서울大學校 師大論叢 第55集 (1997. 12. 30)

심동적 영역에서 준거지향기준의 설정*

김 종 택
(체육교육과)

Ⅰ. 연구의 목적 및 필요성

體育의 궁극적인 목적은 신체활동의 과학적 규명으로 기존의 여러 학문을 망라한 종합과학 또는 인간 움직임(human movement)과 관련된 예술로서 정의된다(Nixon과 Jewett, 1980). 그리고 體育은 全人 완성을 위한 신체의 교육 또는 신체를 통한 교육으로 오늘날에는 학문적 영역으로 분류되는 스포츠 과학 혹은 운동과학과 구분되는 체육학의 전문영역으로 독자적인 학문적 체계를 이루면서 발전하고 있다.

학교 체육에서 주요한 주제는 심동적 영역의 내용으로 運動能力이나 體力 등이 대표적이며 심동적 영역의 평가는 필기검사보다는 실기검사가 일반적이다. 평가는 교육과정이나 학습 프로그램뿐만 아니라 학생의 성취도나 수업목표 달성 여부를 판단하는 주요한 교육의 한 과정으로 학교 체육의 정상화를 위한 주요한 요인 중의 하나이다.

지금까지 체육 현장에서 실기평가는 교육적 관점에서 학습자의 學習目標 달성 여부를 진단하는 준거지향검사 혹은 절대평가의 유용성이 인정되면서도 학생간의 서열과 개인 차이를 구분하기 위한 규준지향검사 혹은 상대평가가 실시되었다. 그러나 교육적 관점에서 절대평가가 강조되고 체육교사의 약 80% 그리고 일반교사의 약 86%가 준거지향검사를 선호하는 것으로 선행연구에서 보고되었다(신승윤, 1989; 신세호, 1992). 학교 체육에서 상대평가가 실시되고 있는 이유는 실기평가를 위한 準據指向基準(CRS; criterion referenced standard)의 개발이 이루어지지 않았고 CRS 작성을 위한 체육교사의 능력 부족으로 사료된다. 그리고 CRS 작성을 위한 과학적이고 논리적인 이론보다는 체육교사의 경험에 근거하여 전문적 판단에 의존한 CRS의 개발로 제한되었기 때문이다. 또한, 전반적인 교육 여건이 교육이 궁극적으로 지향해야 할 발달적 교육관 보다는 선발적 교육관이 강조되어 왔기 때

* 이 논문은 서울대학교 사범대학 발전기금의 지원에 의해 이루어졌음.

문인 것으로 사료된다.

선발적 교육관에 근거한 상대평가는 학생간의 서열에 의한 지적계급 사회의 이념을 정당화하고, 학우간의 경쟁의식을 고취시키는 비교육적 영향뿐만 아니라 교육목표의 달성 여부에 대한 정보를 주지 못한다는 문제점이 지적된다. 오늘날에는 절대평가를 선호하는 경향으로 전반적인 교육 여건이 변해가면서 학교 체육에서도 이에 대한 관심이 고조되고 절대평가가 권장되고 있다(Hensely, Lambert, Baumgartner 그리고 Stillwell, 1987). 특히, 학생의 체력검사에서 CRS의 필요성이 대두되면서 미국의 체력검사에서 1980년 후반부터 規準指向基準(NRS; norm referenced standard)보다 CRS의 사용이 늘어나고 있는 추세이다(Safrit와 Wood, 1995).

교육학 분야에서 오래전부터 CRS에 대한 많은 연구가 이루어졌으나 체육학에서 CRS에 대한 연구는 미흡한 실정이다. 특히, 학교 체육 현장에서 실기평가를 할 때 체육교사가 활용 가능한 CRS나 CRS의 설정 방법에 대한 연구가 요망된다. 따라서 본 연구에서는 CRS에 관한 선행연구를 고찰하여 실기평가를 위한 CRS 설정 요령과 설정 방법에 따른 CRS의 장·단점 그리고 CRS의 妥當度나 信賴度 검증 요령을 모색하고자 한다. 본 연구의 결과는 체육 현장에서 실기평가를 실시할 때 CRS의 활용과 설정에 대한 이론적 기초를 제공할 것으로 사료된다.

Ⅱ. 연구의 내용 및 방법(국내외의 연구동향 포함)

교육학 분야에서는 CRS에 대한 많은 연구가 이루어졌고 CRS 설정에 대한 여러 가지 방법이 개발되었다(박광배, 1992; Berk, 1976). 대표적인 CRS 설정의 방법은 전문가의 판단에 근거한 判斷的 模型, 경험적 자료를 이용한 經驗的 模型, 그리고 混合 模型 등이 있다. 교육학에서 이루어진 CRS에 관한 연구 결과는 체육학 분야에서 응용 가능성이 높다. 그러나 체육학 평가에서 주된 주제가 심동적 영역의 실기검사로 인지적, 정의적 영역의 필기검사가 주된 주제인 교육학 분야와 다른 특성이 있기 때문에 교육학 분야에서 이루어진 선행연구의 결과를 응용하는데 제한점도 사료된다.

체육학에서도 1970년대부터 심동적 영역의 평가를 위한 CRS에 대한 연구가 이루어지고 있으나, 대부분의 연구는 CRS 설정 방법에 따른 CRS의 실용화 가능성에 대한 이론적 논의이다(Looney, 1987; Safrit, 1989; Safrit와 Stamm, 1980). 최근에는 Kalohn 등(1992)의 테니스 기술의 평가에서 CRS 설정 방법별 비교 등의 연구를 통하여 체육 현장에서 활용 가능한 CRS 설정 방법의 모색이 이루어지고 있

다. 그리고 Looney와 Plowman(1990), Rutherfold와 Corbin(1994)의 CRS의 타당성에 관한 연구가 수행되었으며 1980년후반 부터 미국의 학생 체력평가에서 CRS의 사용이 보편화 되기 시작하였다(AAHPERD, 1988; Institute for Aerobics Research, 1987).

