한국통계학회논문집 제1권1호
1994년 12월 pp. 226-231

토론:

김 우철 [1], 김 성호 [2]

1. "대학별 고사를 위한 문항분석, 표준점수, 검사 동등화" 에 대하여

이 논문은 검사의 평가와 동등화에 관한 기본적 이론을 일목요연하게 소개하고 있다. 전반적으로 통계적 개념이 교육평가 분야에서 어떤 의미로 해석되고 있는가를 알기쉽게 전달해주고 있으며, 적절한 참고문헌과 패키지 프로그램이 소개되어 있어 통계인들이 이 분야에 접근하는 데에 많은 도움이 되리라 생각한다. 성 태제 교수의 노고에 감사드리며, 이 분야의 문외한으로서 갖게된 몇가지 의문사항으로 토론에 대신하고자 한다.

첫째로, 검사의 타당도에 관한 내용중에 내용타당도는 이원분류표를 이용한 "전문가의 주관적 판단" 이라 하였다. 그러나 이에 대한 "객관적 방법" 도 가능한 것이 아닌지? 토론자[2] 에 의하면 인과분석(causal)적 방법으로 로그선형모형 (log linear model) 하에서 이러한 시도가 가능하다고 하는데(Kim,S.H. and Wang,M.(1994) Influence Diagram of Test Performance in GRE-Q. ETS Research Report(in review)), 이러한 객관적 방법의 시도에 대한 저자의 의견은 어떤지?

둘째로, 문항반응이론에 관한 내용의 소개에서 능력수준이 주어진 상태에서의 반응에 대한 로지스틱 회귀모형은 분명하게 주어졌으나 능력수준에 대한 모형이 제시되지 않아서, 이의 이해에 장애가 된다고 생각된다. 이보다 더욱 본질적인 의문사항은 왜 능력수준이 하나의 수 값으로, 즉 종합적인 능력으로만 취급되어야 하는 가이다. 실제로 한 검사지에서 측정하고자 하는 능력은 θ_1, ..., θ_k와 같이 여러가지일 것으로 생각되며, 능력 (θ_1, ..., θ_k) 와 (θ_1', ..., θ_k')은 서로 다르며 종합적 능력은 같을 수도 있을 것이다. 이러한 경우에 각 문항별로는 개별적 능력 (θ_1, ..., θ_k)와 (θ_1', ..., θ_k')을 구별하는 것이 필요한 것은 아닌지? 예를 들면, 로지스틱 다중회귀모형이나 또는 로지스틱 일반화 가법모형 (logistic generalized additive model) 하의 분석은 이 분야에서 시도되고 있지 않는 것인지? 만약 이러한 시도가 있다면 문항의 난이도와 변별도는 어떠한 방식으로 스칼라 (scalar) 화 하는지?

세째로, 차별기능문항의 추출방법으로서 Raju 의 방법과 Mantel-Hanzsel 방법의 비교에 대한 소개를 하고 있는데, Raju 의 방법은 로지스틱회귀모형이 타당한 경우의 방법이고 Mantel-Hanszel 방법은 그러한 모형의 가정이 없이 적용가능한 방법인데 보고된 비교연구는 여타 회귀모형하에서도 비교한 것인지?

마지막으로 검사의 동등화에 대하여는 "각기 다른 과목을 선택한 피험자를 동일한 척도에 의하여 평가하는 것은 교육평가이론에 위배된다" 는 저자의 의견에 전적으로 찬성하며, 이러한 동등화의 분석은 출제의 계획단계에서 출제자를 돕기위한 정보제공 목적으로 국한시켜야 한다는 것이 토론자[1] 의 의견이다.

2. "새 대학입시의 통계적 계획과 분석" 에 대하여

이 연구는 문항분석을 비롯한 모의고사의 평가를 통하여 본고사 출제를 위한 양질의 정보제공과 등화방법의 모색이라는 두 가지 목적을 위한 것으로 생각된다. 우선 이와 같이 세심하게 준비

1) 서울대학교 자연과학대학 계산통계학과.
2) 한국과학기술원 기초과학과.

되고 치밀하게 수행된 좋은 사례연구를 수행한 허 명회 교수와 이를 가능하도록 한 고려대학교의 배려를 높게 평가하고 싶다. 이와 같은 연구의 무경험자로서 연구보고서를 읽으며 갖게된 몇가지 의문사항을 정리하여 보고자 한다.

