韓國數學教育學會誌 시리즈 A <數學教育>
1999. 11. 제 38권, 제 2호, 165-172.

J. Korea Soc. Math. Ed. Ser. A: The Mathematical Education
Nov. 1999, Vol. 38, No. 2, 165-172.

논문형 고사 평가에서 평가치 조정과 평가원의 신뢰도 향상에 유효한 CDM 모형의 응용*

홍 석 강 (동국대학교)

Ⅰ. 서 론

시험 도구의 출제 방법과 평가에 관한 관점에서 볼 때 학습평가를 위한 검사 도구의 제작 또는 선택에서는 두 가지 입장에서의 검토가 가능할 것이다. 첫째, 측정관의 입장에서 볼 때는 그 결과를 주로 선발, 분류, 예언, 실험에 이용하는 일이 중심이 될 것이고, 둘째, 평가관의 입장에서 본다면 학습평가의 목적이나 기능은 학습자 개인을 판단하고 분류하는 기능을 넘어서서 교육과정, 교수방법, 교수프로그램, 수업의 과정, 교사의 효능도 등 교수-학습 전반을 개선하는 데 초점을 둔다. 여기서 논의의 폭을 좁혀 시험도구의 경우로 한정하여 전자의 입장에서 보면 신뢰도와 객관도를 중시하여 선택형 출제방법이 적절할 것이고, 후자의 입장에서 본다면 수험자들 개개인의 학습능력, 학습태도 등을 쉽게 변별할 수 있는 서답형 출제방법이 적절할 것이다. 이 논문에서는 그 관점을 후자의 입장에서 본 평가 문항, 즉 수험자 반응의 자유도가 크고 고등정신 기능을 측정하는 데에 유용하며 문항 제작이 쉽고 학생들의 학습태도를 개선하여 주는 데는 효과적이나, 채점이 비객관적이며 신뢰롭지 못하고, 문항의 표본이 제한되며, 채점의 시간과 노력이 많이 드는 단점을 가진 논술형 문항평가에서 수험자들의 진능력(True Ability)을 보다 객관적으로 공정하게 평가할 수 있게 평가치를 적정화시키고, 평가원들의 신뢰도를 향상시키는 연구에 주요 관심을 두고자 한다. 특히, 이 연구 분야에 관심을 두는 이유는 근년에 우리 나라 대학입시에서 수학능력 시험제도가 채택된 이후, 그 시험이 쉽게 출제되어 고득점 동점자들이 많아지고, 그 결과로 논술고사와 면접 등 다른 전형요소의 비중이 오히려 커지는 경향도 크게 나타나고 있다. 그러므로 이러한 논술형 문항평가에서 평가치 산정에 대한 미비점과 평가원들의 신뢰성 부족에 대한 비판의 우려를 해소하기 위하여 보다 바람직한 개선책을 강구할 필요가 있기 때문이다. 다음으로 여기서 논하고자 하는 신뢰도 향상에 관한 연구에서 그 개관을 간단히 개술하면 Ebel R.L.(1951)이 고전 평가론 분야에서 신뢰도 계수에 대한 정의를 제시한 후, Lord F.M.(1968) 등 많은 학자들이 분산분석법에 의한 평가 모형에서 신뢰도 계수의 추정을 위하여 그 계수의 하한 설정, 평가치 분포의 편기에 따른 신뢰도 계수 크기의 변이 추정 등에 의한 간접적인 접근법을 다루고 있으며 Fleiss J.L.(1971)와 Maxwell A.E., Pillinger A.F.(1968) 및 Conger A.J.(1980) 등이 심리학 통계자료에서 이용되고 있는 신뢰도 계수의 정의들 및 그 요인들에 대한 분산분석 모형들을 제시하고 있다. 또 최근에는 이 분산분석 모형에서 산출된 여러 종류의 신뢰도 계수들을 크기 순으로 비교 배열하여 그 이용도와 처리 효과를 강조한 Rae G.(1984) 등의 연구 결과도 있다. 이 논문에서는 그러한 분산분석 모형을 이용한 평가모형에서 결측치가 있거나 분류별 자료모형(Categorical Data Model)[1]의 경우 Landis R.J.와 Koch G.G.(1977), Agresti A.(1988) 및 Uebersax J.S.(1993)와 Longford N.T.(1994) 등이 제시한 내용인 논술, 면접 등 일반적인 서답형 문항 평가에서 평가치 적정화와 평가원의 신뢰도 향상에 유효한 평가자료 모형 및 그 자료처리 과정들에 관한 연구결과들을 참고로 하여, 우리 나라에서 현재 시행하고 있는 논술형 고사의 평가방법에 있어서 더욱 개선된 평가안들을 제시함으로써 이 연구가 보다 우수한 수험자를 선발하려는 바람직한 평가정책 수립에 조금이나마 도움이 되

* 본 연구는 1999년도 동국대학교 전문학술지 논문게재 연구비 지원으로 이루어졌음.