국내에서는 신승윤(1994)의 볼링 기술 평가에서 CRS 설정 방법별 비교연구나 조기정(1996)의 逐次確率比 檢査 등 실기평가에서 CRS와 관련된 몇 편의 연구가 이루어지고 있다. 그러나 체육 교사들이 현장에서 활용 가능한 실기평가를 위한 CRS나 CRS 설정을 위한 체계적인 이론적 고찰이 미흡한 실정이다.

따라서 본 연구에서는 교육학과 체육학 분야에서 이루어진 선행연구를 고찰하여 실기평가를 위한 CRS에 대한 이론적 체계를 모색하고자 한다. CRS에 관한 이론적 고찰의 세부 내용은 다음과 같다.

1) 준거지향기준의 특성
2) 준거지향기준의 설정 방법
3) 준거지향기준의 평가

Ⅲ. 준거지향기준의 특성

개인이나 집단이 특정 영역에서 가지고 있는 능력을 평가하기 위해서는 측정이라는 능력의 계량화 과정이 필요하며 측정을 통해 수집된 양적 정보는 개인이나 집단의 능력에 대한 질적인 평가의 근거가 된다. 이때 검사의 목적에 따라 검사를 통하여 획득한 측정치를 어떠한 기준에 의하여 평가할 것인가가 결정된다. 검사의 목적이 한 개인의 능력을 다른 사람들과 비교하거나 상대적 서열을 정하는 것이라면 평가의 기준은 검사 집단의 점수 분포이다. 그러나 검사의 목적이 특정 목표의 도달 여부를 파악하는 것이라면 평가의 기준은 검사 집단의 점수분포와 상관없이 설정된 목표 수준이다. 특정 영역에서 성취해야만 한다고 정해진 능력 수준을 準據라 하며 이의 도달 여부를 파악하기 위한 검사를 準據指向檢査(CRT; criterion referenced test)라 한다. 즉, CRT는 피검자가 특정 영역에서 어느 정도의 능력을 가지고 있는지에 대한 직접적인 정보를 제공하며, 미래의 어떤 과제를 성공적으로 수행할 수 있는지를 판단하는 검사이다(김석우, 윤명희, 지은림 1997).

心動的 영역의 검사에서 주요 관심 중 하나는 인간의 運動能力이나 體力을 어떠한 관점에서 평가할 것인가의 문제이다. 예를들어, 한 학생의 턱걸이 수행결과가 5개이고 1000m 달리기의 수행결과가 4분 20초였다면 이 학생의 근지구력과 심폐지

구력에 대하여 어떠한 평가를 내릴 수 있을 것인가? 첫 번째 평가의 근거는 다른 사람에 비하여 잘했다 또는 못했다라는 평가를 내릴 수 있다. 두 번째 평가의 근거는 설정된 목표의 도달 여부에 대한 평가이다. 운동선수들의 경기와 관련하여 요구되는 운동 관련 체력을 평가하는 경우는 NRS를 이용한 평가가 타당하나 일반인의 건강과 관련된 체력이나 운동학습과 관련된 학생들의 체력수준을 평가하는 경우는 CRS를 이용한 평가가 보다 논리적이다. 그 이유는 CRS는 기준집단의 체력수준에 상관없이, 건강을 위해 필요한 또는 그 연령대의 발달 단계에 필요한 절대적 체력수준을 제시해주기 때문이다(Blair, Falls, 그리고 Pate, 1983; Safrit, 1981).

1963년에 Glaser가 CRS에 관한 논문을 최초로 발표한 이래 수많은 CRS 관련 연구가 이루어졌다. Popham(1978)은 CRT를 잘 정의된 행동목표에 대한 피검자의 수행 정도를 해석하기 위하여 제작된 검사라고 정의하였다. 또한 Hambleton(1994)은 CRT에서는 측정하려는 행동목표가 잘 정의되어야 하며, 측정하고자 하는 행동목표에 대한 최저 수행기준을 설정하고 이에 따라 피검자의 수행정도를 해석한다고 하였다. 그러므로 CRT에서 가장 중요한 요소는 측정할 영역을 구체적으로 정의하는 것과 타당하고 해석이 가능한 준거를 설정하는 것이다.

CRT의 특성은 規準指向檢査(NRT; norm referenced test)와의 비교로 파악될 수 있다. CRT가 특정 영역에 대한 학습자의 修行與否를 평가하기 위하여 고안된 반면, NRT는 한 학습자의 수행을 다른 학습자의 수행정도와 관련시켜 설명한다. 그러므로 CRT는 NRT에 비해 행동목표가 보다 구체적이며 검사의 분산 등에 관한 문항통계치보다는 측정하려는 행동목표와 검사의 내용이 얼마나 일치하는가에 더 큰 비중을 둔다(Popham과 Husek, 1969).