첫째, 본 연구에서의 한 평가지표인 고전적 검사이론의 변별도는 문항반응이론에서의 변별도와 어떤 관계에 있는 것인지? 두 변별도는 통계적으로 같게 해석하여도 되는 것인지? 이와 관련하여 과목변별도로는 사분위수범위를 사용하였는데, 앞의 질문에서의 변별도와는 같은 의미인지?

둘째, 내적합치도를 높이기 위하는 한 방법으로 "문항간의 곤란도 및 변별도를 가급적 균일하게 조율하는 것"을 제안하였는데, 이 표현에서의 변별도는 각 문항에서 득점의 산포를 뜻하는 것인지? 만약 여기에서의 변별도가 상관계수의 의미로서의 변별도라면 이는 균일하게 할 것이 아니라 상관이 크도록 해야 하는 것이 아닌지? 또한, "문항수를 늘리는 것"을 제안하였는 데, 이는 문항간의 상관계수를 높이는 것을 간접적으로 표현한 것이 아닌지? 이와 관련하여서는 여러가지의 능력을 한 검사지에서 평가할 때, 각 문항이 가능한 한 많은 능력을 물음으로써 문항과 문항간의 관계를 높임으로써 내적 합치도를 증가시킬 수 있는 것이 아닌지?

세째, 출제지침의 하나로서 "난이도는 정답률 또는 평균 (1 점 만점 환산점수) 이 0.5~0.6 이상이 되도록 하여야 한다."고 하였는데, 오히려 하나의 검사지에는 문항별로 난이도가 골고루 퍼져 있어야 하는 것이 아닌지? 정답률이 0.5 이상인 문항으로만 구성되면, 총점의 분포가 높은 학생쪽으로 몰리는 현상이 생기지 않는지? 특히, 이러한 제안에 따르면 중요한 능력을 묻는 문항이 제외되는 것이 아닌지?

네째, 등화의 과정에서 비교가능집단이라 함은 일반적 학업능력이 비슷한 집단을 뜻하는 것으로 생각되는 데, 이를 본고사 실시 이전에 내신 1, 2, 3 등급으로 국한시킨 경우에 사후 확인과 조정은 전혀 고려하지 않는 것인지? 예를 들어, 필수과목의 성적등으로 동질성에 대한 검토가 없어도 무방한지? 또한, 지나치게 일반적 학업능력이 낮은 수험생들로 인한 선택과목별 집단의 점수분포에 대한 왜곡현상은 없었는지? 만약 있었다면 그에 대한 대처방안은 무엇이었는지?

마지막으로, 등화의 과정에서 자연계의 경우에 물리성적과 생물성적을 등화하는 과정에서 수학성적에 대한 회귀부분 (regression part) 을 제거한 후의 등위 (relative standing) 를 이용하는 것에 대한 의견은 어떤한 지? 이러한 방법의 실제 적용은 현실적으로 수용되기 어렵옵겠지만, 필수과목으로 평가된 능력을 제거하고 추가적인 능력에 대한 평가가 선택과목의 목적이라면 이것이 더 합리적이 아닌지?

토론: "새 대학입시의 통계적 계획과 분석"에 대하여

임 형 3)

허명회교수께서 "새 대학입시의 통계적 계획과 분석"을 주제로 문항분석과 선택과목의 표준점수 문제를 논의하셨다. 문항분석은 고전검사이론을 사용하고 선택과목의 표준점수화는 등사분위수 동등화를 사용하였다.

대학입시상황에서 선택과목의 난이도 문제와 '94학년도를 위한 2번의 대학수학능력시험의 난이도 문제는 다르다. 첫째, 전자는 다른 특성을 측정하는 시험의 난이도가 다른 경우이다. 다른 집단의 피험자가 다른 특성을 측정하는 한 검사 혹은 두 검사를 치를 때 과목간의 난이도 차이에 따른 문제이다. 둘째, 후자는 동일한 특성을 측정하는 여러가지 유형의 시험에서 난이도가 다른 경우이다. 다른 집단의 피험자가 동일한 특성을 측정하는 여러 유형의 시험을 칠 때 난이도에 따른 불공정성을 조정하기 위하여 점수의 변환이 필요하다. TOEFL, SAT, GRE, 수학능력시험 등이 여기에 속한다. 지금까지 대부분의 연구(Angoff, 1971; Braun and Holland, 1982; Angoff, 1982; 남현우, 1992; 황소림, 1993)는 동일한 특성을 측정하는 검사의 동등화 문제를 다루고 있고 동등화 가정에서도 이것이 기술되어 있다.

