응용통계연구 제 9권 2호
1996년 9월 pp. 45-53

논술 채점의 신뢰도에 관한 연구

허 명 회 1) · 한 상 태 2)

요 약

1994년도 대학입시부터 재개된 논술시험이 점차 그 비중을 확대하는 추세에 있다. 논술시험은 특히 1997년도 입시부터 시작될 교육개혁안에 의한 대학입시체제에서 강조된다고 한다. 그러나 논술시험에 대해 우려되는 것이 몇 가지 있으며 그 중 하나는 채점의 신뢰도에 관한 것이다. 본 연구에서는 1995년도 K대학교 입시 논술의 채점 신뢰도를 높이기 위하여 취해졌던 연구과정과 그 결과를 평가·보고한다. 주요결과는 두 차례에 걸친 독립 채점의 결과 그 차이가 정한 값 이상일 때 재검하여 제3차 채점자가 최종점수를 내는 채점절차가 세 차례에 걸쳐 독립적으로 채점하는 절차보다 신뢰도가 높으며 또한 효율적이라는 것이다.

1. 들어가며

그간 우리나라 대학입시가 4지택일형 객관식 시험을 위주로 구성되어 수험생의 지식 평가에 치우쳐 왔다는 비판이 줄곧 있어 왔다. 이에 대한 반성으로 대학입시의 자율화가 허용된 1994년도 입시부터 전국의 주요 대학들이 수험생의 창의력 및 논리력을 평가하고자 1987년부터 2년간 대입학력고사의 일부로 시행되었던 논술시험(또는 국어논술시험)을 부활시켰는데 논술시험을 채택한 대학의 수는 1994년도에 6개교, 1995년도에 28개교, 1996년도에 18개교에 이르렀던 것으로 나타났다. 특히 논술시험은 1997년도 입시부터 시작될 교육개혁안에 의한 대학입시체제에서 강조될 것으로 보인다.

그러나 논술시험의 확대에 따라 몇 가지 문제점도 앞으로 점차 부각될 것으로 예상된다. 그 중 하나는 논술채점의 신뢰도에 대한 의문으로 통계적 문제라고 할 수 있다. 다시 말하여, 채점자(채점교수)에 따라서 상이할 수밖에 없는 논술점수에 대하여 어떻게 그것을 정당화할 수 있겠는가 하는 것이다. 만약 채점의 객관성, 공정성 및 일관성 등에 근본적 문제가 있다고 인식되는 경우 지금과 같은 논술시험에 대한 사회일반의 긍정적 평가는 급속히 반전될 가능성도 있다.

본 연구는 1995년도 K대학교(서울 캠퍼스) 입시 논술의 채점 신뢰도를 통계적 측면에서 평가하고 이를 높이기 위해 고안되었던 방안을 소개하는 한편 그 유효성을 검토할 것이다. 아울러 입시관리지침의 수립에 있어 통계전문가의 역할이 긴요하였음을 보고하고자 한다. 이 사례결과가 국내 타대학에도 유용하게 적용가능할 것으로 기대한다.

---

1) 고려대학교 정경대학 통계학과 교수. [136-701] 서울특별시 성북구 안암동 5가 1.
2) 고려대학교 통계연구소 Post Doctor 연구원.
* 고려대학교 통계연구소의 일부 지원을 받았음.

2. 모의고사

K대학교의 경우 1994년도 입시에서 논술시험이 처음 실시되었다. 입시채점후 일부 교수들이 채점의 신뢰도에 대한 약간의 우려 또는 불안감을 표명하였고 향후 입시에서는 이를 개선하기 위해 노력할 필요가 있다는 데 대부분 공감하였다. 이를 위하여 국어논술출제분과는 40점 만점인 논술을 각 채점교수가 정서법, 표현의 독창성, 구성의 논리성, 주제의 선명성 등 4개 소영역(각 10점씩 배점)으로 분리하여 평가하도록 채점절차를 명시화하고 입시평가분과에 통계적 신뢰도 향상방안을 의뢰하였다. 이에 대해 본 연구자는 1995년도 입시를 위한 모의고사(1994년 10월 실시)에서 필요한 정보를 얻기로 하였다. 이 시험이 모의고사였음에도 불구하고, 논술부문은 신뢰도의 통계적 평가를 위하여 2인의 교수가 독립적으로 채점하게 한 뒤 자료분석에 들어갔다. <표 1>이 채점자료의 분석결과이다. 이 표에서 r은 제1차 채점 X_1과 제2차 채점 X_2 사이의 상관계수를 말하고 α는 내적 합치도(**內的 合致度**; internal consistency)를 재는 크론바흐(Cronbach, 1951)의 신뢰도 계수이다(이종성, 1985, p.114; 허명회, 1991; Cronbach, 1971; Stanley, 1971). 즉

$$\alpha = N/(N-1) \cdot \left\{1 - \sum_{i=1}^{N} Var(X_i) / Var(\sum_{i=1}^{N} X_i)\right\}$$

여기서 N은 채점자의 수, 이 경우 N=2.

