days and raters as multiple error sources, we analyzed the error sources caused by relative magnitude of error variances of interaction between the factors and proceeded with D-study based on the results of G-study for optimal determination of measurement condition. The results showed below. The estimated outcomes of variance component for accuracy among the Taekwondo Poomsae categories with G-theory showed that impact of error was the biggest influence factor in raters conditions and in order of interaction in subjects and between subjects, also impact of variance component estimation error on expression category was the major influence factor in interaction and in order of the between subjects and raters. Finally, the result of generalizability coefficient estimation via D-study showed that measurement condition of optimal level depend on the number of raters was 8 persons of raters on accuracy category, and stable reliability on expression category was gained when the raters were 7 persons.

Keywords: Generalizability theory, g-coefficient, multiple source error, reliability, variance component.

¹ Senior researcher, The Research Institution for Physical Fitness and Sport Science, Sungkyunkwan University, Gyeonggi-do 440-746, Korea. E-mail: taedori1974@hanmail.net