영어교육 51권 4호 1996년 겨울

채점 신뢰도 분석*

김 충 배
(고려대학교)

Kim, Choong Bae. (1996). An analysis of scoring reliability. *English Teaching, 51*(4), 117-125.

This paper purports to analyze the scoring reliability of one question item which composes a Korean university entrance English examination administered in January 1994. Fifty-seven types of examinees' responses to the open-ended question item obtained during the marking period were requested to be rated by five English language or literature professors. The ratings of the five judges were computated in terms of correlation coefficient and Fisher Z transformation in addition to descriptive statistics. The results revealed that the dialogue-completion question item proved undesirable for a question item for the very competitive university entrance examination. The very low inter-rater reliability estimated as .51 is largely due to the construction of the question item itself because the range over which possible answers vary was not restricted. This analysis finally suggested that the test maker should be careful in making subjective question items so that item rating can be as objective as possible.

Ⅰ. 서론

1994학년도부터 5지 선다형 대학 수학 능력 시험이 실시되면서 객관식 시험의 문제점을 보완할 필요성에서 이른바 대학별 본고사에서는 주관식 위주의 문항이 출제되었다. 그 이전 대학 입학 학력고사 실시 때는 8개 정도의 주관식 문항이 있었고, 정답 및 부분 점수 인정 문제가 야기된 적도 있었으나 국가 기관에서 관리한 그 시험에서는 심각히 우려할만한 주관식 문항 채점 신뢰도 문제는 없었다. 다

* 이 논문은 1995년도 고려대학교 특별연구비의 지원을 받았음.

시 없어지게 된 지난 3년간의 대학별 본고사 영어 주관식 문항 중 어떤 문항은 과연 신뢰할만한 채점이 어떻게 가능할까하는 의구심을 필자로 하여금 자아내게 했다. 아무리 주관식 시험에 의한 평가의 중요성이 강조되고, 실제로 어떤 주관식 문항의 출제 의도가 아무리 훌륭하다고 해도, 신뢰할만한 채점이 보장되지 않으면 그 시험 자체가 신뢰될 수 없다.

어느 시험에서나 있을 수 있는 주관식 문항으로 채점 신뢰도가 문제가 될 수 있는 실례로 94학년도 대학별 본고사 문제 중의 하나인 아래 문항을 살펴보자.

주 1. 다음 대화를 읽고 빈 칸에 들어갈 적절한 말을 영어로 넣으시오. (2점)

John:	Hi, Mary! What are you doing this weekend?
Mary:	Oh, I'm going to a concert with my roommate, Sally.
John:	_____?
Mary:	It's the Los Angeles Symphony Orchestra concert.
John:	What's on the program?
Mary:	The music of Beethoven and Mozart.
John:	Have a great time at the concert.
Mary:	Thanks.

이 문항에 대한 수험자들의 답의 유형이 대단히 많을 것이며, 그 많은 유형의 수험자 응답을 2점, 1점, 0점으로 채점 기준을 정하고 실제로 신뢰할만한 채점을 하는 작업이 대단히 어려울 것이라는 예상을 하게 한다.

이 논문의 목적은 젊은이의 인생행로를 좌우하는 한 중요한 시험에 출제되었던 한 개의 영어 테스트 문항을 실례로 채점 신뢰도 문제를 제기해 보려는 데 있다. 구체적으로 말하면 앞서 언급한 94학년도 대학별 본고사 영어 주관식 한 문항의 채점 신뢰도를 분석하여 채점 신뢰도가 낮을 수밖에 없는 문제임과 부분 점수를 인정할 경우 수십 가지 응답이 가능하며, 채점자에 따라 부여 점수가 주목할 만하게 달라질 수 있고, 채점 기준을 정한다해도 그 기준이 타당하다는 확신을 줄 수 없음을 재시하고자 한다.

