Journal of the Korean Data & Information Science Society
2010, **21**(1), 87–98

한국데이터정보과학회지

강의평가의 타당성과 신뢰성에 관한 연구 전주대학교 강의평가 결과를 중심으로

이기훈[1]

[1]전주대학교 경영학부

접수 2009년 12월 12일, 수정 2010년 1월 14일, 게재확정 2010년 1월 18일

요 약

본 논문은 강의평가의 타당성과 신뢰성을 측정하는 방법을 소개하고 실제자료를 이용하여 타당성과 신뢰성을 평가하였다. 기존의 강의평가 관련논문이 강의평가에 미치는 외생적인 영향을 통제하는데 주력한 데 반해, 교원의 신분에 직접적인 영향을 미칠 수 있는 중대한 강의평가가 과연 믿을 만한 평가인가에 관한 근원적인 질문에 답하려 하였다. 전주대학교의 강의평가 결과를 실증 분석한 결과 타당성과 신뢰성 면에서 어느 정도 만족할 만한 수준임을 확인할 수 있었다. 본 논문에서는 기존에 간편하게 사용되던 신뢰성 측도가 아닌 일반화가능도 계수를 이용하여 신뢰성을 평가하는 방법을 자세히 소개하고 그 장점을 설명하였다.

주요용어: 강의평가, 신뢰성, 일반화가능도, 타당성.

1. 머리말

우리나라의 대부분의 대학교에서 실시하는 강의평가 (students' evaluation)는 평가의 목적에 따라 형성적 (formative) 평가와 종합적 (summative) 평가로 나눌 수 있다. 형성적 평가는 학기 중에 학생들의 의견을 반영하여 강의의 질을 향상시키기 위한 목적으로 실시하며, 종합적 평가는 교수의 승진, 연봉 산정, 정년보장 임면 등 관리적 목적으로 주로 사용된다 (Crumbley와 Fliedner, 2002). 교수의 강의에 대하여 점검하고 의견을 들어 다음 강의의 개선자료로 활용하는 교수적 기능이 강의평가 실시의 초기단계의 목적이었다면 근래에 들어서는 업적평가, 승진 등에 반영하는 행정적인 기능이 점차 강조되고 있다. 강의평가가 교수의 신분과 처우에 영향을 미치게 되자 과거에 비해 강의평가의 신뢰성에 대한 대학교원들의 관심이 더욱 커지게 되었다. 이러한 강의평가의 신뢰성에 관한 논란은 국내에 일부대학에서 강의평가를 도입하던 시기에서부터 시작되었고 (김영진, 1994), 강의평가 실시가 60년이 넘는 미국에서도 계속되고 있다 (Ahmadi 등, 2001).

믿을만한 평가인가라는 질문은 그 평가가 타당성 (validity)과 신뢰성 (reliability)을 확보하고 있는가에 관한 질문과 일치한다. 고전검사이론 (classical test theory)에 의하면 평가점수는 타당한 점수, 타당하지 않은 점수, 오차점수 등으로 이루어져 있다고 정의한다. 타당한 점수와 타당하지 않은 점수를 합하여 진점수 (true score)라 한다. 전체점수 중에서 진점수에 해당하는 부분이 신뢰도에 관련되며 진점수 중에 타당한 점수 부분이 타당도에 해당하는 부분이다. 또한 오차점수는 우리가 통제하기 어려운 측정과정에서 발생하는 오차로 평가점수 중에서 오차점수의 비중이 낮으면 평가가 신뢰할 수 있다고 판단

[1] (560-759) 전북 전주시 완산구 효자동 3가, 전주대학교 경영학부, 교수. E-mail: khlee@jj.ac.kr

한다 (성태제, 2002). 즉 강의평가가 믿을만한 평가가 되기 위해서는 강의평가의 타당성과 신뢰성 등이 차례로 확보되어야 할 것이다.

우리나라에서 많은 학교들이 강의평가를 실시하고 이를 교수의 업적에 반영함에도 불구하고 강의평가 제도의 문제점을 인식하고 이를 개선하려는 노력은 크게 이루어지지 않고 있다 (한신일, 2002; 류춘호와 이정호, 2003). 현재까지의 논문은 대부분 자신의 학교자료를 대상으로 어떠한 요인이 강의평가에 영향을 주는가에 대한 기술적 (descriptive) 분석과 강의평가에 영향을 주는 요인을 통제하는 방법을 제시하는 내용이 대부분이다 (김종태, 2004; 김성연과 권치명, 2005; 조장식 등, 2009; 박노진, 2009; Lee와 Lee, 2006; Baik와 Yang, 2008).

강의평가가 단순히 교수의 개인적인 강의개선의 피드백 자료로 사용되던 과거에 비해 교수의 교육업적 평가의 자료로 사용되는 빈도가 많아진 현시점에서 강의평가의 타당성과 신뢰성은 교육학 관련 전공자만이 아닌 교육자 모두에게 관심의 대상이 되고 있다. 본 논문에서는 강의평가의 타당성과 신뢰성에 관하여 논의하고자 2008년도 2학기 전주대학교의 일부 교양영어과목의 강의평가결과 자료를 실증분석하고자 한다. 2절에서는 타당성의 개념과 측정방법, 이에 따른 강의평가의 타당성을 평가 (assessment)하고 3절에서는 신뢰도의 개념과 기존의 측정방법을 소개한다. 4절에서는 본 연구에서 제안하는 신뢰성 측정방법인 일반화가능도 이론을 소개하고 강의평가의 신뢰성 평가를 위한 실증연구를 수행하고, 5절에서는 결론과 향후 연구과제에 관하여 언급할 것이다.

2. 강의평가의 타당성

타당성은 검사 (test)가 측정하고자 하는 내용을 얼마나 정확하게 (accurate) 측정하였는가에 관한 측도이다. AERA, APA, NCME (American Educational Research Association, American Psychological Association, National Council on Measurement in Education) (1999)의 Standards for Educational and Psychological Testing에서는 제안된 검사도구에 의해 얻어진 검사점수의 해석에 증거나 이론이 지지하여주는 정도로 정의하고 있다.