심동적 영역에서 CRS의 특성은 다음의 4가지로 요약된다(Cureton, 1990). 첫째, CRS는 해당 준거와 배타적으로 연관되어 있다. 즉, CRS는 어떤 속성에 있어서 바람직하고 세부적인 수행수준을 나타내는 준거이다. 그러므로 평가하고자 하는 속성과 관련된 검사항목으로 검사가 구성되어야 한다. 반면, NRS는 반드시 현재 평가되고 있는 속성에 대해서 만들어 질 필요가 없다. 둘째, CRS는 준거행동이나 속성의 절대적이고 바람직한 수준을 나타낸다. 예를 들어, 건강 관련 체력검사의 준거는 발병의 가능성을 최소화하여 건강한 생활을 유지할 수 있고 일상생활을 수행할 수 있는 최저 수준의 속성이나 기능을 나타낸다. 셋째, CRS는 한 개인의 상태나 현재의 수행능력이 적절한지에 관한 구체적이고 개별적인 정보를 제공해준다. 마지막으로, CRS는 각 개인을 완수자와 미수자로 분류하는 데에만 의미가 있다. 즉 NRT와는 달리 기준치 보다 높은 점수 또는 다른 학생들 보다 높은 점수가 보다 의미있는 것은 아니다.

CRS 설정과 이용에서의 문제점은 기준 설정이 자의적이고 분류오류에 따른 영

향이 심각하다는 것이다(Glass, 1978; Meskauskas, 1976; Millman, 1973). CRS를 설정하는 많은 방법들은 기준 결정의 과정이 주관적이라는 비난을 받아 왔다. 그러나 기준 설정의 과정이 주관적이지만 전문가의 판단에 근거한 기준은 사전적 의미에서 이유없이 무작위로 선택되었다는 뜻보다는 긍정적인 의미에서 판단에 의해 결정된다는 측면으로 생각해야 한다는 지적도 있다(Pohpam, 1978).

기준점수가 정해진 후에는 이 기준에 따라 피검자들을 분류함에 따라 필연적으로 분류오류가 발생한다. 誤完遂者와 誤未遂者의 형태로 나타나는 두 가지 분류오류로 인한 결과는 졸업 여부에 관한 준거나 인명구조 자격에 관한 준거 등 특수한 상황에서는 심각한 위험을 초래할 가능성이 있기 때문에 기준 설정시에는 의사결정의 형태와 의사결정이 개인에게 미치게 될 예측되는 영향을 반드시 고려해야만 한다(신승윤, 1994). 한 예로, 심동적 영역의 주요 주제 중의 하나인 체력평가에서 분류오류로 인한 결과는 다음과 같이 가정할 수 있다(Cureton, 1990). 첫째, 실제로는 적절한 체력수준을 가지고 있으나 최저 한도의 체력을 가지고 있지 않다고 잘못 분류되는 오미수자의 경우이다. 이 경우 보다 적극적인 신체활동을 유도할 수도 있다. 그러나 자신의 체력에 비해 목표성취가 너무 어렵다고 판단하거나 신체활동에 더 많은 노력을 기울일 가치가 없다고 생각한다면 더 많은 신체활동을 기대하기는 어렵다. 둘째, 실제로는 적절한 체력수준에 도달해 있지 않으나 적절한 체력을 가지고 있다고 분류되는 오완수자의 경우이다. 이 경우 발생할 수 있는 결과는 바람직한 수준 이하에 신체활동 수준을 한정함으로서 결과적으로 질병에 걸릴 확률을 증가시킨다. 분류오류의 가능성은 반복측정과 2개 이상의 기준점수 설정, 그리고 타당성의 검증에 따른 기준값의 적절한 수정으로 감소시킬 수 있다(Novick과 Lindley, 1978; Safrit 등, 1980).

체력평가에서 CRT에 대한 또 다른 우려는 이 기준이 바람직한 체력의 최저 수준을 나타내므로 보다 높은 수준의 체력을 성취하기 위한 자극을 주지 못한다는 것이다(Safrit, 1988). 그러나 CRT의 사용 목적에서도 언급된 바와 같이 보다 높은 체력 수준이 요구되는 사람들은 경쟁적인 스포츠 활동에 참가하는 운동선수들이다. 그러므로 일반 학생들이나 일반인들이 바람직한 수준 보다 더 높은 체력을 갖는다 해서 반드시 더욱 건강해지는 것은 아니기 때문에 체력평가에서 NRS보다는 CRS가 사용되어야 한다.

Ⅳ. 준거지향기준의 설정 방법

교육측정 분야에서 개발된 CRS의 설정 방법은 다양하다(Glass, 1978; Hambleton과 Powell, 1983; Shepard, 1984; Berk, 1986). Hambleton(1980)은 지금까지 제시된 기준 설정 방법들을 크게 판단적, 경험적, 그리고 혼합적 방법의 세 가지 범주로 구분하였으며, 이는 판단모형, 경험모형, 혼합모형으로 구분되기도 한다(신승윤, 1994). 判斷模型은 기준설정을 위한 근거를 전문가의 판단에 의존하는 방법을 말하며, 經驗模型은 기준설정 과정에 경험적 자료를 주로 활용하는 방법을 의미하고, 混合模型은 전문가의 판단과 경험적 자료를 함께 사용하여 기준을 설정하는 방법이다. 심동적 영역에서 CRS를 설정하는 방법은 판단적 방법, 규준적 방법, 경험적 방법, 그리고 이들 세 가지를 혼합한 방법이 주로 사용된다(Cureton과 Warren, 1990). 이러한 분류 방법은 Hambleton(1980)의 분류에서 규준적 방법을 세분화하여 제시한 것으로 나머지 방법들간에는 별 차이가 없다(강상조, 1994).