그 시험을 구성하는 여러 문항을 모두 다 고려하지 않고 단 한 개의 문항을 선택하여 분석하는 것은 다른 문항에 대해서는 응시자의 응답 예를 알 수 없고, 이

문항은 실제로 채점을 함으로써 수험자의 응답 자료를 입수할 수 있었기 때문이다. 한 개 문항만을 가지고 구체적으로 분석해 볼 지면의 여유밖에 없을 뿐만 아니라 한 개 문항의 채점 신뢰도 분석만으로도 파급되는 문제점을 시사할 수 있다고 생각한다.

이런 실증적 분석보고는 4지 또는 5지 선다형 객관식 시험(TOEFL, TOEIC, 대학 수학 능력 시험 등)을 보완할 목적으로 실시되는 주관식 시험(대학별 본고사, 대학원 입학시험, 취업시험, 교원임용고사 등)의 출제와 채점에 보다 신중하고 보다 전문적인 접근에 다소간 보탬이 될 것이다. 우리나라에 영어 시험의 실증적 채점 신뢰도에 관한 연구나 보고가 필자가 아는 한 별로 없다.

II. 채점 신뢰도 확보 방법

시험의 신뢰도에 영향을 끼치는 가장 중요한 요인 중의 하나가 채점이며, 객관식 시험의 경우 채점자 신뢰도가 문제되지 않으나 주관식 시험의 경우 채점자 신뢰도 문제가 제기되는데 시험이 신뢰받을 수 있으려면 무엇보다도 측정에 일관성이 유지되어야 한다 (Heaton 1988, pp. 162-164).

Hughes(1989, pp. 36-42)는 시험의 신뢰도를 높일 수 있는 7가지 방법과 채점자 신뢰도를 확보하는 7가지 방법을 제시하고 있는데 그 영문 제목과 내용의 요지를 괄호에 넣어 간략히 소개하면 아래와 같다.

How to Make Tests More Reliable

(1) Take enough samples of behaviour. (서로 독립된 문항수가 많을 것)
(2) Do not allow candidates too much freedom. (여러 항목을 제시하여 선택하게 하거나 가능한 답의 변수의 폭이 크지 않도록 할 것)
(3) Write unambiguous items. (문제의 의도가 분명하고 출제자가 기대하지 않았던 정답이 없도록 할 것)
(4) Provide clear and explicit instructions. (수험자가 다른 해석을 하지 않도록 지시사항이 분명할 것)
(5) Ensure that tests are well laid out and perfectly legible.
(시험지 만듦새와 인쇄 등이 선명할 것)
(6) Candidates should be familiar with format and testing techniques.
(수험자가 시험 형식에 친숙하도록 할 것)
(7) Provide uniform and non-distracting conditions of administration.
(시험 시행에 시간, 소음, 잘 들림 등 동일한 조건을 마련할 것)

Ways of Obtaining Scorer Reliability

(8) Use items that permit scoring which is as objective as possible. (객관적 채점이 가능한 문항을 만듬)

(9) Make comparisons between candidates as direct as possible. (여러 문제 항목중 하나를 선택하여 답을 쓰게 하지 않고 동일한 문제 항목에 대한 응답을 채점하게 함)

(10) Provide a detailed scoring key. (정답, 부분정답 및 배점을 상세히 규정함)

(11) Train scorers. (정확하게 채점할 줄 모르는 사람에게 채점을 맡기지 않음)

(12) Agree acceptable responses and appropriate scores at outset of scoring. (여러 수준 차이를 보이는 샘플을 골라 예비 채점을 해보고 미리 기준을 정함)

(13) Identify candidates by number, not name. (응시자의 이름, 성별, 국적 등이 누구인지 모른 상태로 채점함)

(14) Employ multiple, independent scoring. (2인이 각기 별도로 채점하고 제3의 고참자가 두 성적을 비교하고 불일치를 조사하게 함)

이 밖에 채점자 신뢰도에 영향을 끼칠 것으로는 본질적인 문제는 아니나 채점자의 실수와 고집을 들 수 있을 것이다. 장시간 많은 분량의 채점 경우 피로로 지속적인 주의 집중력이 떨어져 일관된 채점기준 적용에 실패한다든지, 정해진 채점기준에 동의하지 않고 자신이 설정한 다른 기준을 고집하는 하는 경우 채점 신뢰도가 영향을 받게 됨으로 채점자의 지나친 자유를 허용하지 않아야 한다.

