일반화가능도 이론과 FACETS 모형을 적용한 채점모형 개발 연구

- 음악 실기평가를 중심으로 -

김성숙 · 채선희
(한국교육과정평가원)

I. 서 론

최근 교육현장에서는 중요한 의사결정 상황에서 특정검사의 일회성 점수에만 의존함으로써 발생할 수 있는 문제들을 예방하면서, 검사점수의 제한적이고 단편적 표현으로 인한 문제점을 해소하기 위한 방안으로 수행평가의 활용이 적극적으로 권장되고 있다 (한국교육평가연구회, 1997). 그러나 수행평가가 학교현장에서 보편적으로 활용되기에는 아직도 해결되지 못한 많은 문제점을 내포한다. 그 중 우선적으로 고려해야할 과제가 바로 채점에 대한 객관성 확보의 문제이다. 평가대상에 대한 타당하고 신뢰로운 채점이 보장되지 않는한 수행평가는 불완전한 평가방법에 그치고 말 것이기 때문이다.

그동안 수행평가 채점의 타당도와 신뢰도를 높이기 위한 다양한 연구들이 지속적으로 이루어져 왔다 (배호순, 1997; 채선희, 1997; 김성숙, 1995; Baxter, Shavelson, Goleman & Pine, 1993; Koretz, Stecher, Klein, & McCaffrey, 1994; Shavelson, Baxter & Gao, 1993). 본 연구는 지필검사만으로 학생을 평가하는데 한계가 있고, 지속적인 관찰과 실기수행에 대한 피드백이 권장되는 분야의 하나인 음악 교과에 있어서 '기악연주'에 대해 실기평가를 시행한 자료를 수집하여 보다 합리적이고 공정한 채점 모형을 개발하는데 그 목적이 있다. 특히 본 연구는 최근 들어 수행평가 결과점수에 대한 측정상황을 분석하기 위해 적용되는 이론 중 가장 대표적인 일반화가능도이론과 문항반응이론을 적용한다는 점에서 그 독특성을 지닌다. 우선 일반화가능도 이론(또는 G 이론으로 칭함)은 고전검사이론에서 설명하지 못한 측정의 중다요인 오차(multiple sources of error)를 측정상황의 오차국면(facet)에 근거하여 각각 분해하여 설명하고 있다 (Cronbach, 1972; Brennan, 1983). G 이론에서 관찰된 단일 행위는 관찰행동 전집을 대표하는 표집으로 보고 이러한 표집행위가 얼마나 일반화되어질 수 있는가 하는 일반화가능도를 추정한다. 특히 일반화연구(G study)는 각 피험자 점수 또는 관찰점수를 구성하는 허용가능한 관찰의 전집(universe of admissable observation)과 관계된 조건의 집합들의 분산성분추정값을 산출하는 과정이다. 오차분산성분의 상대적 크기를 비교하여 각 오차원의 영향력을 분석함으로써 어떤 요인이 측정의 일반화과정을 저해하는지 설명할 수 있고, 결정연구(D study) 설계의 조건을 결정하는 근거로 활용한다. 결정연구에서 각 오차의 상대적 영향력에 따라 일정 수준의 신뢰도를 보장하기 위한 측정조건을 제시한다. 그러나, 일반화연구에서 산출하는 각 국면의 분산성분은 피험자 원점수에 관한 정보를 제공하는데 그치며 특정 오차국면, 즉 채점자, 문항, 또는 과제 등을 고려하여 재조정된 피험자 점수를 산출하지는 않는다.

한편 Linacre(1993)는 문항반응이론 가운데 피험자의 능력과 과제의 곤란도만으로 피험자의 과제 수행정도를 확률적으로 추정하는 라쉬이론의 기본모형을 확장시켜, 기본공식에 채점자가 엄격한 정도를 포함시킴으로써, 과제가 곤란한 정도와 채점자가 엄격한 정도로부터 비교적 자유로운 피험자 능력 모수치를 추정해낼 수 있는 Facets 모형을 개발하였다. 최근 많은 연구들이 이러한 Facets 모형의 효율성을 보여주는

결과를 산출하고 있다. Linacre(1993)는 또한 G 연구의 결과로 제공되는 정보를 기본으로 Facets 모형을 적용하여 각 개인 점수에 영향을 미치는 오차국면의 효과를 제거한 후의 개인 점수를 추정함으로써 두 가지 분석방법의 접목을 유도하기도 하였다.

II. 연구 자료

음악 실기평가의 채점모형을 개발하기 위한 목적으로, 캐나다 왕립음악학교(Royal Conservatory of Music)에서 연중 2차례에 걸쳐 시행하는 기악 연주 평가의 결과를 사용하여 분석하였다. 구체적 연구자료는, 피아노를 연주한 15명의 학생에 대한 평가결과와 바이올린을 연주한 15명의 평가결과로 구성되었다. 평가 실시를 위하여 한 학생에 대하여 2명의 다른 채점자(전문평가사)가 각각 동원되었고, 6개월 간격으로 2번 시행한 평가의 결과 자료이다. 그리고, 학습자(연주자)는 평가 받은 이후 결과에 따라 다음 등급으로 오르게 된다. 따라서, 2회에 걸쳐 실시된 평가에 있어서 실기과제(곧, 평가항목)은 6가지로 동일하지만, 두 평가시기에 연주해야 할 곡(과제)이나 수행내용은 다르게 된다. 분석을 위한 자료의 구성과 배열은 다음 <표 1>과 같다.

한편, RCM의 실기평가 방법은 등급에 따라 한 학생(연주자)이 6가지 8가지 영역에 대하여 수행하도록 한다. 채점기록 방법은 각 영역에 대한 점수 부여와 함께 연주자 수행 특성에 대하여 상세하게 기술하도록 되어있다. 실제 채점표의 예는 <부록>에 첨부하였다. 이 연구에서 사용된 실기 평가 항목 6가지 영역 중 3가지는 학생이 실제로 연주해야 하는 곡에 대한 평가이고, 나머지 3영역은 각각 테크닉(technique) 중 음계(scale)영역, 그리고 청음(ear tests)과 시창(sight reading) 이다. 연주해야 할 연주곡은 학생이 속해있는 등급에 따라 바로크·고전주의, 낭만주의 그리고 현대음악 각각 영역에서 추천하는 여러곡 중 3곡을 선정하여 연습·연주하도록 되어있다. 실기평가에는 암보(memory)와 테크닉(technique)중 연습곡(studies)도 포함되어 있으나 이번 연구의 분석대상에서는 제외하였다. 자료의 점수는 연주에 대하여 바로크·고전 18점, 낭만주의 24점, 현대음악 12점, 그리고 테크닉(음계), 청음, 시창은 각각 10점으로 배점되어 있다.