判斷的 方法(judgemental approach)은 전문가의 판단에 의존하여 기준을 설정하는 방법으로 지금까지 심동적 영역에서 가장 많이 이용되어왔다. 1970년대 초반까지는 심동영역에서의 준거지향검사 설정에 전문가의 경험과 식견에 기초한 판단에 의존하는 판단적 방법을 주로 사용하여 왔다(Safrit, Baumgartner, Jackson, 그리고 Stamm, 1980). 즉 전통적으로 심동적 영역의 CRS를 설정할 때 체육 전문가나 체육 교사들에게 특정 영역을 성취한 사람이라면 최소한 어느 정도의 점수를 획득할 수 있을 것인지를 묻고 이들의 평균을 CRS로 설정하는 방법이다. 교육측정 분야에서 개발된 판단적 방법에는 Nedelsky 모형, Angoff 모형, Jaeger 모형, Ebel 모형, Rasch 모형 등이 있다(Nedelsky, 1954; Angoff, 1971; Jaeger, 1978; Ebel, 1972; Rasch, 1968). 이들 모형의 공통점은 기준 설정시 최소 능력을 가진 피검자를 판단하고 이들이 맞힌 문항을 고려하여 기준을 설정하거나 問項分析을 통하여 최소한의 기준을 설정한다는 점이다.

판단적 방법은 전통적으로 심동적 영역에서 CRS를 설정하기 위하여 이용되어온 방법이나 전문가의 수나 성향 등에 따라 그 기준이 달라질 수 있기 때문에 임의적이며 타당성이 결여된다는 지적을 받고 있다. 교육측정 분야에서 판단적 방법으로 사용된 모형의 대부분은 다문항 일회 측정이라는 인지적 영역의 문항검사에 적합한 이론으로 심동적 영역의 적용에는 어려움이 있으나 보다 신뢰로운 판단을 위해 적용 가능성을 모색해볼 여지가 있다. 특히 Angoff 모형은 선다형 문항이 아닌 다른 유형의 검사에서도 기준설정이 가능하고 문항반응이론에 의한 방법은 집단 분류의 오류를 최소화하여 안정성(stability)을 보장하므로 심동적영역에 적용 가능성

이 크다.

規準的 方法(normative approch)은 검사점수와 타당한 준거를 서로 연결시킬 수 있는 과학적 근거 자료가 없는 경우에 특정 검사의 CRS를 설정하기 위하여 기존의 NRS를 사용하여 기준점을 설정하는 방법이다(강상조, 1994). 심동적 영역에서 규준적 방법을 이용하여 설정된 기준은 대부분 19~50 백분위 점수에 위치하고 있다(강상조, 1994; 신승윤, 1994). 예를 들어, 미국에서 널리 이용되는 체력검사의 하나인 FITNESSGRAM의 준거지향 체력기준도 남자 75%, 여자 50% 정도가 성취하는 것으로 알려졌다(Looney와 Plowman, 1990). 또한 강민수(1997)와 김종명(1997)은 상완근지구력과 유연성의 준거지향검사을 설정하기 위하여 30, 40, 50백분위점수를 이용하였다. 그러나 규준적 방법에 의해서 설정된 CRS는 준거와 관련된 경험적인 증거가 없기 때문에 판단적 방법과 마찬가지로 임의적인 기준에 불과하다. 經驗的 方法(empirical approach)은 경험적으로 수집된 자료를 근거로 기준을 설정하는 방법이다. 1970년대 후반부터 교육측정 분야에서 CRS 설정을 위한 다양한 경험적 방법이 개발되면서(Berk, 1976; Hambleton과 Novick, 1973; Livingston과 Zieky, 1982) 심동영역의 준거지향검사 설정방법에 경험적 방법을 적용하기 시작하였다(Douglass와 Safrit, 1983; Safrit, Wood, Ehlert, Hooper, 그리고 Patterson, 1985; Shifflett와 Schuman, 1982). 또한 미국의 각종 건강 관련 체력검사에서 CRS의 타당도와 신뢰도를 제고하기 위하여 경험적 자료를 활용하여 다양한 경험적 방법이 개발되었다(AAHPERD, 1988; American Health와 Fitness Foundation, 1986; Institute for Aerobics Research, 1987).

심동적 영역에서 주로 활용되는 경험적 방법은 準據集團模型이다. 준거집단모형은 Berk(1976)가 개발한 방법으로 준거기준이 알려져 있는 경우에 準據檢査와 決定檢査를 실시하여 기준을 설정하는 방법이다. 즉, 준거검사에서 준거기준의 도달여부에 따라 피검자를 준거도달 집단과 준거미달 집단으로 나누고 두 집단의 결정검사 점수분포 곡선의 교차점을 기준으로 설정하는 방법이다(Berk, 1976; Livingston과 Zieky, 1982; Popham, 1990). 그러므로 준거집단모형을 적용하기 위해서는 準據測定(criterion measure)을 실시하여야 한다. 준거측정이란 잠재적 특성(latent trait)의 이상적인 준거로서 과학적인 방법에 의하여 실험실 상황에서 측정된다. 이와 함께 실용적인 현장검사를 이용하여 실험실 상황에서 얻은 준거측정과 비교하여 가장 타당도가 높은 기준을 설정한다. 그러나 심동적 영역에서는 이와같은 준거가 이미 제시되어 있는 경우가 많지 않고, 또한 준거측정이 불가능한 상황이 많다는 단점이 따른다. 이와 같이 준거측정이 어려운 경우 단순히 피검자를 심동적 영역의 특정 프로그램의 이수 여부나 건강과 관련하여 발병인자의 보유 여부 등에 따라 준거분류하고 두 집단의 동일 영역 관련 검사점수의 분포를 예측분류에 사용하므로써 기

준을 설정하기도 한다.

混合的 方法(mixed approach)은 판단적 방법, 규준적 방법, 경험적 방법을 모두 종합하여 기준을 설정하는 방법으로 준거분류는 판단적 방법에 의존하고 현장검사를 통한 예측분류는 경험적 자료에 의존하여 合致度가 높은 점수를 기준으로 설정하는 방법이다. 심동적 영역에서 주로 사용되는 혼합적 방법은 對比集團模型(Nedelsky, 1954), 境界集團模型(Livingston과 Zieky, 1982; Nedelsky, 1954) 등이 있으나 단순히 판단적 방법, 규준적 방법, 경험적 방법에 의해 설정된 기준의 산술평균으로 설정하기도 한다.