앞서 (10)항에서 만점 정답과 부분 점수 부여 기준을 상세히 규정하여야 한다고 했는데 대단히 중요한 이 작업은 여러 경우 그리 간단치 않을 수 있다. 국부적(local) / 전국적(global) 오류 여부,[1] 문법적 / 의미적 / 화용적 견해, 정확성 / 유창성 등의 입

---

1) 실례로 같은 시험의 주관식 7번 문항은 밑줄 친 부분을 우리말로 번역하는 4점짜리 문제인데 밑줄 쳐진 영문과 채점기준은 다음과 같다.

No one can think intelligently without knowing the facts; and if the facts are controlled by interested men, the very idea of democracy is destroyed and becomes a farce.

(주7) 4점
사실이 이해관계가 있는 사람들(이해집단)에 의해 좌우(지배)된다면
(1)
(1)
민주주의 개념 그 자체가 무너진다.
(1)
(1)

위 채점 기준에 의하면 'interested'만을 '흥미를 가진'으로 잘못 번역했다면 3점을 받게 된다. 이 4점짜리 번역 문제에서 'interested'의 의미 파악 잘못은 과연 local error로 간주해도 좋을 것인지는 의문이 간다.

장 정리가 된 채점 기준 설정은 복잡하게 되기 마련이다.

다시 제약이 없는(open-ended) 빈 칸 채우기 1번 문항으로 돌아가 이 문항의 채점 신뢰도가 문제가 되는 것은 (2), (3), (8)항이 지켜지지 않았고 (10)항의 이행이 사실상 불가능하다고 볼 수 있다. 문제 자체가 다양한 응답 가능성이 있고 출제자가 기대하지 않았던— 기대했던 정답은 What / Which concert (is it)? 4가지임— 정답 내지 부분점수가 가능한 응답의 변이의 폭이 크며, 이 많은 변이는 채점자로 하여금 지속적인 집중력과 기억력으로 일관된 채점을 어렵게 할 것이 틀림없다.

III. 채점자 신뢰도 분석

1. 수험자 응답 예 수집

필자는 채점 중 채점기준을 정하기 위해, 또한 실제 점수를 부여할 때 애매한 것이 자꾸 나와서 0점 처리된 것을 포함하여 80여개의 수험자 응답을 메모하게 되었다. 공식적인 채점 기준에 따라 사흘간의 채점을 아무튼 끝낸 후 이 문항의 채점 신뢰도를 분석해 볼 목적으로 이 80여개의 응답 예중 0점이 확실시되는 20여개를 삭제하고 정확히 57개의 응답 예를 확보했다.

2. 5인 채점 조사

94년 2월 22일부터 3월 9일에 걸쳐 대학교수 다섯 분(A, B, C, D, E로 칭함)에게 주관식 1번 문항을 그대로 제시하고, 수험자가 57개 유형으로 응답했을 경우 2점, 1.5점, 1점, 0.5점, 0점 중 어떤 점수를 부여하겠는지를 조사했다.[2] 한국인 교수 A, D, E는 각기 전공이 언어테스팅, 영어교육, 영문학이며 세 분 모두 영어 원어민 수준의 영어 실력(near native proficiency)을 갖추고 있으며, B와 C는 영어 원어민으로 영어교육 전공 석사 학위 이상을 지닌 객원교수이다.

채점 설문 조사지를 회수하고서 다섯 채점자의 평균 점수가 0.5점 미만인 27개 응답 항목을 삭제시켰다. 0점 처리되는 응답으로 간주될 수 있고, 반올림하여 1점이상의 득점 가능성이 있는 것만 다루는 것이 편리해서였는데, 채점 결과는 표 1과 같다.