<표 1> 실기 평가결과 자료 구성과 배열

변인	악 기	학생 (피험자)	평가 시기	평가 항목	채점자 (전문평가사)
구성	피아노	15명	2회 (4월,10월)	6영역	각 2명
	바이올린	15명	2회 (4월,10월)	6영역	각 2명

III. 일반화가능도 이론의 적용

음악, 체육, 미술 등 예체능 교과에서 실시해 온 실기평가를 포함한 수행평가 방법은 학생들이 무엇을 어떻게 배우고 실제로 무엇을 수행할 수 있는가를 교수·학습활동과 병행하여 평가할 수 있고, 결과에 대한 활용이 직접적이고 효율적일 수 있다. 이러한 강점에도 불구하고 평가과정의 공정성, 결과점수의 신뢰성, 한정된 표집활동, 그리고 시간과 노력의 소요 등의 해결해야 할 많은 과제를 안고 있다. 평가과정의 공정성과 결과점수의 신뢰성 문제는 평정의 오차를 최소화함으로써 극복해야 하며 그것은 구조화된 평가준거, 평가항목, 채점기준표, 체계적 절차 등을 개발해야 하는 또다른 숙제를 수반한다. 수행평가가 현장적용에서 대안적 평가로 자리매김하려면 평가방법으로서의 객관성 확보와 결과점수에 대한 신뢰도, 즉 평가자 간 그리고 평가자 내 일관성 문제의 중요성과 심각성을 부각시켜야 할 것이다. 평가결과에 대한 신뢰성과 측정상황의 오차를 동시에 설명할 수 있는 일반화가능도 이론을 적용함으로써 기악 실기평가의 최적 채점조건을 제시하고자 한다.

1. 일반화가능도 이론과 연구설계

피험자의 행동특성(학습성취와 정의적 특성)을 얼마나 안정적으로 일관성있게 정확하게 측정하는가 하는 고전검사이론 (classical test theory)의 신뢰도 개념만으로는 측정에서 발생할 수 있는 변동 요인에 대한 설명이 충분하지 않다. 고전검사이론의 신뢰도 계수는 진점수와 관찰점수의 상관관계로 정의되어진다. 그리고, 신뢰도 계수, 즉, 관찰 점수 분산에 대한 진점수 분산의 비율($\rho^2 = \sigma t^2 / \sigma x^2$)은 측정이 어느 정도 측정대상의 眞점수를 반영하는가 나타낸다. 그러나, 고전 측정 이론에서 신뢰도 추정의 이론적 제한은 측정에서 개인 眞점수(True score)를 고정시키고, 오차점수 분산을 단일차원으로만 설명함으로써, 측정 또는 검사의 목적에 따라 측정 대상의 全集(universe) 이 달라질 수 있다는 확장 가능성을 배제하고있다. 一般化可能度 理論 (Generalizability theory)은 1972년 理論 정립 이후 (Cronbach, et al.), Brennan(1983), Shavelson((1986, 1991) 등에 의하여 적절하게 활용되고 있으며, 여러 연구에서 관찰자료, 수행검사 또는 작문, 논술 등의 측정에서 주지하여야 할 오차要因을 복합적으로 분석하였다 (김성숙, 1992,1993; Colton, 1993; Thompson & Melancon, 1987; Shavelson, 1989).

실기평가는 평가상황에서 발생할 수 있는 오류의 종류가 다양하다. 어떤 개인이 받은 어떤 점수는 선택가능한 수많은 과제 중 하나를, 여러 경우 중 한 시기에, 여러 채점자 중 한 명에 의하여, 그리고 수많은 평가준거 중 하나에 준하여 평정되어진 점수이다. 이와 같이 학생이 평가되는 상황은 학생의 능력 또는 학생이 표현하는 행동특성에 직접 또는 간접적으로 영향을 미칠 수 있는 여러 가지 오류의 조건으로 구성되어진다. 이렇게 측정상황을 구성하는 국면, 즉 채점자, 시기, 평가준거, 평가항목, 수행과제, 방법 등이 학생의 점수를 설명할 수 있는 근거가 되는 것이다. 따라서, 본 연구는 일반화가능도 이론을 적용하여 기악 실기평가의 채점상황에서, 학생에 대한 관찰 점수를 형성하는 진점수와 오차점수의 구성원을 설명하고, 오차의 상대적 기여정도에 따라 점수의 일반화가능성을 어떻게 확보할 수 있는가를 살펴본다. 그리고, 일정 수준의 신뢰성을 보장하기 위하여 최적의 측정조건을 제시하고자 한다. 구체적인 분석과정으로 첫째, 검사 점수의 誤差源으로 지적할 수 있는 피험자, 채점자, 평가문항, 평가시기와 그들의 상호작용의 관계를 분석하는 G(일반화) 연구, 둘째, 각 국면(facet)의 오차분산성분의 상대적 크기에 따라 각 국면의 차원을 조정함으로 적정 수준의 일반화가능도 계수를 산출하는 D(결정) 연구를 수행한다. 일반화가능도 이론을 적용하여 연구대상 자료를 분석하기 위한 연구질문은 다음과 같다.

① 음악실기 평가의 채점과정에서, 피험자 점수에 영향을 미칠 수 있는 오차, 즉 평가 시기, 평가 항목(연주곡, 과제내용), 채점자와 각 효과들과의 상호작용의 분산성분 크기와 각 효과의 영향력은 어떻게 나타나는가?

② 분산성분의 상대적 크기에 따라, 각 오차국면(채점자, 평가시기, 평가영역(항목)) 의 차원(수)을 조절함으로 일반화가능도 계수는 어떤 추세로 개선되는가?