1) 이하 약자로써 CDM으로 표기함.

는 연구가 되었으면 한다.

Ⅱ. 연구의 내용

A. 신뢰도 차이의 유발요인과 평가원 표본의 랜덤화

일반적으로 논술 등의 주관식 문항 평가를 시행할 때, 평가원들은 대개 다음과 같은 채점의 핵심을 고려하고 있다. 즉 그 요점은 제시문의 정확한 이해, 논의의 일관성, 문장 구성의 정확성 등이며 면접이나 구술고사에서는 개인 신상과 지원 동기, 가치관, 기본 소양이나 전공과 관련된 시사 상식 등 다양하고 개성있는 답변을 요구하고 있다. 이때 평가원들은 대개 2~5명 정도이고, 각 평가 문항들은 각 학교의 채점 비율에 따라 각각 다르게 반영되겠지만, 현행 대학입시에서는 전형요소별 성적을 토대로 합격공헌도가 학생부 성적을 1로 했을 때 수능 1.43, 논술 1.13, 면접 0.76 순으로 배분되고 있다. 이 경우 수험생간 점수차를 나타내는 변별력은 수능 0.95, 논술 1.88, 면접 2.51로써 배점이 낮은 논술과 면접이 오히려 합격의 당락을 좌우하고 있음을 고려할 때, 이러한 서답형이나 면접고사 등의 평가에서 보다 객관적이고 공정한 평가의 시행이 더욱 강조되고 있는 현실이다. 지금 이 연구에서 논의하고자 하는 신뢰도 차이의 유발요인에는 평가 내용의 차이, 평가 현장의 상황 또는 평가원들의 심리적 안정 정도의 차이 등 여러 가지 요인이 많으나, 대체로 논술형 고사 평가 현장에서 자주 접하게 되는 평가원들의 평가에 대한 시행의 차이를 고려하여 다음과 같이 세 가지 유형의 신뢰도 차의 유발 요인을 들 수 있다.

(1) 평가원들의 채점에 대한 엄격함(Severity)의 차이와
(2) 문제의 정답에 대한 평가의 견해에 대한 불일치 (Disagreement)
(3) 많은 고사지들을 반복해서 채점하므로 생기는 평가 결과의 변이(Temporal Variation)

이러한 신뢰도 차이의 유발 요인들은 평가를 시행하기 전 채점 회의에서 평가의 요령 및 시행 지침을 미리 평가원들에게 시달함으로써, 평가원 간의 점수 차이에 대한 편차를 어느 정도 줄일 수는 있으나, 이러한 신뢰도 차이의 유발 요인을 완전하게 제거할 수 없는 것이 현실적으로 인정되고 있다. 그리고 현행 논술형 고사 평가에서는, 일반적으로 두 명의 평가원이 평가한 평가 결과를 단순히 산술평균한 것으로 최종 평가치를 결정하고 있으며, 또 평가원들을 고정적으로 배치하여 한 분량의 고사지들을 계속적으로 평가하게 함으로써, 그 결과 평가원들의 신뢰성이 매우 저하되고 있다는 지적도 많다. 그러므로 이 연구에서는 그러한 미비점을 보완하기 위하여 평가원들을 교체하여 배치하는, 즉 평가원 표본의 랜덤화(Randomization)과정을 통하여 신뢰도의 향상을 먼저 시도하고 그것을 기초로 보다 양호한 최적의 평가치를 산정하는 것을 연구의 주안점으로 두었다.

B. 주 연구 내용

(1) CDM 모형의 표본자료구조 및 신뢰도 계수의 추정

여기서는 앞 절에서 논한 바와 같이 Uebersax J.S.(1993)와 Longford N.T.(1994)가 제시한 CDM 모형의 분산분석법(Variance Components Model of Categorical Data)을 이용하여 다음과 같은 처리모형을 가정하였다.

지금 수험자가 $I(i=1, \cdots, I)$명, 평가원이 $J(j=1, \cdots, J)$명이라 할 때, 그들이 평가한 평가 관측치로써

$$X_{ij} = \alpha_i + \beta_{ij} + \varepsilon_{ij} \quad \cdots\cdots (1)$$

인 가산모형[2](Additive Model)을 기초로

(가) 횡열효과 α_i는 수험자들의 능력차의 변수로, 미지의 모평균 μ와 모분산 σ_a^2인 임의 표본($i.\ i.\ d.$ random sample)을 구성하고,

(나) 종렬효과 β_{ij}는 평가원들의 평가결과 차이 즉 j 평가원이 i 수험자를 평가하여 얻은 평가결과의 관측치로 평가원의 엄격함을 표현하는 크기의 값으로써 0인 평균과 분산 σ_b^2을 갖는 임의 표본을 구성하며,

---

2) Longford N.T.(1994)와 Uebersax J.S.(1993)는 CDM 모형에서 관측되어진 평가치를

$$X_{ij_{ik}} = \alpha_i + \beta_{j_{ik}} + \varepsilon_{ij_{ik}}$$

(단, $X_{ij_{ik}}$에서 첨수 j_{ik}는 평가원 j가 수험생 i를 순서 k회에 평가한 관측치임.)
형으로 복잡한 첨수를 표기하고 있으나 여기서는 간편하게 평가원 총수 J와 선정된 평가원수 K의 구분으로 표기하여 이원배치법과 유사한 구조의 표기를 사용하였음.