대비집단모형은 평가자의 판단에 의해 피검자를 완수자집단과 미수자집단으로 분류하여 두 집단의 검사점수분포의 교차점을 기준으로 설정하는 방법이다(Livingston과 Zieky, 1982; Popham, 1990; Wiersma와 Jurs, 1990). 그러므로 이 방법의 타당도는 평가자가 피검자 개개인을 얼마나 적절하게 분류하느냐에 달려 있다. 이 방법의 단점은 피검자 집단을 숙달 학습자와 미숙달 학습자로 분류할 때 두 집단의 능력 특성에 따라 기준점수가 변한다는 점과 기준점수가 불안정하다는 점이다. 즉, 미숙달 학습자로 분류된 상위 득점 피검자들에 의해 높아지거나, 숙달 학습자로 분류된 하위 득점 피검자들로 인하여 낮아질 수 있다. 그리고 숙달 학습자로 분류되었으나 기준점수에 미달하는 경우와 미숙달 학습자로 분류되었지만 기준점수를 초과하는 경우의 두 가지 분류오류가 발생할 수 있다. 그러므로 검사의 특성에 따라 검사자는 부정적 분류오류와 긍정적 분류오류의 상대적 감소 여부를 결정하여야 한다. 또한 대비집단모형에서 대비가 되는 각 집단은 최소 100명 이상이어야 한다(Popham, 1990).

경계집단모형은 대비집단모형에서는 분류된 두 집단의 점수분포에 따라 기준점수가 변화한다는 단점을 보완하기 위해서 Zieky와 Livingston(1982)이 제안한 방법이다. 즉, 완수자나 미수자로 분류되지 않은 집단의 중앙값을 기준으로 설정하는 방법이다(Livingston과 Zieky, 1982; Norcini와 Shea, 1992; Wiersma와 Jurs, 1990). 이 방법의 장점은 경계집단의 중앙치를 기준점수로 설정하므로 기준점수가 미숙달 학습자집단의 점수분포에 의해 영향을 받지 않는다는 점이다. 그러나 경계집단의 정의가 임의적이고 기준설정을 위한 피검자 집단에 따라 기준이 달라진다는 단점이 있다. 즉, 경계집단모형은 사용과 해석이 용이하나 일반적으로, 경계집단에 속하는 피검자는 소수인 경우가 많고 경계집단에 있는 피검자를 한정하기가 쉽지 않다(Zieky와 Livingston, 1977).

심동적 영역에서의 준거지향검사 설정은 전통적으로 전문가들의 학식과 경륜 또는 현장 교사들의 경험에 기초한 판단에 의하여 이루어져 왔으나 대부분의 경우 관련 영역에 대한 피검자의 규준자료나 사전 연구의 자료들을 함께 사용하게 된다.

이러한 기준설정은 혼합적 방법으로 간주된다(신승윤, 1994).

V. 준거지향기준의 평가

심동적 영역의 검사점수는 검사목적에 따라 피검자의 학습목표 도달여부나 건강수준 등에 대한 정보를 제공해줄 수 있어야 한다. 즉 타당하고 신뢰로운 CRS를 통하여 피검자의 학습목표 도달여부나 건강 상태에 대한 해석이 가능할 때 검사점수의 활용가치가 제고된다. 예를 들어, 건강 관련 체력검사의 경우 높은 수준의 유산소 능력을 가진 사람은 심장질환에 걸릴 가능성이 상대적으로 낮다는 경험적인 증거에 기초하여 건강 관련 체력검사의 점수를 해석할 수 있는 CRS를 개발하여야 한다. 또는 고등학교 1학년의 정상적인 발달단계에 따른 근지구력의 CRS가 합리적으로 설정되었다면 팔굽혀펴기검사의 점수로 근지구력을 평가할 수 있다. 그러므로 CRS의 평가는 준거를 타당하게 반영하고 있는가와 준거가 신뢰롭게 설정되었는지에 대한 분석이다.

1. 타당도

CRS의 타당도는 얼마나 정확하게 분류하였느냐로 정의된다(Safrit, 1989). 예를 들어 경험적인 증거에 의하여 건강한 심폐지구력의 준거가 일정한 유산소 능력으로 제시되었을 때, 오래달리기검사의 CRS에 의한 분류와 준거에 의한 분류의 일치 정도가 오래달리기검사의 CRS에 대한 타당도를 나타내는 지수이다. 이는 分類正確確率(probability of correct decision)로서 진완수자와 진미수자의 비율을 더한 값으로 표시되며 우연에 의해 획득될 수 있는 값을 포함하여 .50에서 1.00의 범위를 갖는다.

CRS의 타당도를 나타내는 두 번째 방법은 이들 두 분류간의 상관계수를 구하는 것이다. 즉, 이원이분 유목변인 사이의 상관계수(ϕ)로 타당도를 추정할 수 있다. 세 번째 방법은 有用度(utility)를 분석하는 것이다. 이는 분류오류의 역기능에 대한 상대적 중요성을 감안하여 진완수자와 진미수자에 서로 다른 가중치를 부여하여 분류정확확률을 구하는 것이다. 건강한 심폐지구력을 가진 사람을 건강하지 못한 심폐지구력을 가진 사람으로 잘못 분류한 것을 뜻하는 오완수자 혹은 2종 오류보다는 건강하지 못한 심폐지구력을 가진 사람을 건강한 심폐지구력을 가진 사람으로 잘못 분류했을 때를 뜻하는 오미수자 또는 1종 오류가 더욱 심각한 결과를 초래할 수 있으므로 진완수자에는 1의 가중치를 부여하고 진미수자에는 2의 가중치

를 부여하여 분류정확확률을 구하는 것이 검사의 활용가치를 고려한 타당도 분석의 한 방법이다.