<표 1> 모의고사 논술시험의 채점결과 (2인이 독립적으로 채점)

계열	자료수 n	상관계수 r	크론바흐 신뢰도 α	비고
인문계	594	0.27	0.43	40점 만점
자연계	651	0.54	0.70	40점 만점

계열	점수차(제1차채점 - 제2차채점) 분포의 분위수										
	min	1%	5%	10%	25%	50%	25%	10%	5%	1%	max
인문계	-19	-13	-10	-8	-5	-1	+3	+7	+8	+14	+20
자연계	-15	-12	-8	-6	-2	+1	+5	+8	+10	+13	+21

<표 1>로부터 두 채점 점수사이의 차이가 심한 경우 ±20점 정도나 된다는 것은 알 수 있다. 40점이 만점인 논술에서 누가 언제 채점하는가에 따라 이렇게 큰 차이가 난다는 것은 해당 수험자의 수가 한두명이라고 하더라도 심각한 문제라고 하겠다. 두 채점 점수사이 차이가 ±10점 이상인 사례를 집계한 결과, 인문계에서 57례(전체 594례 중 9.6%)와 자연계에서 62례(전체 651례의 9.5%)가 잡혔다. 따라서 전체의 약 10%쯤에서 심각하게도 ±10점 이상의 점수차이가 발생하는 것을 알게 되었다. 이를 해결하기 위한 방안으로 생각된 것이 다음 두가지이다.

제1안: 3인의 교수가 독립적으로 채점하여 평균점수를 내는 방안,

제2안: 2인의 교수가 독립적으로 채점한 뒤 그 차이가 ±10점 미만인 경우에는 평균점수를 내고 ±10점 이상인 경우 제3의 채점교수가 최종점수를 내는 방안.

여기서 "독립적 채점"은 선행채점의 결과를 후속채점에 앞서 봉함으로써 채점자간 직접적인 정보 교환이 전혀 없음을 말한다.

만약 제1안을 채택한다면 신뢰도가 얼마나 좋아질 것인가를 검토하기로 하자. 이를 위하여 크론바흐의 신뢰도 α에 관한 다음의 공식(Spearman-Brown 예언식의 일반형)을 적용하여 볼 수 있을 것이다(허명회, 1991; Carmines and Zeller, 1979, p.42).

$$\alpha = \frac{N \cdot r}{1 + (N-1)\, r},$$

여기서 N은 채점자의 수, r은 채점점수간 상관계수. (이 식은 $Var(X_1) = \cdots = Var(X_N)$ 이라는 가정하에서 앞서의 α에 관한 공식으로부터 유도되는 것으로 논술자료의 경우 대체적으로 이런 가정이 만족된다). 숫자를 대입하여 보면, 인문계(N=3, r=0.27)에서는 α가 0.43에서 0.53으로 증가하게 되고 자연계(N=3, r=0.54)에서는 α가 0.70에서 0.78로 커지리라는 것을 예상할 수 있다. 그러나 이에 대한 대가로 수험자의 수를 n이라고 할 때, 모두 $3n$ 번의 채점이 있어야 한다.

제2안의 배경에는 처음 두 채점의 차이가 10점이상일 때는 두 채점 중 적어도 하나가 측정오류일 것이라는 전제가 깔려있다. 만약 이 안이 채택된다면, 신뢰도가 어떻게 달라지는가를 보기로 하자. 이때 제3차 채점에는 오류가 없다는 가정을 한다. 채점자 패널 중 가장 유능한 채점자에게 제3차 채점을 맡겨 충분한 시간을 갖고 정밀하게 논술답안을 검토하게 하면 이 가정이 충족되리라 생각한다. <표 2>는 점수차이가 ±10점 미만인 자료로부터 산출된 결과이다.