(다) ε_{ij}가 수험자와 평가원 사이의 교호 작용 오차로써 $N(0, \sigma_e^2)$의 정규분포에 따름을 가정하면, 평가 관측치의 표본자료 구조로써 다음 <표 1>과 같은 표를 얻을 수 있다.

여기서 $\sigma_a^2=0$인 경우는 모든 수험자가 평가원들이 그들을 평가할 때, 모두 동일한 평가결과를 얻는 퇴화분포인 경우이며, $\sigma_b^2=0$인 경우는 평가원들이 그들이 평가하고자 하는 수험자들을 평가할 때, 그들간의 엄격함의 크기에서 차이가 없이 모두 같은 평가결과를 산출한 경우이고, $\sigma_e^2=0$인 경우는 한 조의 평가원들이 한 명의 수험자를 평가할 때, 모두 엄격함의 차이가 없이 동일한 평가결과를 산출해 낸 경우를 의미하고 있다. 다음으로 모든 평가원에 대한 일반적인 신뢰도 계수의 식들을 들어보면 먼저 한 조의 평가원들 간의 평가 결과로 계산한 신뢰도 계수는

$$r_1 = Cor(X_{ij_1}, X_{ij_2}) = \frac{\sigma_a^2}{\sigma_a^2+\sigma_b^2+\sigma_e^2} \quad \cdots\cdots\cdots (2)$$

이고, 한 명의 평가원이 각각 다른 수험자들에 대하여 독립적으로 평가한 결과의 신뢰도 계수는

$$r_1 = Cor(X_{i_1j}, X_{i_2j}) = \frac{\sigma_a^2+\sigma_b^2}{\sigma_a^2+\sigma_b^2+\sigma_e^2} \quad \cdots\cdots\cdots (3)$$

이다.

또, 한 수험자에 대한 각 평가원들의 관측치의 평균치인 $\overline{X}_{i\cdot}$와 수험자들의 진능력점수 a_i와의 신뢰도 계수는

$$r_a = Cor(a_i, \overline{X}_{i\cdot}) = \frac{\sigma_a^2}{\sqrt{\sigma_a^2\left(\sigma_a^2+\frac{\sigma_b^2}{K}+\frac{\sigma_e^2}{K}\right)}} = \left(1+\frac{\tau_b+\tau_e}{K}\right)^{-\frac{1}{2}} \quad \cdots\cdots (4)$$

이다. 단 K는 수험자 i를 평가하기 위해 총 J명 가운데서 선정된 평가원의 수이고

$$\overline{X}_{i\cdot} = \frac{1}{m_i}\sum_{j=1}^{J} X_{ij},$$

$$\tau_b = \frac{\sigma_b^2}{\sigma_a^2}, \quad \tau_e = \frac{\sigma_e^2}{\sigma_a^2}.$$

또, 이 식들에서 모수 σ_a^2, σ_b^2, σ_e^2의 표본추정치들은

<표 1> Longford와 Uebersax의 CDM 모형의 표본자료 구조

평가원군 / 수험치군	1	2	3	4	5	…	j	$j+1$	…	J	횡렬의 관측치수	횡렬의 합	횡렬의 평균
1	X_{11}						X_{1j}			X_{1J}	m_1	$X_{1\cdot}$	$\overline{X}_{1\cdot}$
2		X_{22}	X_{23}							X_{2J}	m_2	$X_{2\cdot}$	$\overline{X}_{2\cdot}$
3	X_{31}		X_{33}				X_{3j}				m_3	$X_{3\cdot}$	$\overline{X}_{3\cdot}$
4	X_{41}	X_{42}			X_{45}						m_4	$X_{4\cdot}$	$\overline{X}_{4\cdot}$
5	X_{51}	X_{52}						X_{5j+1}			m_5	$X_{5\cdot}$	$\overline{X}_{5\cdot}$
⋮											⋮	⋮	⋮
i	X_{i1}						X_{ij}			X_{iJ}	m_i	$X_{i\cdot}$	$\overline{X}_{i\cdot}$
$i+1$			X_{i+13}		X_{i+15}					X_{i+1J}	m_{i+1}	$X_{i+1\cdot}$	$\overline{X}_{i+1\cdot}$
⋮											⋮	⋮	⋮
I		X_{I2}			X_{I5}		X_{Ij}				m_I	$X_{I\cdot}$	$\overline{X}_{I\cdot}$
종열의 관측치수	n_1	n_2	n_3	n_4	n_5	…	n_j	n_{j+1}	..	J	총수 N		
종열의 합	$X_{\cdot 1}$	$X_{\cdot 2}$	$X_{\cdot 3}$	$X_{\cdot 4}$	$X_{\cdot 5}$	…	$X_{\cdot j}$	$\overline{X}_{\cdot j+1}$	…	$X_{\cdot J}$	총합 $X_{\cdot\cdot}$		
종열의 평균	$\overline{X}_{\cdot 1}$	$\overline{X}_{\cdot 2}$	$\overline{X}_{\cdot 3}$	$\overline{X}_{\cdot 4}$	$\overline{X}_{\cdot 5}$	…	$\overline{X}_{\cdot j}$	$\overline{X}_{\cdot j+1}$	…	$X_{\cdot J}$	총평균 $\overline{X}_{\cdot\cdot}$		