▶ **G 연구 p x [(r x i) : o] 설계와 D 연구 p x [(R x I) : O] 설계**

연구질문에 따라 다음과 같은 일반화가능도 모형을 설계하였다. 측정의 誤差요인으로 특정 평가시기 내에서 채점자, 평가항목, 피험자를 포함시킨 3국면 내재모형(nested design)이다. 측정 대상은 피험자(p; person)이고, 오차 국면은 평가시기(o: occasion), 채점자 (r; rater), 평가항목 (i; item) 이므로, G 연구 p x ((r x i) : o)설계이다. 즉, 각 평가실시 시기에서 내재된 채점자는 모든 평가항목에 대하여 채점하고, 학생의 연주는 동일 평가시기에서 모두 평가되므로, 2원 교차 1원 내재 모형이다. 단, 연주된 2 종류의 악기(피아노, 바이올린) 수행에 대하여 각각 분석하였다. 이 모형에서 피험자, 채점자, 평가시기, 평가항목 모두 무한히 가능한 전집 중에서 추출된 표본으로 간주 할 수 있으므로 임의 효과(random effect)로 보았다. 따라서, 각 오차국면이 임의 효과이므로 G연구 결과에 대하여 D연구에서 일반화가능한 국면의 설정이 제한되지 않음을 알 수 있다(김성숙, 1995).

G 연구 p x ((r x i) : o) 설계에서 한 학생의 관찰점수는 9 개의 독립적인 분산성분으로 설명된다. 따라서, 어떤 학생의 연주를 어떤 평가시기에 어떤 채점자가 어떤 평가항목을 기준으로 평가한 점수는 수식 (1)과 같은 분산성분으로 설명될 수 있다.

$$\sigma_x^2 = \sigma_p^2 + \sigma_o^2 + \sigma_{r:o}^2 + \sigma_{i:o}^2 + \sigma_{po}^2 + \sigma_{pr:o}^2 + \sigma_{pi:o}^2 + \sigma_{ri:o}^2 + \sigma_{pri:o,e}^2 \quad \cdots\cdots (1)$$

σ^2_p는 측정에서 나타난 피험자 간의 차이, σ^2_o는 평가실시 시기간 차이, $\sigma^2_{r:o}$는 시기내 채점자간의 차이, $\sigma^2_{i:o}$는 시기내 평가항목 점수에 대한 차이를 각각 의미하고, σ^2_{po}는 시기에 따라 피험자가 다르게 평가된 정도, $\sigma^2_{pr:o}$는 평가 시기내 채점자가 피험자를 다르게 채점하는 정도, $\sigma^2_{pi:o}$는 시기내 피험자가 평가항목에 따라 다르게 평가받는 정도, $\sigma^2_{ri:o}$는 평가시기내 채점자가 항목에 따라 다르게 반응하는 정도, 그리고, $\sigma^2_{pri:o,e}$는 잔차로 설명된다.

위에서 나타난 G 연구 결과는 D 연구 설계로 연결됨으로써 결정과정에 참여한다. 연구자는 각 분산성분의 상대적 크기에 따라 국면의 차원을 조정한다. 이 연구의 D연구 설계는 G연구 설계와 동일한 모형으로 일반화전집을 추리하였다. 측정 대상은 피험자이므로 전집 분산(universe score variance)은 피험자 분산성분이고(σ^2_p), 오차 점수는 상대적 결정과정으로 상대오차분산은 측정 대상과 관계되는 변동, 즉 피험자-평가시기 상호작용 효과(σ^2_{po}), 평가시기 내 피험자-채점자 상호작용 효과($\sigma^2_{pr:o}$), 평가시기내 피험자-평가항목 상호작용($\sigma^2_{pi:o}$), 그리고 잔차 ($\sigma^2_{pri:o}$)를 포함한다. 모든 국면이 임의효과이므로, 전집분산성분과 상대오차분산성분은 다음과 같다.

$$\hat{\sigma}(\tau)^2 = \sigma_p^2$$

$$\hat{\sigma}(\delta)^2 = \frac{\sigma_{po}^2}{n_o} + \frac{\sigma_{pr:o}^2}{n_r n_o} + \frac{\sigma_{pi:o}^2}{n_i n_o} + \frac{\sigma_{pri:o}^2}{n_r n_i n_o}$$

따라서, 일반화가능도 계수는 수식 (2)를 기준으로 산출된다.

$$\rho^2(\delta)^2 = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{(\delta)}^2}$$

$$= \frac{\sigma_p^2}{\sigma_p^2 + \left(\frac{\sigma_{po}^2}{n_o} + \frac{\sigma_{pr:o}^2}{n_r n_o} + \frac{\sigma_{pi:o}^2}{n_i n_o} + \frac{\sigma_{pri:o}^2}{n_r n_i n_o}\right)} \cdots(2)$$

산출된 일반화가능도 계수는 각 국면의 기본조건 ($n_o = n_r = n_i = 1$)에 대한 지수이므로, 각 국면의 조건 또는 수의 조합에 따라 오차분산과 계수의 변동을 계산하였다. 적정 수준의 일반화가능도 계수를 산출하기 위한 각 국면의 최적의 조건을 제시하였다.

2. 분석 결과

음악 실기자료에 대한 채점에서의 오차원과 상대적 영향력, 그리고 측정조건을 제안하기 위한 G 연구와 D 연구 분석 결과는 다음과 같다. 피아노를 연주한 피험자와 바이올린을 연주한 피험자 각각에 대하여 분석하였다.