제2안을 채택할 경우, 재검율이 약 10%가 될 것으로 예상되므로 필요한 총 채점수는 $2.1n$, 즉 응시자수 n의 약 2.1배로 제1안의 총 채점수 $3n$ 보다 작다 (제3차 채점이 제1,2차 채점에 비하여 답안지 1매당 채점시간을 두 배 가량 필요로 하므로 총 채점시간 면에서 본다면 제2안을 채택하는 비용은 $2.2n$이라고 하겠다). 그러나, <표 2>로부터 제2안이 제1안보다 오히려 신뢰도를 약간 더 좋게 하는 것을 볼 수 있다. (제2안이 갖는 채점 신뢰도는 실제 여기에 나타난 수치 이상이라고 봐야 할 것이다. 왜냐하면 제3차 채점의 신뢰도가 1이라고 전제하기 때문이다.) 이런 이유로 본 연구자는 본고사에서 논술시험의 채점지침으로 제2안을 권고하게 되었고 이 안을 국어논술출제분과가 받아들였다.

<표 2> 모의고사에서 2인이 채점한 결과 그 차이가 ±10점 미만인 경우

계열	자료수 n	상관계수 r	크론바흐 신뢰도 α	비고
인문계	537	0.40	0.57	40점 만점
자연계	589	0.68	0.81	40점 만점

제2안과 같이 제1,2차 채점의 어떤 차이 이상을 재검대상의 기준으로 하는 것이, 제1안 (3차례 독립채점 방식)에 비하여 신뢰도를 항상 더 좋게 하는 것은 아니다. 그 이유는 다음과 같다: r을 재검분류 이전의 두 채점간 상관계수라고 하고 α를 3차례의 독립채점으로부터 파생되는 신뢰도라고 하자. 이와 비슷하게 재검대상이 아닌 논술답안지 자료에서의 두 채점간 상관계수를 r^*라고 하고 신뢰도를 α^*라고 하자. $\alpha^* > \alpha$ 일 필요충분조건이

$$\alpha^* > \alpha \Leftrightarrow 2r^*/(1+r^*) > 3r/(1+2r) \Leftrightarrow r^* > 3r/(2+r)$$

이다. 예컨대 <표 1>에서와 같이 r이 0.27과 0.54인 경우 r^*가 각각 0.36과 0.64 이상이 나와야 신뢰도 측면에서 제2안이 제1안보다 낫다고 할 수 있다. 참고로, <표 2>에서 r^*는 각각 0.40과 0.68이다.

3. 본고사

실제 1995년도 K대학 입시의 논술채점에서 점수차이가 ±10점 이상인 답안은 516례(총 7,604례 가운데)로 드러났다. 즉 실제 재검율이 6.8%(인문계 7.2%, 자연계 6.3%)로 나타나 예상치 9.6%보다 약간 작았다. 이와 같이된 이유는 모의고사 때와는 달리 본고사에서는 채점교수들이 채점에 좀더 신중을 기했기 때문일 것이다. <표 3>과 <표 4>를 보라.

<표 3>에서 자연계의 신뢰도가 인문계의 신뢰도에 비해 현저히 낮게 나타났는데 실제 본고사 채점시 자연계의 논술채점에 인문학 및 자연과학 전공교수의 혼합팀이 투입된 결과로 생각된다. 이에 비해 인문계 논술의 채점에는 상대적으로 덜 이질적인 인문학 및 사회과학 분야의 교수들이 투입되었다. 논술채점의 소영역별로 보면, 정서법에 대한 신뢰도가 높은 편인 반면, 특히 자연계에서 표현의 독창성, 구성의 논리성, 주제의 선명성 등에 대한 신뢰도가 상대적으로 낮은 편이었다.

<표 3> 본고사 논술시험의 채점결과 (2인이 독립적으로 채점)

계열	자료수 n	상관계수 r	크론바흐 신뢰도 α	비고
인문계	4,065	0.50	0.67	40점 만점
	정서법	0.42	0.59	10점
	표현의 독창성	0.40	0.56	10점
	구성의 논리성	0.40	0.56	10점
	주제의 선명성	0.38	0.55	10점
자연계	3,539	0.33	0.50	40점 만점
	정서법	0.37	0.54	10점
	표현의 독창성	0.20	0.34	10점
	구성의 논리성	0.21	0.35	10점
	주제의 선명성	0.25	0.39	10점