적율법을 이용하여 구할 수 있으므로, 다음의 통계량들로써 그 계산과정을 간단히 증명하기로 한다.

지금

1. 수험자군 평가자료들의 평방합

$$S_E=\sum_{i=1}^{I}\sum_{j=1}^{J_i=K}(X_{ij}-\overline{X}_{i\cdot})^2 \qquad (5)$$

(단, $\overline{X}_{i\cdot}=\frac{1}{m_i}\sum_{j=1}^{J}X_{ij}$)

2. 평가원군 평가자료 내의 평방합

$$S_R=\sum_{i=1}^{I}\sum_{j=1}^{J_i=K}(X_{ij}-\overline{X}_{\cdot j})^2 \qquad (6)$$

(단, $\overline{X}_{\cdot j}=\frac{1}{n_j}\sum_{i=1}^{I}X_{ij}$)

3. 총 평방합

$$S_T=\sum_{i=1}^{I}\sum_{j=1}^{J_i=K}(X_{ij}-\overline{X}_{\cdot\cdot})^2 \qquad (7)$$

(단, $\overline{X}_{\cdot\cdot}=\frac{1}{N}\sum_{i=1}^{I}\sum_{j=1}^{J}X_{ij}$)

이라 하면, 각 통계량들의 기대치는

$$E(S_E)=(N-I)(\sigma_b^2+\sigma_e^2)$$

$$E(S_R)=(N-J)(\sigma_a^2+\sigma_e^2)$$

$$E(S_T)=(N-K)\sigma_a^2+\left(N-\frac{\sum_{j=1}^{J}n_j^2}{N}\right)\sigma_b^2+(N-1)\sigma_e^2 \qquad (8)$$

이다. (단, $n^{(2)}=\frac{n_1^2+n_2^2+\cdots+n_J^2}{N}$)

여기서 α_i, β_{ij}, ε_{ij}의 오차 분산의 추정치를 각각 $\hat{\sigma}_a^2$, $\hat{\sigma}_b^2$, $\hat{\sigma}_e^2$이라 하면

$$(N-I)(\hat{\sigma}_b^2+\hat{\sigma}_e^2)=S_E$$

$$(N-J)(\hat{\sigma}_a^2+\hat{\sigma}_e^2)=S_R \qquad (9)$$

$$\hat{\sigma}_a^2(N-K)+\hat{\sigma}_b^2(N-n^{(2)})+\hat{\sigma}_e^2(N-1)=S_T$$

이므로

$$\hat{\sigma}_e^2=\frac{\frac{N-n^{(2)}}{N-I}S_E+\frac{N-K}{N-J}S_R-S_T}{N-n^{(2)}-K+1}$$

$$\hat{\sigma}_b^2=\frac{S_E}{N-I}-\hat{\sigma}_e^2 \qquad (10)$$

$$\hat{\sigma}_a^2=\frac{S_R}{N-J}-\hat{\sigma}_e^2$$

이다.

(2) 평가치 적정화와 신뢰도 계수 안정성의 검정을 위한 진단 과정

이 절에서는 앞 절에서 논한 평가원 신뢰도 계수의 크기를 측정한 후, 그 계수의 신뢰구간을 이용하여 바람직한 신뢰도 계수 크기에 대한 판정과 그 안정성을 진단하고, 동시에 각 평가원들이 평가한 평가치들로써 가장 양호한 최적 평가치를 산정하기 위하여, Agresti A.(1988)와 Uebersax J.S.(1993) 등이 제시한 정리들을 이용하였다. 지금 한 조의 평가원들이 동일한 수험자의 논술형 고사에 대하여 내린 평가치들의 조건부 분산을

$$Var(y_{ij_1}-y_{ij_2}/\beta_{j_1},\ \beta_{j_2})=2\sigma_e^2 \qquad (11)$$

이라 하면 그들 두 평가원들의 평가치 차의 표본 평방합은

$$S_{j_1,j_2}^2=\frac{1}{m_{j_1,j_2}}\sum_{i:\,j_1,j_2}(y_{ij_1}-y_{ij_2})^2 \qquad (12)$$