---

2) 설문지의 회수 후 보완할 것이 있어 재채점을 부탁한 경우도 있었는데 처음 채점과 일치하지 않는 항목이 주목할 만하게 있었다. 필자는 이를 예측했었는데, 폭이 좁은 정답과 객관성 있는 채점을 전혀 고려하지 않은 출제 때문에 시차를 초월하는 일관된 채점이 원초적으로 불가능하다고 판단한다. 이 글에서는 다루지 않는 intra-rater reliability에 대해선 Henning(1987, p. 76)과 Bachman(1990, pp. 178-180)을 참조할 것.

표 1
채점표

	수험자 응답 예	부여점수					평균 점수
		2	1.5	1	0.5	0	
1	What concert is it?	a de	b	c			1.7
2	What concert?	abcde					2.0
3	Which concert is it?	e	ab d	c			1.5
4	Which concert?	b e	a cd				1.7
5	Whose concert (is it)?	abcd				e	1.6
6	Which one (is it)?	b de	a	c			1.7
7	What is / What's the concert?	e		a	bcd		0.9
8	What / Which orchestra concert (is it)?	d		a	e	bc	0.7
9	What / Which orchestra is it?	d		a c	e	b	0.9
10	What kind of concert (is it)?	abc	de				1.8
11	What kind of a concert (is it)?	b	c	a de			1.3
12	What kind of concert are you going to?	a	bc	e	d		1.3
13	What / Which concert will you go to?			a c e	b	d	0.7
14	What / Which concert will you go?		b		a e	cd	0.5
15	What concert is it that you are going to?	a c		e	b	d	1.1
16	Whose concert are you going to?	a	bcd			e	1.3
17	What / Which concert are you going to?		a de	bc			1.3
18	What / Which concert you are going to?		de			abc	0.6
19	What / Which concert will you be going to?		a c	e	b	d	0.9
20	What is / What's that?		d		a c	b e	0.5
21	What concert is that?	a c		e	d	b	1.1
22	Which one is that?		a c	d	e	b	0.9
23	Who's / Who is playing?	bcd	a			e	1.5
24	Who is playing in / at the concert?	d	abc			e	1.3
25	Who is going to play in / at the concert?	d	abc			e	1.3
26	What is / What's the name of orchestra?		c	ab	de		0.9
27	What is / What's the name of the orchestra?	d	c	ab e			1.3
28	What concert are you going to listen to?		abc			de	0.9
29	What concert are you going to see / hear?		abc		d	e	1.0
30	A concert?	b	a c	d		e	1.2

정교한 통계적 분석 없이 표 1만으로도 다섯 채점자간의 차이가 상당히 많이 나는 것을 직감할 수가 있다. 완전일치는 2번 항목(What concert?) 1개뿐이며, 비교적 일치된다고 간주되는 항목은 1, 3, 4, 6, 10, 11, 13, 17, 26, 27번의 10개 항목에 불과하다. 반면 상당한 불일치를 보이는 항목이 8, 9, 12, 15, 19, 20, 21, 22, 30번의 9개 항목이나 된다.

종합적인 비교를 위해 만들어진 아래 표 2에서 보듯이 A가 다소 후한 채점(총점 42점)을 하고 E가 (총점 26점) 다소 박한 채점을 하고 있으며 B, C, D가 각기 총점 34.5, 38.5, 36점으로 중간치(35.4점)에 접근한다. 그러나 총점은 별 의미가 없고 점수대별 빈도에서 의미 있는 차이를 보이고 있음이 주목된다. 즉 최빈치의 경우 A, B, C는 1.5점이고, D는 2점이며, E는 0점으로 각 채점자가 특정 점수를 많이 부여했음을 알 수 있다.