이들 세 가지 분석법을 비교하여 CRS의 타당도를 평가할 수 있다(Berk, 1976). 분류정확확률, 상관계수, 그리고 유용도를 비교하여 같은 기준점에서 이들 모두 최대가 된다면 이 기준의 타당도는 높은 것으로 평가할 수 있다. 그러나 세 가지 지수를 최대로하는 기준점이 각각 달리 설정된다면 검사의 목적, 검사기준, 자료의 오류 등, 검사의 특성을 종합적으로 고려하여 CRS의 타당도를 평가해야 한다.

2. 신뢰도

CRS의 신뢰도는 분류의 一貫性으로 정의된다(Looney, 1989). 즉, 다른 시기에 검사가 실시되었다 하더라도 피험자들이 유사하게 분류되어야 신뢰도가 높은 검사라 할 것이다. 나아가 CRS의 신뢰도 개념은 객관도로 확장될 수 있다. 즉, 두 사람 이상의 평가자간의 합치도나 한 사람의 평가자가 두 번 이상의 반복 측정했을 때의 일치정도인 평가자 객관도를 측정하므로서 CRS의 신뢰도를 평가할 수 있다.

CRS의 신뢰도를 나타내는 대표적인 통계치는 合致度(P)로서, 첫 번째 검사와 두 번째 검사에서 모두 완수자와 미수자로 분류된 비율의 합인 P계수를 계산함으로서 추정될 수 있다. P계수는 사용이 간편하나 우연에 의해 첫 번째와 두 번째 검사에서 완수자와 미수자로 분류될 수 있는 확률이 포함되는 단점도 있다. P계수의 범위는 우연에 의하여 분류될 최대 확률인 .50을 제외한 .50에서부터 1.00까지이다.

신뢰도를 나타내는 또 다른 통계치로서 Kappa를 이용한다. p계수에 비하여 Kappa계수는 두 번의 검사에서 우연에 의하여 일치된 분류결과를 제거하고 순수하게 분류가 일치된 정도를 계산해주는 통계치이므로 .00에서 1.00까지의 전 범위가 해석가능하다. 대개의 경우 CRS의 신뢰도를 보다 효과적으로 평가하기 위해서 P계수와 Kappa계수를 함께 제시해준다.

Ⅵ. 결론 및 제언

체육 분야에서도 CRT의 필요성이 고조되면서 현장에서 활용가능한 CRS의 개발에 관한 연구가 활발해지고 있다. 심동적 영역에서 CRS는 운동능력이나 체력수준 등과 관련된 교육목표를 의미있게 평가하기 위함이다. 한 예로 한 학생이 획득한 특정 체력요소에 대한 검사 점수가 CRS 이상일 경우 그 학생은 발단단계에서 필요한 최소한의 건강 상태나 체력 수준에 도달한 것으로 판정된다.

선행연구의 고찰을 통하여 CRS의 특성, CRS의 설정, 그리고 CRS의 평가 요령을 살펴보았다. 이를 통하여 학교 체육 현장에서 필요한 CRS의 작성이 이루어져야 하며 그 과정을 요약하면 다음과 같다.

첫째, 현장 검사와 타당하게 관련된 준거(criteria)를 결정한다.

둘째, 준거에 따라 집단을 분류하고 CRS를 설정한다.

셋째, CRS의 타당도 및 신뢰도를 구한다.

교육적 장점에도 불구하고 학교 체육 현장에서 보편화되지 못하는 CRT의 개발을 위해서 다음의 연구들이 수행되어야 한다.

첫째, 관련된 준거를 규명하기 위하여 충분한 경험적 연구가 필요하다.

둘째, 교육과정상에 발달단계별 도달 목표가 구체적으로 제시되어야 한다.

셋째, 목표 도달 여부의 평가를 위하여 국가수준의 CRS 설정 연구가 필요하다.

참 고 문 헌

강민수(1997). 상완 근지구력 검사의 준거지향기준 설정에 관한 연구, 서울대학교 석사학위논문.

강상조 등(1987). 국민체력평가 기초연구, 체육부.

강상조(1994). 건강관련 체력검사의 준거지향 기준설정, 한국체육대학교 체육과학연구소.

김석우, 윤명희, 지은림(1997). 준거지향검사의 개념 및 기준설정 방안. 한국교육평가학회 학술세미나 발표논문집, pp. 1-22.

김종명(1997). 유연성 검사의 준거지향기준 설정에 관한 연구, 서울대학교 석사학위논문.

박광배(1992). 절대평가 기준과 상대평가 기준, 한국심리학회(편), 심리검사 제작의 이론과 실제(pp. 131-163), 한국심리학회 동계연수회 자료.

신세호(1992). 교육의 본질을 위한 학교교육평가체제 연구(Ⅲ). 서울: 배영사.

신승윤(1989). 학교체육평가의 제문제, 서울대학교 체육연구소논집, 10(1), 49-57.

신승윤(1994). 준거지향검사의 기준설정방식 비교, 서울대학교 박사학위논문.

조기정(1996). 농구자유투검사에서 축차확률비검사의 유용성 검증. 서울대학교 박사학위논문.

AAHPERD(1980). AAHPERD Health-related physical fitness test manual. Reston, Va: American Alliance for Health, Physical Education, Recreation and Dance.

AAHPERD(1988). The AAHPERD Physical Best Program. Reston, Va: American Alliance for Health, Physical Education, Recreation and Dance.