가. G 연구 분석결과

피험자가 부여받은 점수에 미치는 영향을 분석하기 위하여 각 평가시기 내에서 두 채점자가 동일한 평가영역에 대하여 채점한 개개인 음악연주에 대한 실기평가 자료를 분석하였다. 피아노와 바이올린 연주의 G 연구 결과를 각각 <표 2>과 <표 3> 에 정리하였다. 전집 분산과 각 오차 分散成分의 크기를 상대적으

로 비교함으로써 변동요인의 영향력을 설명하고 있다. 피아노 연주와 바이올린 연주에 대한 G 연구 결과는 유사하게 나타났다. G 연구 결과를 보면, 평가 시기, 채점자, 평가항목 각 1 개의 차원을 기준으로, 피아노 연주와 바이올린 연주 각각의 전집점수 분산성분은 상대적으로 작다(.2663, 1.2%; .2722, 1.2%). 피험자 점수 분산은 측정대상이고 전집점수분산이므로 변동의 크기가 클수록 신뢰도 계수에 긍정적 영향을 미친다. 그러나, 상대적으로 다른 오차국면의 분산성분과 비교하여 그 의미를 찾을 수 있다. 그리고, 평가시기 내 문항의 분산성분이 상대적으로 크게 산출된 것은 (20.72, 93.12%; 21.58, 93.94%) 시기에 따라 연주해야 할 연주곡과 수행해야 하는 문항의 내용이 차이가 있음을 의미한다. 이것은 평가시기에 따라 학생이 배우는 내용이 향상되고 그에 따라 연주해야 할 연주곡과 수행해야 할 시창, 청음 등의 내용이 다르기 때문이라 볼 수 있다. 그러나, 평가시기내 피험자와 문항의 상호작용, 채점자와 문항의 상호작용의 분산이 0 에 가깝게 산출되었으므로 연주곡이나 수행해야 할 항목의 내용이 바뀔지라도 피험자나 채점자의 수행이 문항에 의하여 발생되는 변동은 거의 없는 것으로 나타났다. 반면, 평가시기내 채점자(.018, 0.1%; 0.0, 0%)와 피험자-채점자 상호작용 분산성분(.4967, 0.2%; .4540, 0.2%) 은 상대적으로 매우 작으므로, 평가시기 안에서 피험자에 대한 채점자 간의 기준은 일정하다고 해석할 수 있다. 분산성분이 0 에 가까운 것은 평가시기에 따라 채점자 평가의 변동이 적은 것으로 해석 할 수 있다. 3국면이 혼재(confounding)되어있는 상호작용 분산성분은 각각 .7497, .5968 로 전체 분산의 3.4%, 2.8% 정도는 설명하기 어려운 잔차 성분으로 나타났다. 단, 전집점수인 피험자 분산성분이 상대적으로 크지 않은 이유는 등급을 달리하여 학습하고 연주하게 되므로 동일한 학생이 동일한 곡을 반복하여 연주함으로써, 수행(performance)의 변화된 정도를 수리적으로 파악할 수 있는 여지는 이 설계에서 이미 통제되었다고 볼 수 있다.

따라서, G 연구 결과 피아노 연주와 바이올린 연주 실기 평가 채점에서 평가시기에 따른 문항의 변화이외 채점자 변동이 약간 보일 뿐 매우 안정적인 채점 상황으로 해석할 수 있다. 물론, 피험자가 어떻게 향상되었는가에 초점을 맞추지 못한 설계 상의 아쉬움이 있지만, 시기와 채점자가 달라짐에 따라 채점에 미치는 오차의 변동을 동시에 분석하여 측정조건을 제시할 수 있는 틀을 제공하였다. 위의 결과에서 보여지듯이 채점자와 관련된 분산성분은 상대적으로 매우 작으므로, 채점자 사이에서 파생되는 변동은 측정오차에 미치는 영향이 적다고 해석할 수 있다.

<표 2> 피아노 실기평가 p x ((r x i) : o) 설계 G(일반화) 연구 결과

효과	자유도	제곱합 (SS)	제곱평균 (MS)	분산성분	비율
P	14	141.72222	10.12302	.2663690	.012
O	1	.71111	.71111	(0.0)	
R:O	2	10.72222	5.36111	.0181217	.001
I:O	10	6223.87778	622.38778	20.721269	.9312
PO	14	15.78889	1.12778	(0.0)	
PR:O	28	104.44444	3.73016	.4967460	.0023
PI:O	140	97.62222	.69730	(0.0)	
RI:O	10	5.87778	.58778	(0.0)	
PRI:O, e	140	104.95556	.74968	.7496825	.0337
TOTAL	359	6705.72222			

p: 피험자 o: 평가시기 r: 채점자 i: 평가항목

<표 3> 바이올린 실기평가 p x ((r x i) : o) 설계 G(일반화) 연구 결과

효과	자유도	제곱합 (SS)	제곱평균 (MS)	분산성분	비율
P	14	139.88889	9.99206	.2722222	.0118
O	1	7.80278	7.80278	(0.0)	
R:O	2	.27222	.13611	(0.0)	
I:O	10	6481.36111	648.13611	21.5800397	.9394
PO	14	16.15556	1.15397	(0.0)	
PR:O	28	92.97778	3.32063	.4539683	.0198
PI:O	140	102.88889	.73492	.0690476	.003
RI:O	10	5.69444	.56944	(0.0)	
PRI:O,e	140	83.55556	.59683	.5968254	.0277
TOTAL	359	6930.59722			

p: 피험자 o: 평가시기 r: 채점자 i: 평가항목

나. D 연구 분석결과

G 연구(일반화연구) 결과로 산출 된 각 분산성분은 D 연구(결정연구)를 수행함으로 결정과정에 필요한 정보를 제공한다. 일반화가능도 계수는 단일 차원(수)의 평가시기, 채점자, 평가항목을 기준으로 산출되므로, 각 국면의 수에 따라 계수는 상향 조정될 수 있다. 이 연구에서 각 국면을 임의 효과로 설정한 G 연구 설계와 동일한 D 연구 설계를 수행하여 평가시기의 수, 채점자 수, 그리고, 평가항목을 확장시킴으로써 계수 변동의 추이를 비교하였다. G 연구 결과 각 분산성분의 상대적 크기를 비교 해 볼 때, 채점자의 수를 증가시키는 것보다 평가시기나 평가항목를 늘리는 것이 더 효율적으로 계수를 개선시킬 수 있다. 피아노 연주와 바이올린 연주에서 평가시기, 채점자, 평가항목 증가에 따라 오차분산성분의 변동과 일반화가능도 계수 추정값을 산출한 D 연구결과를 <표 4>와 <표 5>와 같이 요약하였다.