(단, m_{j_1,j_2}는 j_1평가원과 j_2평가원이 공통으로 평가한 총 고사지의 수)

이므로 이는 식 (11), (12)로써 나타낸 비율 $\frac{S_{j_1,j_2}^2}{2\sigma_e^2}$의 값이 1이면 두 평가원의 평가치가 비슷하거나 거의 같은 평가를 내린 것으로 진단할 수 있고, 여기서 특히 한 명의 평가원의 경우로 한정하면 그 경우의 평가 관측치의 분산은

$Var(X_{\cdot j}/\beta_j)=\sigma_a^2+\sigma_e^2$ 로 표현되고 j평가원들의 그 군내 표본분산은

$$V_j=\frac{1}{n_j-1}\sum_{i=1}^{I}(X_{ij}-\overline{X}_{\cdot j})^2 \qquad (13)$$

이다.

이때, 각 평가원들의 군내 표본분산인 V_j들을 모두 정규분포에 따르는 변량으로 가정하면

$$\frac{\sum_{i=1}^{I}(X_{ij}-\overline{X}_{\cdot j})^2}{Var(X_{\cdot j}/\beta_j)}=\frac{(n_j-1)V_j}{\sigma_a^2+\sigma_e^2}$$

은 자유도 (n_j-1)을 가지는 $\chi^2(n_j-1)$의 분포에 따름을 알 수 있다.

따라서 우리는 다음의 정리들을 얻을 수 있다.

[정리 1]

$W_j=\frac{V_j}{\sigma_a^2+\sigma_e^2}$은 근사적으로 $N\left(1,\sqrt{\frac{2}{n_j-1}}\right)$에 따른다.(신뢰도 계수 안정성의 진단과정)

과정 1. $\frac{S_{ij}^2}{2\sigma_e^2}$의 비율이 1이면 동일한 수험자에 대한 한 조의 평가원이 각각 평가한 평가치의 크기는 차이가 없다는 판정을 할 수 있다.

과정 2. 만일 W_j의 값이 신뢰구간 $\left[1\pm c\sqrt{\frac{2}{n_j-1}}\right]$(단 c는 신뢰계수), 내에 있으면 그 평가원 j의 신뢰도 계수의 안정성은 더욱 양호하다.

[정리 2] (최적 평가치의 산정공식)

한 수험자의 바람직한 평가치로써의 최적 평가치는

$$\hat{a}_{i_0t}=(1-t_i)\,\overline{X}_{i\bullet}+t_i\frac{1}{J}\sum_{j=1}^{J}\overline{X}_{\bullet j} \quad (14)$$

로 주어진다.

(증명) 지금 위의 공식 (14)를 유도하기 위하여 $\hat{a}_{i,t}$의 평균자승오차(MSE)를 고려하면,

$$E(\hat{a}_{i,t}-a_i)^2=Var\left\{-t_ia_i+\sum_{i=1}^{I}r_ia_i\right\}+Var\{\beta_{ij}\}$$
$$+Var\left\{\left(\frac{1-t_i}{K}\right)\varepsilon_{ij}+\frac{t_i}{K}\overline{n_i}\sum_{ij}\varepsilon_{ij}\right\}$$
$$=\left(\frac{\sigma_b^2+\sigma_e^2}{K}\right)-2t_i\frac{\sigma_e^2}{K}(1-\overline{n_i})$$
$$+t_i^2\left\{\sigma_a^2(1-2\overline{n_i}+\frac{\overline{n_i}}{K})+\frac{\sigma_e^2}{K}(1-\overline{n_i})\right\}$$

(단, $\overline{n_i}=\frac{1}{n_{j_i}}$, 여기서 n_{j_i}는 j평가원이 수험자 i를 평가한 총 고사지이며, r_i는 indicator로 평가원이 한 수험자를 채점한 경우 1 아니면 0으로 표기한 것임.)

그러면

$$E(\hat{a}_{i,t}-a_i)^2=E_{i,0}-2E_{i,1}t_i+E_{i,2}t_i^2,$$

$$E_{i,0}=\frac{\sigma_b^2+\sigma_e^2}{K},\quad E_{i,1}=\frac{\sigma_e^2}{K}(1-\overline{n_i}),$$

$$E_{i,2}=\sigma_a^2\left(1-2\overline{n_i}+\frac{\overline{n_i}}{K}\right)+\frac{\sigma_e^2}{K}(1-\overline{n_i})$$

이므로 최소자승법에 의하여

$\frac{\partial E(\hat{a}_{i,t}-a_i)^2}{\partial t}=0$을 만족시키는 t_i는 $t_i=\frac{E_{i,1}}{E_{i,2}}$,

$0\leq t_i\leq 1$이고 그 최소치는 $E_{i,0}-\frac{E_{i,1}^2}{E_{i,2}}$이다.