AERA 등 (1999)는 타당도를 다섯 가지로 분류하였다. 이는 '검사내용에 기초한 근거', '반응절차에 기초한 근거', '내부구조에 기초한 근거', '다른 변수들과의 관계에 기초한 근거', '검사결과에 기초한 근거' 등이다 (성태제, 2002).

과거에 내용타당도라고 명명되던 '검사내용에 기초한 근거'는 계량화된 정보를 제공하지 못하고 전문가의 주관에 의한 판단에 의존하게 되며 타당성의 정도를 계량적으로 표시할 수 없다. 전주대학교의 강의평가 문항 (표 2.1)이 과거 여러 대학의 강의평가 문항과 유사하고 수년간 시행하면서 전문가들의 의견을 적절히 반영한 점을 고려한다면 내용타당도에서는 문제가 없는 것으로 판단된다.

'반응절차에 기초한 근거'는 피험자의 응답에 대한 이론적이고 경험적인 분석을 통해 평가자의 반응과 피험자의 수행이 일치하는가를 판단하게 된다. 이는 각 교수들의 경험적 의견으로 휴강이 없는 경우 1번 문항이 높게 평가 (역코딩)되고, 난해한 과목일수록 난이도에 대한 평가가 낮다는 점 등에서 수행과 반응이 일치한다는 사실을 추론할 수 있다. 그러나 이와 관련한 계량적인 자료가 존재하지 않으므로 본 논문에서는 내용타당도와 더불어 이에 대한 상세한 분석을 진행하지 않겠다.

'내부구조에 기초한 근거'는 구인 (構因, construct) 타당도 또는 구성타당도라고 불리며 문항과 요인들 사이의 관계들이 검사점수 해석에 기초하는 구조를 검증하는 정도를 의미한다. 이를 검증하기 위해 가장 많이 사용하는 통계적 방법이 요인분석 (factor analysis)이다. 표 2.1에 의하면 강의평가 설문지는 강의를 구성하는 속성과 관련된 구체적인 변수들로 구성되어 있어 요인분석을 할 경우 2-3개의 요인이 검출될 것으로 예상할 수 있다. 그러나 본 논문에서 분석하는 자료에서는 고유값 (eigen value)이 1 이상이 되는 공통요인을 하나밖에 검출하지 못하였다. 이는 온라인상에서 학생들이 성적열람 전에 수강

표 2.1 전주대학교 강의평가 문항과 분석자료의 기초통계량(n=640)

문항	설 문 내 용	평균	표준편차
1	교수는 공식 행사 외에 휴강을 몇 번 정도 하였습니까	4.86	.513
2	강의는 강의 계획서에 맞추어 진행되었습니까	4.03	.896
3	교재(학습자료)는 학습에 도움이 되었습니까	3.86	.977
4	교수는 강의 준비를 철저히 하였습니까	3.89	.923
5	교수는 강의내용을 이해하기 쉽게 전달하였습니까	3.86	.988
6	강의의 성격에 맞게 적절한 수업방법을 사용하였습니까	3.86	.924
7	교수는 학생들의 관심과 참여를 유도했습니까	4.00	.928
8	강의의 난이도는 적당하였습니까	3.77	.919
9	성적의 평가기준은 합리적이고 미리 제시되었습니까	3.91	.948
10	강의에 대하여 전반적으로 만족합니까	3.85	.962

과목에 대하여 일괄적으로 평가를 진행하면서 모든 문항에 동일한 평가점수에 체크하는 불성실한 경우가 다수 발생하여 변수들 간에 상관계수가 높게 측정되어 생기는 현상이라 할 수 있다. 그래서 본 논문에서는 불성실한 응답자료를 제외하고자 표준편차가 0.7 이상인 응답자의 자료만으로 요인분석을 하였다 (n=184). 여기서 표준편차가 0.7 이상이 되려면 대략적으로 5점 척도에서 3, 4, 5 등의 평가가 균등하게 분포되어 있음을 의미하는데, 특별히 표준편차 0.7 이상이 성실한 답변이라고 판단할 근거는 없지만 표준편차를 기준으로 자료를 선택하여 반복적으로 요인분석을 한 결과 0.7 이상인 경우에 비로소 여러 개의 요인이 검출되었기 때문에 0.7을 기준으로 자료를 선택하였다. 그 결과 고유값 1 이상인 공통요인이 3개 검출되었고 이들이 설명하는 총분산의 비율은 60.2%에 달하였다. 다음 표 2.2에 표시되어 있는 직교회전에 의한 각 요인의 요인적재 추정값을 통해 요인들을 해석하면 첫 번째 요인은 높게 적재된 변수가 성적의 평가기준, 전반적 만족도, 난이도, 교재 등이기 때문에 강의 구성에 관한 요인으로 해석할 수 있다. 두 번째 요인은 교수의 성실성 (휴강횟수, 강의계획서 등)에 관련한 요인이고, 세 번째는 강의내용의 전달, 수업방법, 강의준비, 참여유도 등 교수법에 관련한 요인으로 해석할 수 있다. 이렇듯 세 가지 요인은 직관적으로도 우리가 강의를 구성하는 요인이라 추정할 수 있는 요소들로 이러한 요인분석의 결과를 통해 강의평가 설문내용의 구성타당도가 높다고 평가할 수 있다.