D 연구 결과는 상대 오차분산에 포함되는 분산성분 크기에 따라, 적정 수준 계수 값(.70) 에 도달하는 평가시기, 채점자, 평가항목의 수를 조합하여 얻은 정보를 제공한다. <표 4>와 <표 5> 에서 보듯이 일반화 가능도 계수는 평가 시기의 수가 증가함에 따라 급속도로 개선되는 반면, 채점자 수의 증가에 따른 변화는 둔하게 나타났다. 즉, 평가시기의 효과가 기록의 일반화 과정에 더 강하게 작용하는 것이다. 따라서, 각 국면의 기본조건 ($n_o = n_r = n_i = 1$)에 의한 일반화가능도 계수는 적어도 .7 의 일반화가능도 계수를 산출하기 위해 피아노연주와 바이올린 연주 모두 평가시기 2번 이상, 채점자 3명 이상, 그리고 평가항목(연주 또는 수행과제) 4 가지 이상 또는 6가지 정도의 조합이 필요하다. 또한 동일한 조건에서 평가항목의 수를 늘리는 것보다 채점자 수를 늘리는 것이 효율적으로 일반화가능도 계수를 개선시키고 있다.(평가시기 2번, 채점자 2명, 평가항목 4번에서 평가시기 2번, 채점자 2명, 평가항목 6 가지와 평가시기 2번, 채점자 3명, 평가항목 4가지로 조정하였을 때 계수 변화에 대한 비교를 참조하라.)

Ⅳ. FACETS 모형의 적용

1. FACETS 모형의 소개 및 적용

Linacre(1988; 1989; 1994)에 의해 개발된 Facets 모형은 최근 발전된 측정이론의 하나인 문항반응이론(Item response theory)에 기초한 모형으로, 일모수 라쉬모형(Rasch, 1960; 1980; Wright & Stone, 1979)을 확장시킨 것이다. 수식 (3) 에서 볼 수 있는 것처럼 실기평가 상황의 모든 변인-- 여기에서는 과제의 곤란도와 연주자의 능력수준, 그리고 채점자의 엄격한 정도 --를 동일한 선형척도 위에 수량화시킬 수 있는 준거틀을 제공한다. 이때 과제의 곤란도와 채점자의 엄격한 정도를 동간척도상의 영점(zero)에 고정시키면 이들 두 변인으로부터 자유로운-- 즉, 두 변인의 영향력이 제거된 -- 연주자 실기능력 수준에 대한 로짓 점수를 얻을 수 있다 (Lunz, et al., 1990).

$$\log(P_{nijk} / P_{nijk-1}) = B_n - D_i - C_j - F_k \quad (3)$$

P_{nijk} = 연주자 n이 과제 i에 대해 채점자 j로부터 k 점수를 받을 확률
P_{nijk-1} = 연주자 n이 과제 i에 대해 채점자 j로부터 k-1 점수를 받을 확률
B_n = 연주자 n의 능력수준
D_i = 과제 i의 곤란도
C_j = 채점자 j의 엄격한 정도
F_k = 점수 k-1에 비해 점수 k를 받기가 어려운 정도

·Facets 모형을 실제 자료분석에 적용하기 위해서는 몇 가지 전제조건이 만족되어야 한다. 하나는 과제가 단일능력을 측정하는 동질적인 과제들로 구성되었다는 일차원성(unidimensionality)의 가정을 만족해야 하는데, 이는 이 모형이 문항반응이론을 기반으로 하고 있기 때문이다. 검사가 일차원성의 가정을 만족하는지의 여부는 Facets 분석 후 제시된 모형적합도 검정결과를 통해 판단할 수 있다 (Chi-square 검정 결과 본 연구의 음악 실기평가 자료는 일차원성의 가정을 만족하는 것으로 나타났다).

또다른 조건은 모든 채점자가 일정한 분량의 공통답안에 의해 서로 연결되는 채점을 해야 한다는 것인데, 본 연구의 자료는 두 명의 채점자가 동일한 연주자를 채점함으로서 모든 자료가 서로 연결되므로 두 번째의 전제조건 역시 만족하고 있다. 따라서 과제의 곤란도와 채점자의 엄격한 정도를 통제한 연주자의 능력모수치를 추정해 낼 수 있는 Facets 분석이 가능해지는 것이다.

2. Facets 모형을 적용한 음악 실기평가 분석 결과

가. 실기평가의 양호도에 대한 전반적인 분석 결과

피아노(1회) 평가, 피아노(2회) 평가, 바이올린(1회) 평가, 바이올린(2회) 평가의 양호도를 검증하기 위하여 각각에 대한 Facets 분석을 수행한 결과는 다음과 같다.

1) 피아노(1회) 평가

[그림 1]은 피아노(1회) 평가의 과제 곤란도와 채점자의 엄격한 정도, 그리고 연주자의 능력 수준을 단일한 동간척도 상에 제시하고 있다. [그림 1]을 통해, 피아노(1회) 평가의 6 과제는 전반적으로 15명의 연주자의 능력수준에 알맞는 곤란도 수준의 과제들임을 알 수 있다. 그러나 연주자 11번이나 7번 등 연주능력이 매우 높은 연주자에 알맞는 약간 어려운 과제가 첨부된다면 능력수준이 높은 연주자의 능력 모수치 추정에 도움이 될 것으로 보인다. 피아노(1회) 평가 과제들의 적합도 지수는 0.6-1.5이고 분리신뢰도 값이 .86으로 피아노 실기평가를 위한 과제들이 전반적으로 양호함을 알 수 있다.

2) 피아노(2회) 평가

[그림 2]는 피아노(2회) 평가의 과제 곤란도와 채점자의 엄격한 정도, 그리고 연주자의 능력 수준을 단일한 동간척도 상에 제시하고 있다. [그림 2]를 통해, 피아노(2회) 평가의 6 과제는 15명의 연주자의 능력수준에 비해 전반적으로 쉬운 과제들임을 알 수 있다. 연주자 11번, 7번, 3번, 12번 등 연주능력이 높은 연주자들에게 피아노(2회) 평가에서 제공된 과제들은 지나치게 쉬운 경향이 있어서 능력수준이 높은 연주자의 모수치 추정을 위해 약간 어려운 과제를 보충할 필요가 있다. 반면, 과제6은 대부분의 연주자에게 너무 쉬운 과제라서 연주자의 능력 모수치 추정에는 별 도움이 되지 않는 과제임을 알 수 있다. 피아노(2회) 평가 과제들의 적합도 지수는 0.6-1.5이고 분리신뢰도 값이 .86으로 피아노 실기평가를 위한 과제들은 전반적으로 양호한 것으로 나타났다.