대학입시상황에서 4개의 선택과목 시험은 서로 다른 피험자 집단(P,Q,R,S)에 실시되어서 동등화 설계는 표 1과 같다. 내신성적, 수학능력시험 점수, 공통과목 점수를 가교검사점수로 가정하면 동등화 설계는 표 2와 같다. 허교수 논문에서는 선택과목에 영향을 주는 요인으로 내신등급이 고려되어서, 비교가능집단 즉 동등집단(P1, Q1, R1, S1)을 인문계와 자연계의 안암캠퍼스에 소재한 학과를 1지망으로 하는 일반고교 출신의 응시생으로 설정하였다(표 3). 비교가능집단을 일반고등학교를 졸업한 내신 1, 2, 3등급의 피험자로 결정했다. 현실적으로 각 고등학교의 수준차이를 감안할 때, 비교가능집단을 결정하는 준거로 내신성적만을 사용하는 것에는 좀 더 깊은 고려가 필요할 것으로 생각된다. 또 고3을 지도하는 몇몇 선생님의 의견을 참조하면 1등급과 2등급은 동등집단으로 분류하나, 3등급까지를 동등집단으로 분류하지는 않았다. 따라서 동등집단을 선정하는 기준을 정할 때, 각 피험자의 내신성적과 더불어 수학능력시험 점수와 본고사의 공통점수가 유용하므로 이 세가지 점수를 모두 함께 사용하면 보다 동등한 비교가능집단을 설정하는데 도움이 될 것으로 생각한다.

비교가능집단이 설정되면 이 집단의 선택과목점수를 백분위수 동등화와 선형동등화 방법을 혼합한 등사분위수 동등화로 변환했다. 백분위수 동등화와 선형동등화 방법은 고전검사이론을 이용한 동등화방법으로 특별한 가정이 필요없다. 백분위수 동등화는 표집의 크기가 동등화 결과의 안정성에 영향을 주며 선형동등화는 점수분포의 모양이 동등하게 영향을 준다. 그러므로 표집집단이 크고 검사점수의 분포모양이 유사하면 등사분위수 동등화 방법은 비동등집단 설계를 위하여 안정적인 결과를 보여줄 것으로 생각된다. 결론적으로 대학입시 상황에서 처럼 선택과목간의 난이도 차이가 존재하여 과목선택에 따른 우연한 이익 또는 불이익이 존재하지 않도록 과목점수의 동등화는 적용되어야 할 것이다. 끝으로 동등화에 대한 필요성과 그 효율성에 대한 인식이 높아지기를 기대하며 원점수에 고착되어있는 우리의 사고방식에 전환이 있기를 기대한다.

문항분석방법은 고전검사이론을 이용하여 난이도, 변별도, 신뢰도를 구하고 문항의 양호도는 난이도와 변별도 측면에서 고려하였다. 변별도는 문항과 총점간의 상관계수, 신뢰도는 Cronbach-알파로 계산되었다. 문항반응이론을 사용하면 문항의 난이도, 변별도, 추측도, 문항정보함수와 원점

3) 국립교육평가원 교수

수에 대응하는 피험자의 능력모수치가 추정된다. 그리고 고전검사이론의 신뢰도에 대응하는 검사정보함수가 구해진다. 검사정보함수는 검사의 난이도가 피험자의 능력수준과 동일할 때 최고값을 가져서 개발된 문항들이 피험자의 능력수준에 알맞는지 판단하는데 도움을 준다. 특히 다음해 대학별고사를 제작할 때, 고전검사이론과 문항반응이론으로 문항분석한 결과를 함께 이용하면 해당 대학에 지원하는 피험자의 능력범위에 해당하는 난이도 수준의 문항과 변별도가 높은 문항으로 구성된 검사를 개발하는데 도움이 될 것이다.