표 2.2 직교회전 (varimax)에 의한 요인적재 추정값

설 문 내 용	성분의 적재값		
	1	2	3
9. 성적의 평가기준은 합리적이고 미리 제시되었습니까	.758	.136	-.032
10. 강의에 대하여 전반적으로 만족합니까	.724	.110	.142
8. 강의의 난이도는 적당하였습니까	.704	-.068	.307
3. 교재(학습자료)는 학습에 도움이 되었습니까	.524	.339	.345
1. 교수는 공식 행사 외에 휴강을 몇 번 정도 하였습니까	.030	.872	.025
2. 강의는 강의 계획서에 맞추어 진행되었습니까	.171	.852	.127
5. 교수는 강의내용을 이해하기 쉽게 전달하였습니까	-.018	.178	.805
6. 강의의 성격에 맞게 적절한 수업방법을 사용하였습니까	.199	-.133	.690
4. 교수는 강의 준비를 철저히 하였습니까	.260	.472	.575
7. 교수는 학생들의 관심과 참여를 유도했습니까	.359	.183	.483

'다른 변수들과의 관계에 기초한 근거'는 외적변수와 검사점수의 관계를 분석하여 타당성을 검증하는 방법이다. 이는 '수렴과 판별근거 (convergent and discriminant evidence)'와 '검사-준거 관련성 (test-criterion relationship)', '타당도 일반화 (validity generalization)' 등으로 이루어져 있다 (성태제, 2002). 여기서는 강의평가의 특성상 '검사-준거 타당성'에만 관심을 갖도록 하겠다. 강의평가에 대한 준거로 학생들의 학업성취도를 포함할 수 있으며 이는 해당과목의 학업성적으로 대신할 수 있을 것

이다. 강의평가가 교수개인의 교수법과 교과특성에 관한 평가인 동시에 학생들의 만족도에 관한 조사이기 때문에 이는 학생들 개인의 학업성취도와 밀접한 관계가 있을 것이다. 다음 표 2.3에 의하면 학생들의 학업성적과 강의평가점수 (10문항 평균)의 회귀분석 결과, 강의평가가 학업성적에 유의한 영향을 줌으로 '검사-준거 관련성'이 높은 것으로 판단된다.

표 2.3 회귀분석에 의한 회귀계수 추정값과 유의성

종속변수: 학업성적		비표준화계수	표준오차오류	t	유의확률
독립 변수:	(상수)	46.627	4.401	10.594	.000
	강의평가 점수	7.731	1.157	6.682	.000

'검사결과에 기초한 근거'는 결과타당도라 불리며 실시한 검사의 목적과 그 영향에 대하여 고려해야 한다는 것이다. Messick (1998)이 평가점수 사용에 대한 윤리적 문제를 제기하며 결과타당도를 제안하였고 이를 타당도 범주 안에 포함시켜야 하는가에 대한 논란은 Shepard (1997), Linn (1997), Popham (1997), Mehrens (1997) 등의 논문에 의해 찬반이 엇갈리고 있다. 강의평가에서 결과 타당도는 주관적으로 판단될 수밖에 없는데 Linn과 Gronlund (2000)에 의하면 다음과 같이 점검해볼 수 있을 것이다. 첫째, 강의평가가 원래 측정하고자 하는 것이 대학의 교육목표와 부합하는지, 둘째, 교수가 평가를 잘 받기 위해 강의에 더 노력하는지, 셋째, 평가가 교수들의 교육을 인위적으로 제한하지는 않는지, 넷째, 평가가 교수들의 창의적 교육방법이나 연구를 격려하고 있는지 등을 확인하여 그 타당성을 이야기할 수 있다. 이와 관련한 내용은 정성적인 평가에 의해 이루어지므로 본 논문에서는 그 논의를 제외하고자 한다.

3. 강의평가의 신뢰성

신뢰성은 검사도구가 측정하고자 하는 것을 얼마나 일관성 있게 측정하였는가를 의미한다. AERA 등 (1999)는 신뢰도는 피험자들에 동일한 시험을 반복적으로 시행하였을 때 그 측정의 일관성 (consistency)이라고 정의한다.

강의평가의 신뢰성은 대개 평가결과의 내적 일관성 (internal consistency)과 안정성을 통하여 확인한다. 여기서 내적 일관성이란 동일한 학생집단이 여러 가지 형식의 평가도구를 이용하여 특정 강의를 평가했을 때 각 평가도구에서 얻은 결과간의 일치정도를 의미하는 것이다. 그리고 안정성은 동일 교수의 강의를 시기를 달리하여 두 번 이상 평가한 결과 사이에 어느 정도 일치하는 가에 따라 결정된다 (이종승, 1995). 미국의 경우 대체로 학생에 의한 교수강의평가는 내적 일관성과 안정성이라는 관점에서 볼 때 비교적 만족할만한 신뢰도를 나타내고 있는 것으로 보고되고 있다 (Kulik와 McKeachie, 1975).

신뢰도를 측정하는 고전검사이론의 방법은 크게 두 가지로 나눌 수 있다. 첫 번째 방법으로 동일한 피험자에게 동일한 시험을 적당한 시간 간격을 두고 반복 측정하여 두 결과의 유사성을 피어슨 상관계수로 측정하는 것이다. 두 번째는 평가점수의 분산을 처리분산과 오차분산으로 나누어 오차분산의 비율이 낮으면 이 측정 도구는 일관성이 있으므로 신뢰도가 높다고 평가하는 것이다. 그러나 첫 번째 방법은 여러 가지 현실적인 이유 (예. 시험효과, 성숙효과, 적당한 시간간격 결정 문제 등) 때문에 권장되지 않고 두 번째 방법이 주로 사용되어 왔다. 이러한 측도들은 다음과 같은 과정을 통해 정의된다.