계열	점수차(제1차채점 - 제2차채점) 분포의 분위수										
	min	1%	5%	10%	25%	50%	25%	10%	5%	1%	max
인문계	-23	-12	-8	-6	-3	0	+3	+7	+9	+12	+20
자연계	-19	-12	-8	-6	-3	+1	+4	+7	+9	+13	+22

<표 4> 본고사에서 2인이 채점한 결과 그 차이가 ±10점 미만인 경우

계열	자료수 n	상관계수 r	크론바흐 신뢰도 α	비고
인문계	3,771	0.65	0.79	40점 만점
자연계	3,317	0.49	0.66	40점 만점

<표 5> 제1차 채점, 제2차 채점, 제3차 채점간 상관계수

	인문계			자연계		
	1차	2차	3차	1차	2차	3차
1차	1	-0.455	0.377	1	-0.398	0.400
2차	-0.455	1	0.351	-0.398	1	0.309
3차	0.377	0.351	1	0.400	0.309	1

모의고사에서와 달리 본고사에서는 점수차이가 ±10점 이상인 답안은 재채점된다. 이제 제3차 채점에 아무 오류가 없다는 가정하에 탐색적 자료분석의 관점에서 제1차, 제2차 채점이 어떤 오류를 범하는가를 보기로 하자.

제3차 채점에 들어간 사례에 한하여 제1차 채점을 X_1, 제2차 채점을 X_2, 제3차 채점을 X_3로 정의하자. 이 때 X_1, X_2, X_3 간 상관계수는 <표 5>와 같다. 제1차와 제2차간 상관계수가 음이고 (∵ 제1차 또는 제2차 채점 중 적어도 하나가 측정오류이기 때문), 제1차와 제3차간 상관계수 및 제2차와 제3차간 상관계수가 양이나 <표 4>의 상관계수보다 작으므로 (∵ 제1차과 제2차 중 일부가 측정오류이므로 나머지 일부는 그렇지 않으므로) 제3차 채점이 제1,2차 채점선상의 어디에 위치하는지를 알아볼 필요가 있다.

따라서 다음과 같이 제3차 채점 X_3의 표준화된 변환을 생각하기로 하자.

$$Z = 2 \cdot \frac{X_3 - \min(X_1, X_2)}{\max(X_1, X_2) - \min(X_1, X_2)} - 1 .$$

이 변환은

$$\begin{aligned} X_3 &= \min(X_1, X_2) \quad \Rightarrow \quad Z = -1 \\ X_3 &= (X_1 + X_2)/2 \quad \Rightarrow \quad Z = 0 \\ X_3 &= \max(X_1, X_2) \quad \Rightarrow \quad Z = +1 \end{aligned}$$

로 바꾼다. 즉 Z가 0을 중심으로 -1 쪽이면 제3차 채점자가 작은 쪽 점수를 편들었음을 말하고 반대로 +1 쪽이면 제3차 채점자가 큰 쪽 점수를 편들었음을 말한다. <표 6>이 새 변환 Z에 관한 요약결과이다. 인문계의 경우, Z의 중위수가 +0.17에 있는 것으로 봐서 제3차 채점이 선행 두 점수 중 큰 쪽을 편드는 경향이 약하게 있다고 하겠다. 그러나 자연계의 경우에는, Z의 중위수가 -0.33에 있어 제3차 채점이 선행 두 점수 중 작은 쪽을 편드는 경향이 발견되었다.

<표 6> 제3차 채점의 상대적 위치

	Z의 분위수 (n=820)										
계열	min	1%	5%	10%	25%	50%	25%	10%	5%	1%	max
인문계	-3.80	-1.40	-1.00	-0.67	-0.36	+0.17	+0.63	+1.00	+1.00	+1.60	+1.92
자연계	-1.55	-1.31	-1.00	-0.86	-0.64	-0.33	+0.00	+0.41	+0.82	+1.00	+1.18