표 1. 비동등집단 설계

표 본	검사			
	W	X	Y	Z
P	*			
Q		*		
R			*	
S				*

표 2. 가교검사-비동등집단설계

표 본	검 사						
	W	X	Y	Z	수능	내신	공통
P	*				*	*	*
Q		*			*	*	*
R			*		*	*	*
S				*	*	*	*

표 3. 동등집단 설계

표 본	검사			
	W	X	Y	Z
P1	*			
Q1		*		
R1			*	
S1				*

참고문헌 (추가)

[1] 남현우(1992) "문항모수변이에 따른 선형, 동백분위, IRT, 검사동등화방법의 강인성 비교연구". 교육평가연구, 5권 2호, 27-60.

답신: 김우철 · 김성호 교수의 질문에 대하여

성 태 제

첫째 질문인 내용타당도를 위한 log linear model의 적용가능성에 대하여: 내용타당도는 검사내용을 논리적 사고에 입각하여 논리적 분석의 과정으로 판단하는 주관적 타당도이다. log linear model을 타당도에 적용하려면 구인타당도를 검정하는 데 사용될 지 모르나 측정분야에서 구인타당도를 증명하는 보편적인 방법은 아니라고 생각한다.

둘째 질문인 문항반응이론에서 능력이 왜 하나로 측정되어야 하는가에 대하여: 문항반응이론은 일차원성(unidimensionality) 가정을 전제로 한다. 즉, 한 검사는 단일한 능력측성을 측정하여야 함을 전제로 한다. 그러나 어떤 질문지가 여러 개의 특성을 측정한다면 문항반응이론은 적용되지 않는다. 이를 위하여 현재 다차원 문항반응이론(multi-dimensional item response theory)이 개발중이다.

세째 질문인 Raju 방법은 문항특성곡선을 이용한 방법이고, Mantel-Hanszel 방법은 가정없이 적용가능한 방법인데 보고된 비교연구에서 회귀모형에 대한 비교가 없었는지에 대하여: Mantel-Hanszel 방법은 χ^2 통계를 이용하여 승산비(odds ratio)를 계산하는 방법이다. 차별기능문항을 추출하는 방법으로 Swaminathan and Rogers(1990)의 로지스틱 회귀 방법이 있으며, 이 방법의 타당성 검정을 위한 다수의 연구가 있다.

참고문헌 (추가)

[1] Swaminathan, H. and Rogers, H. J.(1990) Detecting differential item function using logistic regression procedures. *Journal of Educational Measurement,* 27, 361-370.

답신:

허 명 회

답신에 앞서 토론자를 비롯하여 심포지움에 참여하신 모든 분들께 감사드린다. 입시는 우리나라에 있어 사회적으로 매우 중요한 문제인데 이제부터는 그동안 등한시되어 왔던 통계적 측면에서 이 문제에 관하여 충분히 연구되면 좋겠다는 생각이다.

김우철 · 김성호 교수의 질문에 대하여 충분하지는 못하겠지만 나름대로 응답해보기로 하겠다. 첫째 질문은 문항 변별도에 대하여 고전검사이론에서의 변별도(A1)와 문항반응이론에서의 변별도(A2)의 관련성(또는 1:1 대응성) 및 과목 변별도(B) 등에 대한 것이다. 우선 고전검사이론과 문항반응이론이 각각 다른 검사측정이론 체계이므로 A1과 A2가 엄밀한 1:1 대응관계에 있다고 생각되지는 않는다. 전자가 묵시적으로 선형적 정규모형을 바탕으로 하는 반면 후자는 명백히 로지스틱 이항모형을 바탕으로 하고 있지 않은가? A1과 B를 본문에서 구별하지 않은 것은 문맥상 혼동이 없기 때문일 뿐이다. 이들 역시 고전검사이론의 틀안에서는 같은 맥락으로 이해할 수 있다고 본다. 둘째 질문은 내적 합치도를 높히는 방안에 관한 것이다. 문항간의 상관도를 높힘으로써 내적 상관도가 높아지는 것은 잘 알려진 사실이다. 그러나 실제 그것을 높히는 것은 출제기술상의 문제가 되는데, 본 발표자의 의도는 문항 변별도를 들쭉 날쭉하게 하는 것보다는 가급적 균일하게