어떠한 교과목에 평가자가 n명이고 평가항목이 k개일 때 평가점수의 모형을 다음과 같이 가정하자.

$$x_{ij} = \mu + \nu_i + \beta_j + e_{ij} = T_{ij} + e_{ij}; i = 1, \dots, n, j = 1, \dots, k,$$

여기서 ν_i는 평가자 효과, β_j는 문항 효과, T_{ij}는 진점수, e_{ij}는 오차항이다. 이때 평가점수의 분산은

다음과 같이 표현할 수 있다.

$$Var(x_{ij}) = \sigma_X^2 = \sigma_T^2 + \sigma_e^2 = Var(T_{ij}) + Var(e_{ij}); i = 1, \dots, n, j = 1, \dots, k. \tag{3.1}$$

고전검사이론에서 신뢰도는 다음과 같이 진점수의 분산과 관찰점수의 분산의 비로 정의된다.

$$\rho^2 = \frac{\sigma_T^2}{\sigma_X^2} = 1 - \frac{\sigma_e^2}{\sigma_X^2}. \tag{3.2}$$

Spearman (1910)은 자료를 양분 (split-half)하여 이들 간의 상관계수로 신뢰도를 추정하였는데, 이 값은 문항을 어떻게 둘로 나누는가 (예. 홀수번 문항 대 짝수번 문항)에 따라 신뢰도가 달리 계산될 수 있다는 단점이 있었으나 시험을 두 번에 걸쳐 반복할 필요가 없이 한번 실험만으로도 신뢰도를 구할 수 있다는 장점을 가졌다. 그 후로 Kuder와 Richardson (1937)은 자료를 양분하는 방법에 무관하게 신뢰도를 계산하는 방법을 제안하였는데 이분문항을 위한 KR-20 (Kuder-Richardson Formula 20), 연속형 문항점수를 위한 KR-21 등이다. 그리고 Cronbach (1951)가 0-1 문항이 아닌 일반적인 연속형 측도에서 사용할 수 있는 신뢰도 계수 α를 제안하였는데, 이 값은 자료를 모든 가능한 경우의 수로 둘로 나누어 상관계수를 구하고 이를 평균한 값과 일치하는 특성을 갖고 있으며 KR-20의 일반형이라 할 수 있다. 크론바하 알파는 다음과 같이 정의된다.

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{j=1}^k \sigma_j^2}{\sigma_X^2}\right),$$

여기서 $\sigma_j^2 (j = 1, \dots, k)$는 j번째 문항의 분산이다.

신뢰도 계수로서 크론바하 알파처럼 많이 사용되면서 그 장단점이 다른 연구자에 의해 빈번히 언급된 통계량도 드물 것이다. Cronbach (2004)의 회상에 의하면 조사의 신뢰도를 언급해야하는 논문에서 2003 년까지 최소한 5,590번 인용되었고 2000년대에도 연평균 약 325번의 사회과학인용빈도 (social science citation)를 기록하고 있다고 한다. 크론바하 본인은 계속적으로 신뢰도 계수가 개선되리라 예상하고 자신의 통계량을 α라 명명했지만 그 뒤로 β, γ 등의 영향력 있는 새로운 계수가 출현하지는 못하였다 (Cronbach, 2004). 크론바하 알파가 그 간결성과 우수한 특성 때문에 신뢰도 계수로 가장 많이 사용되고는 있지만 신뢰도의 과소추정 문제는 계속적으로 지적받아온 사항이다. 본 논문에서는 크론바하 알파가 신뢰도의 하한과 동일하다는 증명 (Novick와 Lewis, 1967; 이기훈, 2008)을 재인용하여 알파의 특성에 관하여 논의하도록 하겠다.신뢰도를 구성하는 식 (3.1)의 진점수 분산은 다음과 같이 정의할 수 있다.

$$\sigma_T^2 = \sum_j^k Var(T_{ij}) + \sum_{j\neq}^k \sum_{j'}^k Cov(T_{ij}, T_{ij'}).$$

여기서, $T_{ij} (i = 1, \dots, n)$는 i번째 평가자의 j번째 항목 평가의 진점수이다.

이때, $\sum_{j=1}^k Var(T_j) \geq \sum\sum_{j\neq j'}^k Cov(T_{ij}, T_{ij'})/(k-1)$ 이므로 진점수 분산의 하한은 다음과 같다.

$$\sigma_T^2 = \sum_j^k Var(T_j) + \sum_{j\neq}^k \sum_{j'}^k Cov(T_{ij}, T_{ij'}) \geq \frac{k}{(k-1)} \sum_{j\neq}^k \sum_{j'}^k Cov(T_{ij}, T_{ij'}).$$

$Cov(T_{ij}, T_{ij'}) = Cov(X_{ij}, X_{ij'})$를 이용하여 신뢰도 공식 (3.2)가 다음과 같이 표현될 수 있다.

$$\rho^2 = \frac{\sigma_T^2}{\sigma_X^2} \geq \frac{k}{(k-1)} \frac{\sum\sum_{j\neq j'} Cov(T_{ij}, T_{ij'})}{\sigma_X^2} = \frac{k}{(k-1)} \frac{\sum\sum_{j\neq j'} Cov(X_{ij}, X_{ij'})}{\sigma_X^2} = \alpha. \tag{3.3}$$

이때 등호는 $Cov(T_{ij}, T_{ij'}) = Var(T_{ij}), Var(T_{ij}) = Var(T_{ij'})$ (모든 j, j')일 때 성립하므로 모든 문항이 동질한 경우에만 크론바하 알파가 신뢰도를 정확히 추정할 수 있고 일반적으로 α는 신뢰도를 과소추정한다고 할 수 있다.