3) 바이올린(1회) 평가

[그림 3]은 바이올린(1회) 평가의 과제 곤란도와 채점자의 엄격한 정도, 그리고 연주자의 능력 수준을 단일한 동간척도 상에 제시하고 있다. [그림3]을 통해, 바이올린(1회) 평가의 6과제는 15명의 연주자의 능력수준에 비해 전반적으로 쉬운 과제들임을 알 수 있다. 연주자 8번, 4번, 15번, 1번 등 연주능력이 높은 연주자들에게 바이올린(1회) 평가에서 제공된 과제들은 지나치게 쉬운 경향이 있어서 능력수준이 높은 연주자의 모수치 추정을 위해 약간 어려운 과제를 보충할 필요가 있다. 반면, 연주자 10번이나 7번과 같이 연주능력이 다소 떨어지는 연주자들의 능력모수치를 추정하는데 알맞는 곤란도가 낮은 과제 역시 보충될 필요가 있다. 바이올린(1회) 평가 과제들의 적합도 지수는 0.7 1.4이고 분리신뢰도 값이 .66으로 바이올린 실기평가를 위한 과제들이 전반적으로 양호한 것으로 나타났다.

4) 바이올린(2회) 평가

[그림 4]는 바이올린(2회) 평가의 과제 곤란도와 채점자의 엄격한 정도, 그리고 연주자의 능력 수준을 단일한 동간척도 상에 제시하고 있다. [그림4]를 통해, 바이올린(2회) 평가의 6과제는 15명의 연주자의 능력수준에 비해 전반적으로 쉬운 과제들임을 알 수 있다. 연주자 8번, 4번, 15번, 1번 등 연주능력이 높은 연주자들에게 바이올린(2회) 평가에서 제공된 과제들은 지나치게 쉬운 경향이 있어서 능력수준이 높은 연주자의 모수치 추정을 위해 약간 어려운 과제를 보충할 필요가 있다. 바이올린(1회) 평가 과제들의 적합도 지수는 0.4 1.2이고 분리신뢰도 값이 .75로 바이올린 실기평가를 위한 과제들이 전반적으로 양호한 것으로 나타났다.

나. 채점자간 신뢰도 분석 결과

피아노와 바이올린에 대한 네 번의 실기평가별로, 두 채점자간 신뢰도를 분석한 결과는 다음과 같다. 피아노(1회)와 피아노(2회) 평가의 경우, 두 채점자의 엄격한 정도의 차이에 대한 Chi-square 검정 결과 통계적으로 유의미한 차이가 있는 것으로 나타났다 (피아노 1회 평가: Chi-square 값 = 6.4, P-value = .01; 피아노 2회 평가: Chi-square 값 = 6.7, P-value = .01). 반면, 바이올린(1회)와 바이올린(2회) 평가의 경우, 두 채점자의 엄격한 정도의 차이에 대한 Chi-square 검정 결과 통계적으로 유의미한 차이가 없는 것으로 나타났다.

(바이올린 1회 평가: Chi-square 값 = 1.0, P-value = .32;
바이올린 2회 평가: Chi-square 값 = 0, P-value = .85).

두 채점자간 신뢰도 분석 결과는 각 [그림]에 제시된 결과와도 일치함을 알 수 있다. [그림 1]와 [그림 2]

의 피아노 실기평가의 경우, 채점자1과 채점자2의 엄격한 정도의 로짓점수가 약 .4만큼의 간격을 두고 떨어져 있음을 보여준다 (피아노 1회평가의 경우, -.08에서 -.46; 피아노 2회 평가의 경우 .04에서 -.44). 반면, [그림 3]와 [그림 4]의 바이올린 실기평가의 경우, 채점자1과 채점자2의 엄격한 정도의 로짓점수가 거의 일치함을 보여준다 (바이올린 1회평가의 경우, -.07에서 -.23; 바이올린 2회 평가의 경우 -.08에서 -.12).

결과적으로, 피아노 실기평가의 경우에는 두 채점자간에 엄격한 정도가 통계적으로 유의미한 정도까지 차이가 나므로, 채점자 변인을 통제한 후의 연주자 능력 모수치 산출이 바람직할 것이다. 반면, 바이올린 실기평가의 경우에는 두 채점자간에 엄격한 정도의 차이를 거의 보이지 않으므로, 달리 채점자 변인의 영향을 통제할 필요가 없게 된다.

다. 연주자의 최종 능력모수치 산출

연주자의 최종 능력모수치를 산출하기 위하여, 과제의 곤란도와 채점자의 엄격한 정도를 영점에 고정시킨 후 Facets 모형에 의한 재분석을 시도하였다. 재분석을 통해 산출된 연주자 능력모수치는 과제의 곤란도와 채점자의 엄격한 정도로부터 자유로운 모수치이다. 채점자간 신뢰도 분석결과로부터 짐작할 수 있듯이 바이올린 실기평가의 경우에는 채점자 변인을 통제하기 전과 통제한 후의 능력모수치가 거의 일치한다 (바이올린 1회 평가의 경우 .15의 차이; 바이올린 2회 평가의 경우 .10의 차이). 반면 피아노 평가의 경우에는 채점자 변인을 통제하기 전과 통제한 후의 능력모수치 간에 어느 정도의 차이(피아노 1회 평가의 경우 .27의 차이; 피아노 2회 평가의 경우 .20의 차이)를 보이지만, 연주자의 능력별 순서를 뒤바꿀 정도의 변화를 보이지는 않는다.