하는 것이 내적 합치도를 높히게 된다는 것이었으므로 오해가 없기를 바란다. 세째 질문은 난이도를 0.5-0.6 이상이 되도록 해야 한다는 출제지침이 별 근거가 없지 않나하는 것이었다. 추측이 불가능한 문항의 경우 난이도가 0.2 이하인 문항과 0.8 이상인 문항은 같은 변별능력 또는 문항정보를 갖는다. 그러나 대입 본고사가 어려워질수록 "과외 망국론" 등 대학에 대한 사회적 비판이 강하게 되고 또한 고교의 정상적 교육을 파행으로 유도하는 등 많은 병폐가 생긴다. 따라서 구태어 난이도가 0.2 이하인 문항을 내는 것은 대학의 위신을 거들먹거리는데 잠시 도움이 될 뿐 사회적 측면에서 바람직하지 않다고 생각된다. 네째 질문은 등화시 선발된 비교가능집단이 내신 외에 수능이라든가 본고사 공통시험에서 어떤 차이를 보이지 않았는가에 대한 우려이다. 본 발표자가 분석한 바로는 어떤 체계적 차이를 보인 적은 없었다는 것이다. 이에 관하여는 모의고사 종합결과 보고서(전성연 · 허명회, 1993, 고려대학교 입시출제위원회)에 보고되어 있다. 물론 약간의 차이는 어쩔 수 없는 것이라고 본다. 그런 경우라고 하더라도 실제입시에서는 입시관리실무자가 미리 정하여진 것이 아니면 어떤 통계적 조정도 할 수 없다고 본다. 마지막 질문은 등화 방법에 관한 새로운 제안이다. 향후 검토해 보겠다. 예상되는 문제점은 각 선택과목이 공통과목과 차별적인 상관성을 보인다는 사실이다. 예컨대 물리는 수학과 상관성이 높고 생물은 수학과 상관성이 낮게 나온다. 조목조목 질문해주신 김우철 · 김성호 교수께 다시 한 번 감사드린다. 앞으로 수리적으로 엄밀한 바탕에서 현실문제에 통계적으로 응용할 것을 권한 것으로 받아들이겠다.

임 형 교수는 선택과목 등화에 대하여 그 필요성을 인정하고 발표자의 연구내용을 보완하였으며 고무적인 지지를 해주셨다. 또한 비교가능집단의 선발에 있어 내신 외에 수학능력시험 및 본고사 공통과목을 이용하도록 권고하였다. 향후 검토하여 보겠다. 이 중에서 본고사 공통과목을 이용하는 방법보다는 수학능력시험을 이용하는 방법이 더 나을 것 같다는 생각이다. 왜냐하면 수능이 수험생의 능력을 일부과목에 국한하지 않고 보다 넓은 영역에서 평가하기 때문이다. 또한 임 형 교수는 본고사의 문항분석에 있어 고전검사이론 뿐만 아니라 문항반응이론에서의 방법론을 이용할 것을 제안하였다. 그러나 문항반응 방법론을 위한 소프트웨어로서 현재 널리 쓰이는 BILOG 가지고는 주관식 문항을 분석할 수 없다는 것이 문제점이다.

마지막으로 선택시험의 등화에 대한 본 심포지움의 일부 논의에 대하여 본 발표자의 주관적 의견을 개진하기로 하겠다. 선택시험의 등화가 서로 다른 차원의 자(尺)를 같게 하려는 것임이 사실이다. 그러나 이것을 제한된 교육측정학적 맥락에서만 보는 것은 적절하지 않다. 선택시험 등화는 일종의 경쟁적 게임에서 꼭 있어야할 규칙과 같은 것이다. 우리나라에서의 대학입시는 대단한 이해관계가 걸려있는 사회적 제도이다. 이에 있어 어느 한 선택과목이 매우 쉬어서 그 과목을 선택한 학생들이 80% 합격한 반면 다른 한 과목을 선택한 학생들은 20%만 합격하였다면 이것을 무엇으로 정당화시킬 수 있을 것인가? 수험생의 재수 탓으로 돌릴 것인가 (선택을 잘못한 죄)? 상대적 난이도가 균일하도록 선택시험 출제를 잘 하면 된다는 생각은 어리석은 낙관론에 지나지 않는다. 출제교수가 그렇게 많은 출제경험을 갖고 있지 않다는 것이 본 발표자의 판단이다. 선택과목을 두는 것이 국영수 위주의 본고사가 갖는 고교교육의 왜곡화를 막는 한 방편이고 또한 입학생의 다양성을 위해서도 도움이 된다. 따라서, 대학 본고사에서 선택과목을 두는 이상 공정한 규칙의 설정은 제도적으로 필수불가결하다고 생각한다.