본 논문의 자료에서 크론바하 알파를 구하면 0.938이다. 그런데 식 (3.3)에서 크론바하 알파가 신뢰도의 하한이므로 강의평가문항의 신뢰도가 0.938 이상이라고 추정할 수 없다. 식 (3.3)을 증명하기 위해서 여기에서는 $Cov(T_{ij}, T_{ij'}) = Cov(X_{ij}, X_{ij'})$을 가정하였는데 앞서 언급한 바와 같이 학생들의 평가 중에서 불성실한 답변 즉 동일한 점수로 모든 문항을 체크하는 경우가 다수 발생하고 있다. 그러므로 다음과 같은 식에서 오차항의 공분산이 0이라고 단정할 수 없다.

$$Cov(X_{ij}, X_{ij'}) = Cov(T_{ij}, T_{ij'}) + Cov(e_{ij}, e_{ij'}).$$

즉, $j \neq j'$에 대해서 $Cov(e_{ij}, e_{ij'}) > 0$인 항이 존재하므로 $Cov(T_{ij}, T_{ij'}) = Cov(X_{ij}, X_{ij'})$이 성립하지 않고 $\sum\sum Cov(X_{ij}, X_{ij'}) > \sum\sum Cov(T_{ij}, T_{ij'})$가 성립하게 된다. 그러므로 식 (3.3)의 신뢰도 하한은 다음과 같은 관계를 가진다.

$$Min(\rho^2) = \frac{k}{(k-1)} \frac{\sum\sum_{j \neq j'} Cov(T_{ij}, T_{ij'})}{\sigma_X^2} \leq \frac{k}{(k-1)} \frac{\sum\sum_{j \neq j'} Cov(X_{ij}, X_{ij'})}{\sigma_X^2} = \alpha. \tag{3.4}$$

식 (3.3)에서는 크론바하 알파가 신뢰도를 과소평가함을 증명한데 반하여 식 (3.4)에서는 알파가 신뢰도를 과대평가할 수 있는 가능성을 증명한 셈이다. 그리고 이러한 현상은 응답시간이 부족하여 불성실한 답변으로 앞서의 답변과 동일한 형태로 응답할 때 발생할 수 있기 때문에 강의평가의 신뢰도를 측정할 때 크론바하 알파의 정확성을 보장하기 어렵다. 분석자료에서도 2절에서와 마찬가지로 성실한 응답이라고 판단되는 표준편차가 0.7 이상인 응답케이스만을 뽑아 크론바하 알파를 구한 결과 0.877로 0.06 정도 감소하는 것을 확인할 수 있었다.

또한 크론바하 알파는 문항의 내적일관성에 관한 측도이고 안정성에 관한 측도가 되지 못한다. 강의평가를 활용할 때는 각 문항의 점수가 아닌 10개 문항의 평균점수를 사용하기 때문에 평가자 간에 차이를 보는 안정성의 평가가 필요하다고 할 수 있고 이는 다음 4절에서 분석하기로 한다.

4. 일반화가능도 이론에 의한 신뢰성 검정

Cronbach 등 (1963)이 제안한 일반화가능도 (generalizability) 이론은 평가점수의 분산이 발생하는 다양한 원인에 대해서 분석하고 이에 따라 신뢰성을 평가하는 방법이다. 분산분석 (ANOVA)에 의해 요소별 분산을 검출하고 각 분산성분 (variance component)을 추정하고 분산의 비율로 신뢰도를 추정하는 일반화가능도 분석은 근래 들어 크론바하 신뢰도 계수보다 신뢰성을 파악하는 데 더 우수한 방법으로 인식되고 있다 (Shavelson과 Webb, 1991; Cronbach, 2004).

그러나 이는 아직 평가이론을 실제로 사용하는 심리학, 교육학 실험자에게는 수학적으로 어려운 내용이기 때문에 보편적으로 사용되는데 한계가 있다. 경험적 심리실험분야에서 고전검사이론, 문항반응이론 (item response theory), 일반화가능도 이론 등이 측정점수 평가에 주로 사용되고 있다. 전문가들은 뒤에 두 가지 이론을 추천하지만 아직도 많은 문헌에서 고전검사 이론을 주로 사용하고 있는 이유는 아직 일반화가능도 이론을 편리하게 구현해주는 통계소프트웨어가 존재하지 않기 때문인 것 같다 (Mushquash와 O'Connor, 2006).

고전검사이론은 신뢰성을 측정하는 각양의 방법 (검사-재검사법, 내적일관성, 평가자간 일치성 등)을 제공하고 있지만 측정오차의 원인을 한 가지 (시기, 문항, 검사자 등)로 밖에 파악하지 못하는 단점이 있다. 이 다양한 방법을 종합하는 방법도 개발되어있지 않기 때문에 각 원인의 중요도, 상호작용 그리

고 이들을 어떻게 결합하여야 최적의 믿을 만한 검사법을 구성할 수 있는가에 대한 정보도 얻을 수 없다 (Webb 등, 1988). 일반화가능도 이론은 이에 대한 해결책을 제시하고 있다. 관측된 점수는 여러 가지 다양한 요인에 의해 오차가 발생할 수 있는데 이러한 각 요인에 따른 오차분산을 추정하여 각 요인의 오차비중을 평가하게 해준다. 이는 또한 평가자, 문항수, 평가횟수 등을 변화시켜 신뢰성을 얼마나 변화시킬 수 있는가에 대한 지침도 주게 된다.

일반화가능도 이론에서 p번째 피험자가 t번째 시기 (occasion)에 r번째 평가자 (rater)에게 받은 평가점수의 모형을 다음과 같이 정의한다 (Brennan, 2001).

$$X_{ptr} = \mu + \nu_p + \nu_t + \nu_r + \nu_{pt} + \nu_{pr} + \nu_{tr} + \nu_{ptr}, t = 1, \ldots, n_t; r = 1, \ldots, n_r. \tag{4.1}$$

여기서 μ는 모집단 (population)과 전집 (universe)에서 나온 전체 평균이고, ν는 각 변인들의 효과라 할 수 있다.

일반화가능도 이론에서 개별적인 평가점수는 단지 무한한 모집단과 가능한 측정의 여러 전집에서 나온 표본 (sample)으로 간주한다. 예를 들어, p번째 피험자를 n_r명의 평가자가 n_r번 반복하여 평가점수를 부여하였다면 무한한 피험자의 모집단으로부터 표본을 얻은 것이고 평가자와 평가시기는 수많은 가능한 평가자에서 선택된 평가자와 가능한 여러 평가시기 중에서 선택된 특정시기이기 때문에 가능한 전집에서 뽑은 표본으로 표현되는 것이다.