여기서 $\hat{a}_{i0,t}$는 $0\leq t_i\leq 1$ 내에서 정해진 t_i의 값에 따라 결정되며, 그 결과로써 최적인 평가치를 산정할 수 있다.

Ⅲ. 계산 예 및 검토

다음 예는 어느 대학 입시 평가 연구소에서 시행한 논술고사의 평가 결과를 이용하여 평가원들의 신뢰도 계수와 최적 평가치를 계산한 것이다. 여기서 논술고사의 평가 항목은 ① 창의성 ② 논리성 ③ 표현력의 3단계로 나누었으며, 각 평가 항목에는 (가)에서 (바)까지의 6개 항으로 점수를 배분하였다. 또 평가 항목 ①, ②, ③에는 가중치를 각각 8점, 7점, 5점으로 두어 20점 만점으로 하였는데, 이 평가 결과를 다시 입학고사 총점에 반영되는 비율로 환산한 결과 10점 만점의 평가 관측치 구조를 얻었다. 이 때 평가원의 배치는 5명의 평가원으로 이루어진 평가원 표본을 랜덤화 과정을 거쳐 2명씩 임의로 배치하여 다음과 같이 <표 2>를 얻었다.

지금 <표 2>에 수록된 통계자료와 $N=60$, $I=30$, $J=5$, $K=2$, $S_E=37.0$, $S_R=227.33$, $S_T=236.6$와 식 (2), (3) 과 (10)으로써 $\hat{\sigma}_e{}^2=1.3262$,

$\hat{\sigma}_a{}^2=2.8070$, $\hat{\sigma}_b{}^2=0.0009$,

$r_1=0.64779$, $r_2=0.64779$,

$r_a=0.8993$을 얻었다. 이들 결과로써 $\hat{\sigma}_b{}^2=0$이므로, r_1과 r_2는 같은 값으로 계산되었으며, 이것을 해석해 보면 한 수험자를 평가하는 데 있어서 두 평가원의 신뢰도 계수는 $r_1=0.64779$로써 상당히 높은 편이며, 한 평가원이 두 명의 수험자를 평가할 때의 신뢰도 계수

<표 2> CDM 모형구조화에 의한 논술형 고사의 평가 관측치[3]

평가원군 / 수험자군	1	2	3	4	5	합	평균	표본수
1	3	5				8	4	2
2	8		6			14	7	2
3	8			9		17	8.5	2
4	4				4	8	4	2
5		7	7			14	7	2
6		9		7		16	8	2
7		7			6	13	6.5	2
8			8	9		17	8.5	2
9			9		6	15	7.5	2
10				5	8	13	6.5	2
11	7	9				16	8	2
12	8		8			16	8	2
13	9			7		16	8	2
14	9				6	15	7.5	2
15		7	8			15	7.5	2
16		7		8		15	7.5	2
17		6			7	13	6.5	2
18			4	5		9	4.5	2
19			3		2	5	2.5	2
20				4	6	10	5	2
21	9	8				17	8.5	2
22	3		4			7	3.5	2
23	4			4		8	4	2
24	5				3	8	4	2
25		4	3			7	3.5	2
26		5		6		11	5.5	2
27		9			8	17	8.5	2
28			4	5		9	4.5	2
29			6		8	14	7	2
30				8	7	15	7.5	2
합	77	83	70	77	71			
평균	6.4166	6.9166	5.8333	6.4166	5.9166			
표본수	12	12	12	12	12			

도 $r_2=0.64779$로써 같은 크기의 신뢰도를 나타내고 있고, 또 수험생들의 진능력치와 평가원의 평가관측치 점수에 대한 신뢰도 계수는 $r_a=0.8993$으로써 매우 높은 것으로 판정되고 있다.

3) 실제로 이 <표 2>의 평가관측치 구조를 구성하는 데는, 평가 시행상 평가원이 K=2인 2명씩 한 조를 이루어 채점을 하므로, 본 연구에서 요하는 CDM 모형의 자료를 수집하기 위해 수험자군과 평가원군의 표본의 크기가 비교적 큰 표본으로 얻은 채점 결과에서 두 명의 평가원이 임의로 배치된 경우들을 추출한 것이다.

다음 <표 3>의 결과를 관찰하면 [정리 1]의 평가원들의 신뢰도 계수 안정성의 진단과정 1에 의하여 평가원 1과 2, 1과 5, 3과 5, 4와 5 외에 각 조의 평가원들은 그 정규화 분산치의 값이 모두 1이하의 값으로써 그들은 매우 공정하고 바람직한 평가 결과를 산출하고 있으며, 앞에서 열거한 4개 조의 평가원들의 분산치들은 모두 1이상의 값을 나타내고 있음을 보아, 그들의 평가결과에는 평가의 차이가 너무 큼으로 인하여 평가에 대한 신뢰성이 매우 낮은 것으로 판정할 수 있다.