4. 부수사항 및 토의

논술시험의 실제 채점에 있어서는 채점자에게 표준분포를 제시하여 주는 것이 신뢰도 향상에 도움이 된다. 왜냐하면, 40점 만점에서 어떤 채점자는 중간수준을 28점쯤으로 생각하고 어떤 채점자는 32점쯤으로 생각할 수도 있는데 가급적 한 단과대학이나 한 학과(또는 학부) 답안지 채점을 동일한 채점자 쌍에 의뢰한다는 채점원칙이 실제로는 엄격히 지켜지지 않기 때문이다. 따라서 본 연구자는 <표 7>과 같이 모의고사 제1차 채점의 분포를 본고사 채점교수들에 참고용으로 제시하였는데, 본고사 채점시 많은 도움이 되었다는 이야기를 논술채점교수들로부터 들을 수 있었다. (더불어 K대학교는 전년도 본고사의 단과대학별 문항분석 통계를 갖고 있다.) 역시 <표 7>에 본고사 제1차 채점결과도 제시하였다. 모의고사에 비해 본고사 점수분포의 분위수가 1-2점 낮은 경향이 있으나 이는 본고사 수험자의 질이 모의고사에 비해 상대적으로 불균일하기 때문으로 생각된다. 제2차 채점결과는 산포(**散布**)가 작아지는 경향을 약간 보였으나 대체로 제1차 채점과 동일하므로 여기에 제시하지 않겠다. (인문계의 경우 제1차 채점의 IQR이 7점이었으나 제2차 채점에서는 5점으로 나왔다. 이와 같이 산포가 작아지게 된 이유는 피로(**疲勞**)로 인한 채점자의 분별력 저하 때문인 것으로 보인다.)

K대학교의 경우에는 본고사에서 점수차이가 ±10점 이상인 답안을 재채점하였으나(재검율 6.8%), 몇 점이상의 차이를 재채점의 기준으로 두느냐의 결정은 개별대학의 입시관리여건에 달려있다고 생각된다. 다시 말하자면, 재채점의 기준을 가급적 엄격하게 하는 것이 좋겠지만 그렇게 하면 할수록 입시관리 자원(인적·물적)의 한계를 쉽게 넘어서게 될 것이라는 것이다. 논술 재검은 특히 신중히 처리되어야 하므로 각 채점자 패널의 대표자가 전담할 필요가 있고 제1,2차 채점에 비해 논술답안 1매당 2배가량의 채점시간이 소요된다는 것을 염두에 두면, 재검율의 실무적 상한선이 최대 25% 정도가 아닐까 한다.

K대학교의 1995년도 입시자료에서 점수차이가 ±8점 이상인 답안을 집계하여 보면 재검대상 답안지 비율이 전체적으로 14.0%(인문계 14.5%, 자연계 13.5%)로 나온다. ±6점 이상을 재검기준으로 하는 경우는 재검대상답안지 비율이 전체적으로 26.4%(인문계 27.0%, 자연계 25.8%)가 된다. 만약 실제 채점시 이에 상응하는 추가적 노력을 투입한다면, ±10점 이상의 점수차이를 재검기준으로 하는 경우에 비해 신뢰도가 상당히 증가하게 될 것이다. <표 8>과 <표 9>를 보라.

<표 7> 논술점수의 분포

	모의고사 제1차 채점 분위수 (전체)										
계열	min	1%	5%	10%	25%	50%	25%	10%	5%	1%	max
인문계	0	13	21	23	26	29	32	35	36	39	40
자연계	0	4	20	23	27	29	32	35	37	40	40

	본고사 제1차 채점 분위수 (전체)										
계열	min	1%	5%	10%	25%	50%	25%	10%	5%	1%	max
인문계	0	14	19	22	25	29	32	35	36	39	40
자연계	0	14	20	22	25	27	30	33	35	38	40

<표 8> 본고사에서 2인이 채점한 결과 그 차이가 ±8점 미만인 경우

계열	자료수 n	상관계수 r	크론바흐 신뢰도 α	비고
인문계	3,474	0.74	0.85	40점 만점
자연계	3,062	0.58	0.73	40점 만점

<표 9> 본고사에서 2인이 채점한 결과 그 차이가 ±6점 미만인 경우

계열	자료수 n	상관계수 r	크론바흐 신뢰도 α	비고
인문계	2,969	0.84	0.91	40점 만점
자연계	2,627	0.71	0.83	40점 만점

논술시험이 그간 지식 평가에 치우쳤던 대학입시를 창의력·논리력 평가로 유도하는 기능을 갖고 있기 때문에 향후 더 큰 비중을 갖게 될 전망이다. 그런 만큼 논술시험의 제도적 확립을 위해서는 다각적 측면에서 많은 연구가 있어야 할 것이다. 이 연구는 논술채점의 신뢰도 문제를 다루었지만 이외에도 논술의 예측타당도(predictive validity)에 관한 실증적 연구도 해볼 필요가 있다. 예컨대 논술점수와 대학에서의 학업성취도(즉 GPA) 사이의 연관성을 확인해 봐야 할 것이다. 이에 관하여 K대학교에서 얻고 있는 연구결과에 대하여는 추후에 보고하기로 하겠다.