Ⅴ. 요약 및 논의

기악연주 실기평가에 G 이론을 적용함으로써 얻을 수 있는 결과를 요약·정리하면 다음과 같다. 첫째, 연구자료 분석결과 연주되는 악기의 종류와 상관없이 동일한 학생의 수행에 대하여 시기를 달리하거나 채점자를 달리하여도 평가되는 어떤 항목에 대한 채점결과는 유사하게 나타났다. 그러나, 평가항목의 변화는 매우 의미 있는 것으로 나타났다. 따라서, 평가요소와 평가영역을 구체화시키고, 즉, 요소의 수를 늘린다든지, 평가기준을 다양하게 한다든지, 평가요소를 세분화한다든지 하는 방법으로 실기평가 채점의 일반화가능성은 향상될 수 있다. 둘째, 本 자료는 채점과정에서 다른 평가 시기가 의미하는 것은 평가되어야 할 평가항목의 내용이 다르므로 피험자 자체의 변동은 설계상 통제되었다. 그리고, D(결정) 연구 결과에서 보여지듯이 일반화가능도 계수는 채점자 수를 늘리는 것보다 평가시기를 다양화하는 것이 일반화가능도 계수는 효율적으로 향상되었다. 적어도 .7 이상의 일반화가능도 계수를 산출하기 위하여, 평가시기 3, 채점자 3, 평가항목 4 또는 6 의 조합이 두 악기 모두에서 공통된 결과로 나타났다. 물론, 단순히 국면의 수를 늘리는 것이 타당하고 양호한 영역 또는 시기를 보장하는 것은 아니다. 따라서, 최적 수준의 계수 산출을 위하여 제시한 차원(수)은 그 내용과 질의 검토가 수반되어야 의미있는 결과를 유도할 수 있다.

실기평가의 채점이 포함하고 있는 문제점을 극복하려면, 채점자 수를 늘리는 것보다 평가영역과 기준을 구체화, 명시화하여 피험자와 채점자 모두에게 발생할 수 있는 오차요인을 축소시켜야 할 것이다. 학생들에게는 평가 영역과 기준에 대한 지침을 공개하여, 각 영역과 평가요소에서의 혼돈을 방지하여야 한다. 또한, 제시된 평가항목, 특히 평가요소의 기준 자체가 모호하여, 채점자의 해석에 따라 결과가 달라질 수 있는 오차를 최소화시켜야 한다. 즉, 채점자에게도 추상적인 평가 기준이 아닌 구체적 지침을 제시하고 순서효과, 채점자내 변동 등을 제거할 수 있는 채점훈련이 요구된다. 이번 연구 자료로 사용된 RCM의 경우 '전문 평가사'라고 불리는 평가자는 인증제도에 의한 자격증을 기반으로 매년 재교육을 통하여 전문성을 검증하고 있다. 채점자 간의 신뢰성을 보장하고 있어 점점 단일 평가자에 의한 평가결과를 반영하려는 추세라고 한다. 주지하듯이 타당성 있는 과제에 대하여 피험자가 표현할 수 있는 내용을 수행하였을 때, 평가영역, 평가요소, 채점자, 채점시기 등 각 채점과 관련된 변동요인이 최소화되면 결과에 대한 일관성과 일반화가능성이 확장되고, 음악연주를 포함한 모든 실기 평

가과정의 質的 향상을 유도할 수 있을 것이다.

한편, Facets 모형을 음악 실기평가 채점결과 분석에 적용하여 얻을 수 있는 점을 요약·정리해 보면 다음과 같다. 첫째, '채점자-연주자-과제(평가항목) 동일척도화,' '과제별 적합도 지수,' '과제의 분리신뢰도' 등을 산출하여 평가과제(항목)가 연주자의 능력수준에 적합한 것인지를 검증할 수 있다. 이러한 결과는 동일한 종류의 후속 평가를 계획할 때, 평가과제의 곤란도 수준을 어느 정도로 해야 하는지에 대한 구체적인 정보를 제공해 준다.

둘째, 채점자간의 신뢰도 지수를 통해, 연주자의 최종 능력 모수치를 산출할 때 채점자 변인의 영향을 통제할 필요가 있는지의 여부를 검증할 수 있다. 본 연구에서는 바이올린 평가의 경우에는 채점자 변인의 영향 통제가 불필요할 정도로 채점자간의 신뢰도가 높았고, 피아노 평가의 경우에는 상대적으로 채점자간 신뢰도가 낮은 편이었으나 연주자 능력 모수치에 의미 있는 영향을 미칠 정도는 아니었다. 이는 G 이론을 적용하여 분석한 결과와도 동일함을 알 수 있다.

셋째, 연주자의 능력 모수치 추정에 영향을 주는 과제 변인과 연주자 변인을 동시에, 그러나 상호 독립적으로 분석하여 이들 두 변인으로부터 통계적으로 자유로운 연주자 능력 모수치를 제공해 준다.

G 이론에 의한 분석은 각 국면(오차)으로 인한 변동을 동시에 분석할 수 있는 틀을 제공하는데 기여한 반면, Facets 모형은 오차를 반영하여 개인의 점수를 조정한다는 장점을 지닌다. 따라서, 채점자 효과나 과제의 효과에 대하여 개인 점수 변동이 심한 경우 각 효과의 오차를 분석하고 그 결과에 준하여 개인점수를 조정한다면 신뢰성에 근거한 개인점수를 반영할 수 있을 것이다. 이때 평가과제 곤란도의 적합성 및 채점자간 신뢰도 지수에 대한 Facets 모형 적용 결과도 함께 활용할 수 있을 것이다. 실기평가에 대한 Facets 모형 분석은, 연주자의 능력 모수치 추정에 영향을 미치는 평가상황의 여러 변인들을 통제함으로써, 보다 신뢰로운 연주자 능력 모수치를 제공해 준다는 데 초점이 있다. 이때 연주자의 능력 모수치에 영향을 미치는 변인들을 파악하는데 있어서 G 이론 적용 결과를 활용할 수 있을 것이다.