<표 3> 한 조의 평가원들의 평가 관측치로써 정규화 시킨 분산치

평 가 원	1 과 2	1 과 3	1 과 4	1 과 5	2 와 3
$\frac{1}{2}S_{jj'}/\sigma_e^2$	1.1310	0.6283	0.6283	1.6337	0.2513
평 가 원	2 과 4	2 와 5	3 과 4	3 과 5	4 와 5
$\frac{1}{2}S_{jj'}/\sigma_e^2$	0.6283	0.3770	0.3770	1.7593	1.7593

<표 4>는 식 (13)의 $W_j=\frac{V_j}{\sigma_a^2+\sigma_e^2}$를 <표 1>에서 계산한 평가원들간의 표본분산들인

$V_1^2=5.90152$, $V_2^2=2.81061$,

$V_3^2=4.69697$, $V_4^2=3.35606$,

$V_5^2=3.90152$들의 값으로 대입하여 계산한 값들을 수록한 것인데 이 W_j치들이 앞 절의 [정리 1]의 신뢰도 계수 안정성 진단 과정 2에 의하여 95%의 신뢰구간 $\left[1\pm1.96\sqrt{\frac{2}{n_{j-1}}}\right]$ 내에 있는지 검토한 결과, 모든 평가원들의 통계cl W_j가 구간 [0.4852, 1.5147]내에 포함되고 있으므로 각 평가원들이 그들에게 할당된 총 고사지를 평가할 때 신뢰도에서 차이가 없이 모두 안정적이고 공정한 평가를 시행한 것으로 판정할 수 있다.

<표 4> 평가원 군 간의 통계량 W_j

평가원	1	2	3	4	5
W_j	1.4277	0.6799	1.1363	0.8119	0.9439

끝으로, 아래의 <표 5>는 현재 일반적으로 평가 현장에서 시행되고 있는 평균화에 의한 평균치 점수와 이 연구에서 제시한 [정리 2]의 최적화 공식에 의한 최적 평가치를 각각 비교하여 수록한 것이다. 여기서 t_i는 0.198382

로 계산되어졌으며 식(14)에 의하여 수험생들의 최적 평가치를 산정한 것인데, 이 값들은 모두 평가원들의 신뢰도 크기가 잘 반영되었으며, 동시에 최적화 공식에 의하여 선형화에 의한 자료 적합도 잘 이행되어지고 있음을 보여주고 있다. 따라서 이 최적 평가치가 수험자들의 진능력을 평가함에 있어서 보다 신뢰성이 높으며 더욱 양호한 대표치로 인정되어진다.

<표 5> 평균치와 최적 평가치의 비교

수험자군	1	2	3	4	5	6	7	8	9	10
평 균 치	4	7	8.5	4	7	8	6.5	8.5	7.5	6.5
최적 평가치	4.45	6.86	8.06	4.45	6.86	7.66	6.46	8.06	7.26	6.46

수험자군	11	12	13	14	15	16	17	18	19	20
평 균 치	8	8	8	7.5	7.5	7.5	6.5	4.5	2.5	5
최적 평가치	7.66	7.66	7.66	7.26	7.26	7.26	6.46	4.85	3.25	5.25

수험자군	21	22	23	24	25	26	27	28	29	30
평 균 치	8.5	3.5	4	4.5	3.5	5.5	8.5	4.5	7	7.5
최적 평가치	8.06	4.05	4.45	4.85	4.05	5.65	8.06	4.85	6.86	7.26

Ⅳ. 결론 및 제언

본고에서는 현재 시행되고 있는 면접이나 논술형 고사 등 주관식 문항 평가에서 평가원의 신뢰도 향상에 유효한 CDM 평가 자료 모형의 기술, 표본자료 모형의 구성법, 평가원의 신뢰도 안정성의 진단과정 및 최적 평가치 산정에 대한 자료처리 과정과 통계적 해석법을 제시하였는데, 이 논문의 주요 연구 결과와 기대 효과 및 활용 방안을 요약하면 다음과 같다.

(1) 주관식 문항 평가에서 CDM 모형의 효과적인 운용은 수험생들의 진능력을 판정하는데 있어서, 평가원들의 신뢰도 차이의 유발요인들을 최소화 시켜 최적 평가치의 산정을 용이하게 하고, 평가판정의 기준인 신뢰도와 객관도에 입각한 공정한 평가의 근거를 제공하게 한다.

(2) Ⅱ. A절에서 제시한 바와 같이, 평가원 표본의 랜덤화 과정은 그들 간의 신뢰도 차이의 유발요인을 더욱 최소화시키는데 효과적이다. 특히 수험자 군이나 평가원 군이 소표본인 경우, 평가원의 교체와 채점횟수 증가 등의 반복을 통한 시행으로 수험자들의 고등정신 기능을 평가함에 있어서 보다 정밀한 평가 결과를 도출할 수 있다.