그러므로 개인의 평가점수는 개인의 진점수 (true score)의 추정값이고 용인된 관측값의 전집 (universe of admissible observation)의 부분이다. 일반화가능도 이론에서는 문항, 평가자, 평가시기 등을 국면 (facet)이라 하는데 통계학의 요인 (factor)과 유사한 개념이지만 무한히 수준을 확장할 수 있다는 점이 다르다. 측정의 개체는 주로 피험자이며 이들의 개인적 차이는 분명한 사실이고 실험자는 이런 차이에 관심을 갖는 것이기 때문에 개체간의 차이는 오차의 요인으로 간주하지 않는다. 식 (4.1)의 모형은 오차에 영향을 주는 국면이 평가자와 평가시기이기 때문에 2국면 설계 (two-facet design) 모형이라 정의한다. 일반적으로 측정의 대상, 즉 학교, 기업, 집단 등은 국면으로 간주되지 않는다. 각 피험자에게 수차례 국면의 수준에서 측정한 값의 평균은 개인의 진점수 (또는 전집점수)의 추정값인데 모든 오차는 국면에 의해 생기는 것이므로 분산성분 (variance components)으로 그들의 영향력을 측정할 수 있다. 식 (4.1)의 피험자 평가점수의 분산성분은 다음과 같다.

$$\sigma^2(X_{ptr}) = \sigma_p^2 + \sigma_t^2 + \sigma_r^2 + \sigma_{pt}^2 + \sigma_{pr}^2 + \sigma_{tr}^2 + \sigma_{ptr}^2. \tag{4.2}$$

일반적인 분산분석에 의해 각 요인, 국면에 대한 평균제곱합 (mean squares)을 구하고 평균제곱합 (expected mean squares)의 기댓값 공식을 이용하여 각 분산성분의 추정값을 구할 수 있다. 실제 자료에서 분산성분을 추정하는 연구를 일반화가능도 연구, 즉 G 연구 (G study)라 하며 이러한 결과는 국면에 변화가 있을 때 각 오차성분이 어떻게 변화하는지를 분석하고 최적의 실험설계를 결정하는 D 연구 (decision study)에 활용된다 (Cronbach 등, 1972). D 연구를 위하여 식 (4.1)의 피험자 평가점수를 각 국면에 대한 기댓값으로 다음과 같이 바꾸어 표현하도록 한다.

$$X_{pTR} = \mu + \nu_p + \nu_T + \nu_R + \nu_{pT} + \nu_{pR} + \nu_{TR} + \nu_{pTR}. \tag{4.3}$$

그리고 각 분산 성분은 다음과 같이 표현된다.

$$\sigma^2(X_{pTR}) = \sigma_p^2 + \sigma_T^2 + \sigma_R^2 + \sigma_{pT}^2 + \sigma_{pR}^2 + \sigma_{TR}^2 + \sigma_{pTR}^2.$$

이때 각 분산성분의 추정값은 식 (4.2)의 각 성분의 분산 추정값을 이용하여 구하게 된다. 즉, $\widehat{\sigma}_T^2 = \widehat{\sigma}_t^2/n_t$, $\widehat{\sigma}_R^2 = \widehat{\sigma}_r^2/n_r$, $\widehat{\sigma}_{pT}^2 = \widehat{\sigma}_{pt}^2/n_t$, $\widehat{\sigma}_{pR}^2 = \widehat{\sigma}_{pr}^2/n_r$, $\widehat{\sigma}_{TR}^2 = \widehat{\sigma}_{tr}^2/n_t n_r$, $\widehat{\sigma}_{pTR}^2 = \widehat{\sigma}_{ptr}^2/n_t n_r$.

D 연구를 위하여 사용할 일반화가능도 계수 (G 계수)는 전집점수분산을 관측점수분산으로 나눈 것이다. G 계수가 크면 실험자는 관측점수를 연구국면에 일반화할 수 있다는 의미로 일반화가능도 이론이라 명명한 것이다. G 계수는 상대 G 계수와 절대 G 계수로 분류할 수 있다. 상대 G 계수는 원자료의 가능한 변화를 고려하지 않고 단지 측정개체의 국면의 변화에 따른 정도만을 반영하고 있어 고전 검사이론의 신뢰성 계수와 유사하여 일반적으로 이 값을 일반화가능도 계수 (generalizability coefficient)라 부른다. 절대 G 계수는 좀 더 엄격하게 원자료의 가능한 변화까지 반영한 것으로 운전면허시험, 정신병 진단평가의 결과처럼 피험자에게 의미가 중대할 때는 절대 G 계수를 사용한다. 절대 G 계수는 파이계수 또는 의존계수 지수 (index of dependability coefficients)라 부르기도 한다. 상대 G계수의 분모에는 피험자에게 해당하는 분산성분만 포함하고 절대 G 계수에서는 모든 분산성분이 포함된다. 국면이 모두 변량 국면 (random facet)일 때 식 (4.3) 점수의 G 계수는 다음과 같이 정의 된다.

일반화가능도 계수 (상대 G 계수): $E\rho^2 = \dfrac{\sigma_p^2}{\sigma_p^2 + \sigma^2(\delta)}$,

의존계수 지수 (절대 G 계수): $\Phi = \dfrac{\sigma_p^2}{\sigma_p^2 + \sigma^2(\Delta)}$,

여기서, $\sigma^2(\delta) = \sigma_{pT}^2 + \sigma_{pR}^2 + \sigma_{pTR}^2$, $\sigma^2(\Delta) = \sigma_T^2 + \sigma_R^2 + \sigma_{pT}^2 + \sigma_{pR}^2 + \sigma_{TR}^2 + \sigma_{pTR}^2$.