American Health and Fitness Foundation.(1986). FYT program manual. Austin, TX: Author.

Angoff, W. H.(1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational measurement(2nd ed., pp. 508-600). Washington, D.C.: American Council on Education.

Baumgartner, T. A., & Jackson, A. S.(1987). Measurement for evaluation in physical education and exercise science (2nd ed.). Dubuque, IA: Wm. C. Brown.

Berk, R. A.(1976). Determination of optimal cutting scores in criterion-referenced measurement. The Journal of Experimental Education, 45, 4-9.

Berk, R. A.(1986). A consumer's guide to setting performance standards on criterion-referenced tests. Review of Educational Research, 56, 137-172.

Blair, S. N., Clark, D. G., Cureton, K. J., & Powell, K. J.(1988). Exercise and Fitness in Childhood: Implication for a Lifetime of Health. In C. V. Gisolfi & D. L. Lamb(Eds.), Perspectives in exercise science and sports medicine: Youth, exercise and sport. Indianapolis: Benchmark Press.

Blair, S. N., Falls, H. B., & Pate, R. R.(1983). A new physical fitness test. The physician and Sport Medicine, 11, 87-91

Cureton, K. J., & Warren, G. L.(1990). Criterion-Referenced Standards for Youth Health-Related Fitness Test: A tutorial. Research Quarterly for Exercise and Sport, 61, 7-19.

Douglass, J. A., & Safrit, M. J.(1983). An empirical approach to the validation of a criterion-referenced measure of motor performance. Journal of Human Movement Sudies, 9, 57-69.

Ebel, R. L.(1972). Essentials of educational measurement(2nd ed.), Englewood Cliffs, NJ: Prentice-Hall.

Glaser, R.(1963). Instructional technoloy and the measurement of learning outcomes: Some questions. American Psycholgist, 18, 519-521.

Glass, G. V.(1978). Standards and criteria. Journal of Educational Measurement, 8, 321-326.

Grosse, M. E., & Wright, B. D.(1993). How to set standards. Rasch Measurement Transaction of the Rasch Measurement SIG, American Educational Research Association, 7(3), 315-316.

Hambleton, R. K. & Novick, M. R.(1973). Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 10, 159-170.

Hambleton, R. K.(1980). Test score validity and standard-setting methods. In R.A. Berk(Ed.), Criterion-references measurement: The state of the art. 80-123. Baltimore, MD: Johns Hopkins University Press.

Hambleton, R. K.(1994). Criterion-referenced measurement. In T. Husen, & T. N. Postlethwaite(Ed.), The internationna encyclopedia of education, Kidlington, Oxford: Elsevier Science Ltd, 2, 1183-1189.

Hambleton, R. K., & Powell, S.(1983). A framework for viewing the process of standard-setting. Evaluation and Health Professions, 6, 3-24.

Hensley, L. D., Lambert, L. T., Baumgartner, T. A., & Stillwell, J. L.(1987). Is evaluation worth the effort? Journal of Physical Education, Recreation and Dance, 58(8), 59-62.

Hughes, F. P., Schumacher, C. F., & Wright, B. D.(1984). Estimating criterion referenced standards for multiple choice examinations. National Board of Medical Examiners: Philadelphia, Pennsylvania.

Institute for Aerobics Research(1987). FITNESSGRAM user's manual. Dallas, TX: Author.

Kalohn, J. C., Wagoner, K., Gao, L., Safrit, M. J., & Getchell, N.(1992). A comparison of two criterion-referenced standard setting procedures for sport skills testing. Research Quarterly for Exercise and Sport, 63(1), 1-10.

Livingstone, S. A., & Zieky, M. J. (1982). Passing scores: Amanual for setting standards of performance on educational and occupational tests. Princeton, NJ: Educational Testing Service.

Looney, M. A., & Plowman, S. A.(1990). Passing Rates of American Children and Youth on the FITNESSGRAM Criterion-Referenced Physical Standards. Research Quarterly for Exercise and Sport, 61(3), 215-223.

Looney, M. A. (1987). Threshold loss agreement indices for criterion-referenced measures: A review of applications and interpretations. Research

Ouarterly for Exercise and Sport, 58, 360-368.

Looney, M. A.(1989). Criterion referenced measurement: Reliability. In M. J. Safrit & T. M. Wood(Eds.), Measurement concepts in physical education and exercise science. Champaign, IL: Human Kinetics.

Meskauskas, J. A.(1976). Evaluation models for criterion-referenced testing: Views regarding mastery and standard setting, Review of Educational Research, 46, 133-158.

Millman, J.(1973). Passing scores and test lengths for domain-referenced measurements. Review of Educational Research, 43, 205-215.

Nedelsky, L.(1954). Absolute grading for objective tests. Educational and Psychological Measurement, 14, 3-19.

Nixon, J. E., & Jewett, A. E.(1980). An Introduction to Physical Education. Philadelphia: Saunders College Publishing/Holt, Rinehart and Winston.

Norcini, J. J., & Shea, J. A.(1992). Eouivalent estimates of borderline group performance in standard. Journal of Educational Measurement, 29(1), 19-24.

Novick, M. R., & Lindley, D. V.(1978). The use of more realistic utility functions in educational applications. Journal of Educational Measurement, 15, 181-191.

Pate, R. R.(1983). A New Definition of Youth Fitness. The Physician and Sports Medicine.

Pate, R. R.(1983). South Carolina physical fitness test manual(2nd ED.). Columbia : South Carolina Association for Health Physical Education, Recreation and Dance.

Plowman, S. A.(1992). Criterion referenced standards for neuromuscular physical fitness test: An analysis. Pediatric Exercise Science. 4, 10-19.