음악 실기평가 채점 결과 분석을 위해 G 이론과 Facets 모형을 최대로 활용할 수 있으려면, 연주영역별(피아노와 바이올린), 연주시기별, 연주과제별로 공동 연주자와 공동 채점자에 의한 연계설계(linking design)가 필요하다. 예컨대, 연주시기별로 한 두 개의 공통 연주과제를 포함시키게 되면, Facets 분석을 통해 평가시기에 따라 연주자들의 연주능력이 어느 정도로 변화했는지에 관한 능력 모수치 변화량까지 추정할 수 있다. 연주자의 능력 모수치 변화량에 대한 정보는 지속적 평가를 강조하는 수행평가의 경우 특히 가치로운 정보일 것이다. 이와 같이 이 연구에서는 실제 수행한 악기연주 실기 평가에 대하여 일반화가능도 이론과 Facet 모형을 적용하여 채점상황에 미치는 오차분석을 수행하고 점수조정 과정을 제시함으로써 후속 평가설계에 도움이 되는 정보를 제공하였다는데 의의를 두고 있다. 단지 각각의 분석 결과에 대한 장단점은 비교되었지만, 자료입력 과정부터 각 이론의 기본 가정이나 결과 검증 등 수리적 배경에 대한 설명은 구체화하지 않았다. 채점자, 평가시기, 평가과제 등 여러 국면(효과)에 대한 오차를 분석하고 있는 두 이론을 단일 자료에 적용하여 분석·설명하고자 하는데 초점을 두었기 때문이다. 두가지 접근방법을 통하여 실제 사용하고 있는 여러 가지 수행평가 결과를 위한 채점모형으로 발전되려면 평가 설계의 통합적 접근과 용어의 통일, 그리고 결과해석의 단일화 문제 등에서 보완하여야 할 것이다.

참 고 문 헌

김성숙(1997). 고전검사이론의 극복: 일반화가능도 이론,「한국심리학회 동계세미나」.

김성숙(1995). 논술문항 채점의 변동요인 분석과 일반화가능도 계수의 최적화 조건.「교육평가연구, 8(1)」, 35-57.

배호순(1997). 포트포리오 평가의 질관리 방안의 모색.「교육평가연구」, 10(1), 75-104.

채선희(1996). 논술시험 채점의 공정성과 효율성 확보 방안: 채점과정의 엄격성과 FACETS 모형에 의한 채점결과 분석.「교육평가연구」, 9(1). 5-29.

한국교육평가연구회(1997).「수행평가의 이론과 실제」. 학술세미나 발표논문집.

Baxter, G.P., Shavelson, R.J., Goldman, S.R., & Pine, J.(1992). Evaluation of procedure-based scoring for hands-on science assessment. *Journal of Educational Measurement*, 29, 1-17.

Baker, E. L and others (1996). Dimensionality and generalizability of Domain-independent performance assessments, Journal of Educational Research, v89, n4, 197-205.

Brennan, R. (1992). *Elements of Generalizability Theory (2nd.)*. Iowa City, 1A: The American College Testing Program.

Chae, Sunhee (1998). Controlling the judge variable in grading essay-type items. *Journal of Outcome Measurement*. Vol. 2. No. 2.

Crick, J.E. & Brennan, R.L. (1983). Mannual for GENOVA: A Generalized analysis of variation system. *ACT technical bulletin, 43*, The American College Testing Program. Iowa city, IA.

Cronbach, L.J., Glesser, G.C., Nanda, H., & Rajaratnam, N.(1972). *The generalizability of behavioral measurement : The theory of generalizability for scores and profiles*. New York : John Wiley.

Koretz, D., Stecher, B., Klein, S., & McCaffrey, D. (1994). The Vermont portfolio assessment program: Findings & Implications. *Educational Measurement: Issues and Practices !3(3)*, Fall, 5-16.

Lehmann, R.H. (1990). Evaluation Studies : Reliability and Generalizability of ratings of compositions, Studies in Educational Evaluation, 16, 501 512.

Linacre, J. M.(1988). *FACETS*. Chicago: MESA Press.

Linacre, J. M.(1989; 1994). *Many-Faceted Rasch Measurement*. Chicago: MESA Press.

Lunz, M. E., Wright, B. D. & Linacre, J. M.(1990). Measureing the impact of judge severity on examination scores. *American Measurement in Education 3(4)*. 331-345.

Rasch, G.(1960; 1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.

Reckase, M. D.(1995). Portfolio assessment: A theoretical estimate of score reliability. Educational Measurement: *Issues and Practice. 14(1)*. 12-14.

Shavelson, R., Baxter, P. & Gao, X. (1993). Sampling variability of performance assessment, *Journal of Educational Measurement, 30(3)*, 215 235.

Shavelson, R., & Webb, N.M. (1991). *Generalizability theory : A Primer*. Newbury Park, CA: Sage.

Wiggins, G.(1989). Teaching to the (authentic) test. *Educational Leadership*. April. 41-7.

Wright, B. D. & Masters, G. N.(1982). *Rating scale analysis*. Chicago: MESA Press.

Wright, B. D. & Stone, M. H.(1979). *Best test design*. Chicago: MESA Press.

ABSTRACT

Application of Generalizability Theory and Many-Facet Rasch Model for Performance-Based Assessment of Music

Sungsook Kim
(KICE)
Sunhee Chae
(KICE)

Performance-based assessment is an authentic evaluation which can grasp what and how well students' can do in real-like situation. Since it is an impending task to control reliability of performance-based assessment, however, we need more technical developments to improve it. The purpose of this study was to examine both Generalizability theory and Many-facet Rasch model which have been the most frequently applied psychometric methods lately in evaluating the reliability of performance assessment. It also intends to figure out a way to combine these two methods so that we get more reliable scores in performance assessment of Music.

The data used in this study consists of 15 examinees' subscores as well as total scores obtained in the performance-based assessment in the domains of piano and violin provided by the Royal Conservatory of Music Examinations of Canada.

For the first part of the analysis, the generalizability theory was applied. We estimated the variance component of instruments, raters, occasions, and evaluative factors to compare the relative influence of each facet, and determined the optimal number of grading conditions of each facet that maximizes the generalizability coefficient.

For the next part of the analysis, the many-facet Rasch model was applied. We detected the intensity of inconsistency in examinee scores caused by raters and provided examinee logit scores after adjusting the effects of a rater facet on raw scores.

The results show that both generalizability theory and many-facet Rasch model would be useful to produce more reliable examinee scores in performance assessment of music. However, it is also found out that they have the relative strengths and weaknesses in producing more reliable scores. G theory is useful in detecting various error components in performance assessment. On the other hand, Facets model makes it possible to produce examinee scores after controlling these error components.