(3) Ⅱ. B절의 [정리 1]과 [정리 2]에서 평가원들의 신뢰도 계수 안정성에 대한 진단과정과 수험자들의 최적 평가치를 산출하는 공식들을 제시하였으며, Ⅲ절에서 그 계산과정 및 검토로써 구체적인 예를 들었다.

(4) 이 연구에서 제시된 수험자들의 최적 평가치는 현행 단순 평균화법에 의한 평균점수보다 평가원의 신뢰도 크기들이 반영된 더욱 바람직한 평가치로써, 평가 판정의 기준에 매우 적합한 대표치로 이용될 수 있다.

(5) 실제로 <표 2>의 평가자료 모형 구성에는 주 3)에서 부기한 바와 같은 어려움이 있었으므로, 그러한 미비점을 보완하기 위하여 Ⅱ. A절에서 제시한 바의 시행을 사전에 이행하기 위한 지침을 세우면 더욱 좋을 것이다.

참 고 문 헌

홍석강 (1988). 수학교육에서 이해력심도의 측정과 방법, 교육문제연구 5, 동국대학교 교육문제연구소, pp.83-95.

홍석강 (1994). 결측치를 가진 목표지향형 평가 모델에서 수학 학습 능력의 평가에 관한 연구, 한국수학교육학회지 시리즈 A <수학교육> 33(2), 서울: 한국수학교육학회, pp.167-175.

홍석강 (1997). 이원화 평가자료의 최적 평가치 산정과 평가원의 신뢰도에 관한 연구, 한국수학교육학회지 시리즈 E <수학교육 프로시딩> 6, 서울: 한국수학교육학회, pp.159-171.

홍석강 (1998). 이원화 평가자료의 최적 평가치 산정법에서 유효성에 관한 연구, 운강 김연식 교수정년기념논집, pp.47-59.

Agresti, A. (1988). A model of agreement between ratings on an ordinal scale, *Biometics* 44, pp.539-548.

Conger, A.J. (1980). Integration and generalization of Kappa for multiple raters, *Psychological Bulletin* 88, pp.322-328.

Ebel, R.L. (1951). Estimation of the reliability of ratings, *Psychometrika* 6(4), pp.407-424.

Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters, *Psychological Bulletin* 76,

pp.378-388.

Fleiss, J.L. & Cohen, J. (1973). The equivalence of weighted Kappa and the intraclass correlation coefficient as measures of reliability, *Educational and Psychological Measurement* 33, pp.613-619.

Landis, R.J. & Koch, G.G. (1977). The measurement of observer agreement for categorical data, *Biometrics* 33, pp.159-174.

Longford, N.T. (1994). Reliability of essay rating and score adjustment, *Journ. of Educational Statistics* 19, pp.171-201

Longford, N.T. (1995). Models for uncertainty in educational testing, Springer, Princeton N.J. U.S.A.

Lord, F.M. & Norick, M. (1968). *Statistical theories of mental test scores,* Addison, U.S.A.

Maxwell, A.E. & Pillinger, A.F. (1968). Deriving coefficients of reliability and agreement for ratings, *The British Journ. of Mathematical and Statistical Psychology* 21, pp.105-116.

Rae, G. (1984). On measuring among several judges on the presence or absence of a trait, *Educational and Psychological Measurement* 64, pp.247-253.

Uebersax, J.S. (1993). Statistical modeling of experts' ratings on medical treatment appropriateness, *Journ. of American Statistical Association* 88, pp.421-427.

Application of the Categorical Data Model for Enhanching the Reliability of the Raters' Ratings and Score Adjustment of the Essay Type Test

Hong, Suk-Kang

Dept. of Mathematics Education Dong Guk University 26-3 Ka, Pil-Dong, Choong-Ku, Seoul, Korea, 100-715; email: skhong@kra.dongguk.ac.kr.

In this thesis we suggested to use the categorical data model effective to eliminate the disagreement of the experts' ratings for score adjustment and to enhance the reliability of the essay type test. The three sources of imperfection can be generally referred to as differences in severity, disagreement in merit and temporal variation. We could estimate the sizes of reliability of raters by obtaining the estimates of the inconsistency variances, $\hat{\sigma}_a^2$, $\hat{\sigma}_b^2$ and $\hat{\sigma}_e^2$ that were results from those three sources of such imperfection.

Especially to eliminate the differences in severity among many raters the randomization procedure of raters sample was very effective in enhancing the reliability of ratings with comparatively small groups of examinees and raters.

And we also introduced the new rating methods, i.e. the 2-step diagnostic procedures to check the sizes of the reliability stability of raters and the sore adjustment method to enumerate the optimal mean values in rating the examinees.