Plowman, S. A., & Corbin, C. B.(1994). Muscular strength, endurance, and flexibility. In J. R. Morrow, H. B. Falls, & H. W. Kohl(Eds.), The Prudential FITNESSGRAM technical reference manual(pp. 73-95). Dallas, TX: Cooper Institude for Aerobics Research.

Popham, W. J.(1978). Criterion-referenced measurement. Englewood Cliff, NJ: Prentice Hall.

Popham, W. J.(1990). Modern educational measurement: A pratitioner's perspective. Englewood Cliffs, NJ: Prentice-Hall.

Popham, W. J., & Husek, T. R.(1969). Implications of criterion-referenced measurement. Journal of Educational Measurement, 6, 1-9.

Rasch, G.(1968). A methmetical theory of objectivity and its consequences for model contribution. European Meeting on Statistics, Econometrics, and Management Science, Amsterdam.

Rutherford, W. J., & Corbin, C. B.(1994). Validation of criterion-referenced standards for tests of arm and shoulder girdle strength and endurance. Research Quarterly for Exercise and Sport, 65(2), 110-119.

Safrit, M. J.(1968). Comparison of Four Factor Analysis. Evanston: Northwestern University Press.

Safrit, M. J.(1986). Introduction to measurement in physical education and exercise science. St. Louis: Mosby.

Safrit, M. J.(1988). Methods of measurement of physical fitness in children and youth. Paper presented at a consensus meeting on the measurement of physical fitness sponsored by the public Health Service, Washington, D.C.

Safrit, M. J., & Wood, T. M.(1989). Measurement Concepts in Physical Education and Exercise Science. Illinois, Human Kinetics Books.

Safrit, M. J., Baumgartner, T. A., Jackson, A. S., & Stamm, C. L.(1980). Issues in setting motor performance standards. Quest, 32, 152-163.

Safrit, M. J., Wood, T. M., Ehlert, S. A., Hooper, L. M., & Patterson, P.(1985). The application of sequential probability ratio testing to a test of motor skill. Research Quarterly for Exercise and Sport, 56, 58-65.

Safrit, M. J.(1981). Evaluation in physical education(2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Safrit, M. J.(1989). Criterion referenced measurement: Validity. In M. J. Safrit, & T. M. Wood(Eds.), Measurement concepts in physical education and exercise science(pp. 119-136), Champaign, IL. : Human Kinetics.

Safrit, M. J., & Stamm, C. L.(1980). The reliability of criterion referenced measures of motor behavior: A comparative study. Research Quarterly for Exercise and Sport, 51, 359-368.

Safrit, M. J., & Wood, T. M.(1995). Introduction to Measurement in physical education and exercise science(3rd ed.). St Louis: Mosby.

Sharkey, B. J.(1977). Fitness and work capacity. Washington, D.C. : U.S.

Department of Agriculture Publication. FS-315.

Shepard, L. A.(1984). Setting performance standards. In R. A. Berk(Ed.), A guide to criterion-referenced test construction. Baltimore, MN: The Johns Hopkins University press.

Shifflett, B., & Schuman, B. J.(1982). A criterion-referenced test for archery. Research Quarterly for Exercise and Sport, 53, 330-335.

Smith, R.(1997). Validating standard setting with a modified Nedelsky procedure through common item equating, Journal of Outcome Measurement, 1(2), 164-172.

Stone, G. E.(1996). The construction of meaning: Replicating objectively derived criterion-referenced standards. Paper presented at the annual meeting of the American Educational Research Association at New York.

W.H.O.(1967). Exercise test in relation to cardiovascular function. Report of a W.H.O. meeting, Genora.

Washburn, R., & Safrit, M. J.(1982). Physical performance tests in job selection: A model for empirical validation. Research Quarterly for Exercise and Sport, 53(3), 267-270.

Wiersma, W., & Jurs, S. G.(1990). Educational measurement and testing. Needham Heights, MA: Allyn and Bacon.

Wood, T. M.(1995). Setting Criterion Referenced Standards for Physical Fitness Test: The New FITNESSGRAM, Oregon State University.

Zieky, M. J., & Livingston, S. A.(1977). Manual for setting standards on the basic skills achievement tests. Princeton, NJ: Educational Testing Service.

<Abstract>

Methods for establishing CRS in the Psychomotor Domain

Kim, Jong-Taek

One of current issues in evaluation is with regard to the distinctions between, and relative advantages and disadvantages of norm-referenced and criterion-referenced measurement. However, criterion-referenced measurement is more emphasized with changing the educational circumstances in which a predeterminded standard of performance is used rather than a normative standard.

Criterion referenced standard(CRS) represents desired levels of performance or status on a criterion domain or attribute, and provides diagnostic information about whether status or performance is adequate. The use of CRS in testing is to categorize students into master or nonmaster based on the CRS or the cut-off score. The problems of CRS are that they are arbitrary and that the consequences of misclassifications.

A variety of methods for establishing CRS have been developed. Mehtods that are applicable in the psychomotor domain of physical education are judgemental, normative, empirical, and combination methods. The judgemental mehtod is to establish CRS by the experts' opinion. The normative method is to use normative data along with the information to setting the CRS. A CRS with empirical method is to base on empirical data of predeterminded master and nonmaster. The combination method is to use the combination of the above methods.

An important aspect of developing CRSs is establishing their reliability and validity. The reliability and the validity of CRS are defined as the consistency

and accuracy of classification, master or nonmaster. One type of reliability of CRS is rater reliability that is the same concept of objectivity. And P and Kappa coefficients are used as the indices of the CRS reliability in test-retest apporach. The validity of CRS is evaluated based on a probability of correct decision, Phi coefficient, and a utility.