앞서 설명한 일반화가능도 이론을 전주대학교 영어교양교과목 강의평가 자료에 적용하도록 하겠다. 자료는 8명의 영어교수가 동일한 교과목의 분반인 4반을 담당하고 각반의 수강생이 20명이기 때문에 총 640명의 케이스로 구성되어 있다. 즉 평가시기 (t)국면이 $n_t = 4$, 평가자 (r)국면이 $n_r = 20$인 2국면 설계 모형인데 이들이 서로 교차 (crossed)되어 있지 않고 피험자 (강의교수)아래 교과목 ($t : p$), 교과목 아래 평가자 ($r : t$)들이 지분 (nested)되어 있다. 그러므로 평가점수의 모형은 다음과 같이 표현할 수 있다.

$$X_{ptr} = \mu + \nu_p + \nu_{t:p} + \nu_{r:t:p}, t = 1, \ldots, 4; r = 1, \ldots, 20. \tag{4.4}$$

이들 자료를 분산분석하면 다음 표 4.1과 같은 분산분석표를 얻을 수 있다. 교수간의 차이에 의해 발생하는 분산의 추정값이 0.0084 (=(1.224-0.555)/80)이고, 동일한 교수의 분반 차이에서 생기는 분산은 0.0017 (=(0.555-0.521)/20)로 작은 것을 알 수 있다. 같은 교수가 시간을 달리하여 강의하였을 때 발생하는 분산이 상대적으로 적은 것은 평가가 안정적임을 의미하고 있다. 그러나 평가자간에 분산이 0.521로 상대적으로 큰 값이고, 이를 피험자간의 분산과 절대적 비교를 위해 표준화한 값도 0.0065 (=0.521/80)로 학생들의 평가의 편차가 상당함을 짐작할 수 있다.

표 4.1 강의평가 점수의 분산분석표

	df	SS	MS	EMS	$\widehat{\sigma}^2$	비율
p	7.000	8.568	1.224	$\sigma_{r:t:p}^2 + n_r\sigma_{t:p}^2 + n_r n_t \sigma_p^2$	.0084	.016
t:p	24.000	13.322	.555	$\sigma_{r:t:p}^2 + n_r\sigma_{t:p}^2$	.0017	.003
r:t:p	608.000	316.525	.521	$\sigma_{r:t:p}^2$	.521	.981

D 연구를 위하여 G 계수를 다음과 같이 추정한다. 평가점수의 모형이 식 (4.4)이고 모든 분산 성분에 p가 포함되어 있으므로 $\sigma^2(\delta)$ 와 $\sigma^2(\Delta)$의 추정값은 동일 (0.0069)하고 따라서 절대 G 계수와 상대 G 계수의 값도 동일하다.

$$E\rho^2 = \Phi = \frac{0.0084}{0.0084 + 0.0017/4 + 0.521/80} = \frac{0.0084}{0.0084 + 0.0069} = 0.55.$$

일반화가능도 계수의 값이 얼마 이상이어야 신뢰도가 높다고 할 수 있는가에 대한 기준은 존재하지 않지만 0.6 이상이면 만족할 만한 수준이라 할 수 있다 (Brennan, 2001). 본 논문에서 분석하는 자료에서 일반화가능도 계수가 국면을 조정함에 따라 어떻게 달라지는가를 표 4.2에 표시하였다. 교과목 (평가시기) 국면 수를 증가시키는 것보다 평가자의 수를 증가시키는 것이 현실적으로 더 실현가능하다는 점을 착안하면 현재의 분반수 ($n_t = 4$)에서 평가자의 수를 30명으로 늘린다면 0.6 이상의 일반화가능도 계수를 확보할 수 있을 것으로 추정된다. 현재 전주대학교는 수강생 수가 20명 이상인 경우에만 강의평가 결과를 업적평가에 적용하는데 이 숫자를 30명으로 늘리는 방식도 고려해볼만 하겠다. 각 국면의 변화에 따른 계수 값의 변화는 그림 4.1에 도시되어 있다.

표 4.2 국면의 변화에 따른 일반화가능도 계수

n_r	$n_t = 2$	$n_t = 4$	$n_t = 6$	$n_t = 8$
10	0.237	0.383	0.483	0.554
20	0.376	0.546	0.644	0.707
30	0.467	0.637	0.724	0.778
40	0.532	0.694	0.773	0.819

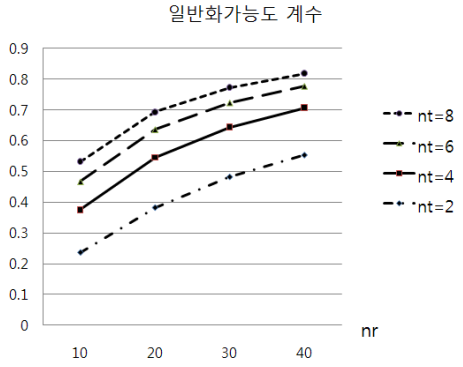

그림 4.1 일반화가능도 계수 변화

5. 결론

우리나라 대학에서 강의평가를 실시한지 상당한 시간이 지나는 동안 대학 내부적으로는 강의평가의 적절함에 대한 많은 논란이 있었지만 학문적인 접근이나 논의는 그리 많지 않은 편이다. 그리고 몇 안 되는 강의평가 관련 논문들의 대부분이 강의평가의 단점을 보완해주는 형식, 즉, 외생변수의 통제, 공평함을 위한 보정 등에 초점이 맞추어져 있었다. 최근 들어 강의평가가 대학교원의 업적평가에 직접적인 영향을 미치게 되자 강의평가를 정교화하고 공정하게 실시하는 문제보다 좀 더 근원적으로 강의평가의 신뢰성에 대한 의문이 제기되고 있다. 학생들의 평가를 믿을 수 있는가에 관한 학생들의 자질문제 뿐만 아니라, 성실성, 진실성, 일관성 등 많은 문제점이 제기되어 왔지만 이에 대한 활발한 논의는 이루어지지 않는 실정이다.