추기 1: 최근 성태제(1995)는 타당도와 신뢰도에 관하여 일반적으로 논의하는 가운데 구체적인 교육통계적 사례를 제시하였다. 또한 평정자(채점자)간 신뢰도에 관한 방법론으로서 1) 채점자간 신뢰도 또는 일치도, 2) 일반화 가능도 이론 등을 체계적으로 정리하였다.

추기 2: 본연구의 2절에서 제시된 제1,2안 외에, 제3안으로 처음 두 채점점수가 10점이상의 차이를 보일 경우 제3차 채점을 하여 최종 점수로 하되 최종점수는 첫 두 채점점수 중 낮은 값보다 더 낮게 줄 수 없도록 하는 규정을 둘 수 있을 것이다. 본 연구의 제2안의 경우에는 제3차 채점점수(최종 점수)가 첫 두 점수보다 낮아질 수 있는데, 이것이 익명의 한 심사위원에 의하여 지적된대로 수험자의 개인적 불이익을 초래할 수 있다. 제3안은 제2안의 그러한 점을 보완한 것이나 한편으로는 약간의 변별력 저하를 결과할 수 있겠다.

참고문헌

[1] 성태제 (1995). 『타당도와 신뢰도』, 서울: 양서원.
[2] 이종성 (1985) 역. 『행동과학연구를 위한 측정이론의 기초』, 서울: 중앙적성출판사.
[3] 허명회 (1991). 설문지·시험지 문항의 신뢰성 분석, 『응용통계연구』 4권 1호, 93-104.

[4] Carmines, E.G. and Zeller, R.A. (1979). *Reliability and Validity Assessment,* Sage Publications, Beverly Hills, California.

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

[6] Cronbach, L.J. (1971). Test Validation, in *Educational Measurement*, 2nd Edition (edited by R.L. Thorndike), American Council on Education, Washington, D.C.

[7] Stanley, J.C. (1971). Reliability, in *Educational Measurement*, 2nd Edition (edited by R.L. Thorndike), American Council on Education, Washington, D.C.

Reliability of Essay-Writing Scoring in University Entrance Exam

Myung-Hoe Huh [1] and Sang-Tae Han [2]

Abstract

Essay-writing, first introduced to Korean university entrance exams in 1994, is gaining its weight year by year. Especially from 1997 when Nationwide Education Reform System begins, it will be a key component of student selection criteria at Korean universities. Essay-writing's future, however, will not be that smooth unless it shows necessary validity and reliability.

This study is on reliability of Essay-writing scoring, mainly from the experience of University K case. To secure solid reliability in Essay-writing scoring for the 1995 University Entrance Exam, the authors started research from the 1994 Autumn Pre- exam which was administerd to potential applicants of University K following year. Total of 1,254 students took Essay-writing exam and, subsequently, their essays were graded by two professors independently. The result was not so good. The correlation between two scores was 0.27[0.54] with Cronbach alpha 0.43[0.70] for Humanity-Social Science [Natural Science-Engineering] field. So, some action for reliability improvement was inevitable. The authors considered and investigated following two alternatives.

Alternative 1[A1]: Essays are to be graded three times independently at the 1995 University K Entrance Exam. Scores will be given as the average of three scores.

Alternative 2[A2]: Essays are to be graded twice independently, followed by a possible third grading only if two gradings show "significant" difference. Scores are given as the third score if done or the average of first two scores otherwise.

From Pre-exam data, it was predicted that A1 will yield Cronbach alpha 0.53[0.78] for Humanity-Social Science[Natural Science-Engineering] field. On the other hand, for A2 with cut-off difference ± 10 points for the third grading (between first two scores of 40 points each), the prediction was that it will need third grading for 10% of the essays and that it will have Cronbach alpha 0.57[0.81] for Humanity-Social Science [Natural Science-Engineering] field. Hence A2 was recommended from the reliability and economic reasons. The University decided to adopt A2 for its 1995 entrance exam. Finally, it turned out that University K Entrance Exam data have 6.8% third gradings out of 7,604 essays and larger Cronbach alpha reliability coefficient than expected.

1) Professor, Department of Statistics, Korea University.

2) Post Doctoral Researcher, Institute of Statistics, Korea University.
Mailing address: Anam-dong 5-1, Sungbuk-ku, Seoul 136-701, Korea.