본 논문에서 전주대학교 특정 교과목에 대한 강의평가 자료를 분석한 결과 강의평가의 타당성이 높은

편이고 신뢰도도 어느 정도 만족할 만한 수준임을 확인할 수 있었다. 그러나 동일한 강의를 평가하는 학생들의 개인 차이에 의한 오차가 크기 때문에 강의평가가 신뢰성을 확보하려면 평가인원을 확대할 필요가 있다. 즉, 행정적 기능을 위한 강의평가 활용은 수강생이 적정 수 이상인 경우에 한해야 함을 의미한다.

일반화가능도 이론을 단순하게 설명하기 위해서 본 논문의 자료는 동일한 교과목에 수강생이 동일하게 20명인 반으로 제한하였으나 각 반의 수강생 수가 다르더라도 각 분산 성분의 추정이 가능할 것이다. 그러나 교과목을 다양화하는 문제는 이론적으로 매우 복잡한 문제를 포함하게 된다. 향후 여러 교과목을 포함하여 교수와 과목, 수강생들이 일부 교차하는 지분법 모형으로 대학에 개설되어 있는 전체 자료를 이용하여 신뢰성을 평가하는 방법의 연구가 필요한 것으로 사료된다.

참고문헌

김성연, 권치명 (2005). 통계적 기법을 활용한 균등화법에 의한 강의평가 개선방안 연구. <한국자료분석학회지>, **7**, 1705-1721.

김영진 (1994). 교수강의평가제-과연 생산성이 있는가. <생산성논집>, **8**, 252-235.

김종태 (2004). A study of reliability of lecture evaluation by students. <한국데이터정보학회지>, **15**, 183-191.

류춘호, 이정호 (2003). 대학의 강의평가에 영향을 미치는 학생관련 요인에 관한 연구. <경영학연구>, **32**, 789-807.

박노진 (2009). 핵심 문항들을 활용한 모델링-강의 평가 자료를 활용한 사례연구. <한국데이터정보학회지>, **20**, 1075-1083.

성태제 (2002). <타당도와 신뢰도>, 학지사, 서울.

이기훈 (2008). 크론바하 신뢰도 계수에 관한 이해. <산경논총>, **28**, 43-54.

이종승 (1995). <교육연구법>, 배영사, 서울.

조장식, 강창완, 최승배 (2009). 강의평가에 대한 균등화방법의 비교. <한국데이터정보학회지>, **20**, 65-75.

한신일 (2002). 강좌규모와 강의평가의 관계분석. <고등교육연구>, **13**, 155-173.

Ahmadi, M., Helms, M. and Ralszadeh, F. (2001). Business students' perceptions of faculty evaluations. *The International Journal of Educational Management*, **15**, 12-22.

American educational research association, American psychological association and national council on measurement in education (AERA, APA and NCME) (1999). *Standard for educational and psychological testing*, American Psychological Association, Washington D. C..

Baik, T. and Yang, G. (2008). Classroom lecture monitoring case study. *Journal of the Korean Data & Information Science Society*, **19**, 1191-1200.

Brennan, R. L. (2001). *Generalizability theory,* Springer, New York.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, **16**, 297-334.

Cronbach, L. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement*, **64**, 391-418.

Cronbach, L. J., Gleser, G. C., Nanda, H. and Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*, Wiley, New York.

Cronbach, L. J., Rajaratnam, N. and Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *The British Journal of Statistical Psychology*, **16**, 137-163.

Crumbley, D. L. and Fliedner, E.(2002). Accounting administrators' perceptions of student evaluation of teaching(SET) information. *Quality Assurance in Education*, **10**, 213-222.

Kuder, G. F. and Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, **2**, 151-160.

Kulik, J. A. and McKeachie, W. J. (1975). The evaluation of teachers in higher education. *Review of Research in Education*, **3**, 210-230.

Lee, K. H. and Lee, S. W. (2006). A study on controlling the external effect in student evaluation of testing. *The Korean Communications in Statistics*, **12**, 589-601.

Linn, R. L. (1997). Evaluating the validity of assessments: The consequences of use. *Educational Measurement: Issue and Practice*, **16**, 14-15.

Linn, R. L. and Gronlund, N. E. (2000). *Measurement and assessment in teaching, 8th Ed.*, Upper Saddle River, NJ: Merrill.

Mehrens, W. A. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice*, **16**, 16-18.
Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research*, **45**, 35-44.
Mushquash, C. and O'Connor, B. P. (2006). SPSS and SAS programs for generalizability theory analyses. *Behavior Research Methods*, **38**, 542-547.
Novick, M. R. and Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika*, **32**, 1. 1-13.
Popham, W. J. (1997). Consequential validity: Right concern-wrong concept. *Educational Measurement: Issues and Practice*, **16**, 9-13.
Shavelson, R. J. and Webb, N. M. (1991). *Generalizability theory: A primer*, Sage, Newbury Park, CA.
Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, **16**, 13-24.
Spearman, C. (1910). Correlation calculated with faulty data. *British Journal of Psychology*, **3**, 271-295.
Webb, N. M., Rowley, G. L. and Shavelson, R. J. (1988). Using generalizability theory in counseling and development. *Measurement & Evaluation in Counseling & Development*, **21**, 81-90.

Journal of the Korean Data & Information Science Society
2010, **21**(1), 87–98

한국데이터정보과학회지

A study on validity and reliability of students' evaluation

Ki Hoon Lee[1]

[1]School of Business, Jeonju University

Received 12 December 2009, revised 14 January 2010, accepted 18 January 2010

Abstract

This research deals the method to assess the validity and reliability of students' evaluation for lectures. Most papers for student's evaluation have focused the procedures for controlling the external effects, but this paper is trying to answer for "How reliable is the student rating?" An empirical study shows that the evaluations in Jeonju University have the fair validity and reliability. The generalizability theory is suggested to obtain the more comprehensive results rather than Cronbach's alpha to examine internal consistency.

Keywords: Generalizability, reliability, students' evaluation, validity.

[1] Professor, School of Business, Jeonju University, Jeonju, Jeonbuk 560-759, Korea.
E-mail: khlee@jj.ac.kr