표 2

기술 통계[3)]

채점자	부여점수 빈도					총점	평균	범위	중앙치	최빈치
	2	1.5	1	0.5	0					
A	8	12	7	2	1	42.0	1.40	0~2	1.5	1.5
B	8	9	3	4	6	34.5	1.15	0~2	1.5	1.5
C	6	13	6	2	3	38.5	1.28	0~2	1.5	1.5
D	10	7	3	5	5	36.0	1.20	0~2	1.5	2.0
E	6	3	7	5	9	26.0	0.87	0~2	1.0	0.0

3) 참고로 표준편차, 평균표준오차 및 신뢰구간(95%)은 아래 표와 같다. 이 통계와 표3의 채점자간 상관계수 및 Fisher's Z transformation은 전지현 선생의 도움을 받았음을 밝힌다. (Inter-Rater) Reliability 산정에 관해선 Henning(1987, pp. 80-85)과 Krzanowski와 Woods(1984)를 참조할 것.

채점자	표준편차	평균표준편차	신뢰구간
A	.515	.094	.192
B	.756	.138	.282
C	.597	.109	.223
D	.761	.139	.284
E	.754	.138	.281

3. 채점자 신뢰도

다섯 채점자간의 상관계수는 표3과 같으며, 채점자간 신뢰도를 측정하는 Cronbach's Alpha는 0.5102이다.

표 3

채점자간 상관 계수

	A	*B*	*C*	*D*	*E*
A	1				
B	.4382	1			
C	.7676	.4373	1		
D	.0747	.2307	.0038	1	
E	.0755	.0515	-.1430	.0782	1

상관계수가 0.50을 넘는 것은 A · C간 1개, 근접하는 것은 A · B간과 B · C간 2개뿐이며, 0.50에 훨씬 못 미치는 것이 A · D간, A · E간, B · E간, C · D간, C · E간, D · E간 등 6개가 된다. 또한 Fisher's Z분석에 따른 *p*값은 아래와 같은데 A와 B 및 C, B와 C간만이 상관관계가 유의할 뿐이다.

A · B: .0146* A · C: <.0001* A · D: .6972 A · E: .6943
B · C: .0148* B · D: .2222 B · E: .7890 C · D: .9843
C · E: .4542 D · E: .6840

채점자간 신뢰도를 종합적으로 나타내는 Cronbach's Alpha값 0.51은, 절대적인 기준은 없으나 다른 상황의 경우 0.7 이상이면 괜찮다고 말할 수 있다. 그러나 1점 차이에 합격과 불합격이 좌우되는 우리나라 대학 입학 시험같은 중요한 경쟁 시험에서는 이 채점 신뢰도는 대단히 낮은 수치라고 해석된다. 적어도 0.8이 넘는 신뢰도는 얻어져야 한다고 본다.

IV. 결론 및 제언

지금까지 응답의 제약이 적어 가능한 답의 변수의 폭이 크고 출제자가 기대하지 못했던 (부분) 정답이 가능한 한 간단한 짧은 대답 시험 문항을 예로 들어 채점 신뢰도 문제를 조사해 보았다. 이 문항은 출제 자체가 채점 신뢰도가 낮을 수밖에 없는 문제로 타당한 채점 기준을 정하는 것 자체가 간단치 않으며, 수십 가

지 응답이 부분 점수 부여가 가능하여 채점의 정확성과 일관성을 보장하기 힘든 문제이다. 다섯 교수의 채점 실험 조사 분석을 통해 통계적 측정 수치로도 이 사실은 증명되었다.

이 문항이 대학입학시험 문제로 적합한 것인가에 대해서는 채점 신뢰도 문제 이외에 내용 타당도에서 보아도 의문이 간다. 이 문항은 문법적으로 기껏 what이나 which의 의문을 유도하는 wh-determiner의 용법(예: *What time* is it? *Which house* do you prefer?)을 알아보는 것같다. 대화중 생활영어의 관용적 표현도 아니고 담화(discourse)적 중요성이 있는 표현도 아니어서 이 문항의 출제 의도는 그리 분명치가 않다.

주관식 문항의 채점 신뢰도를 높이기 위해서는 II장에 열거한 방법을 지켜야 하며, 특히 가능한 한 객관적 채점이 되도록 가능한 답의 범위가 넓지 않도록 출제되어야 한다. 논의해 오고 있는 이 주관식 1번 문항의 경우 빈 칸에 들어갈 적절한 '두 단어'를 넣으라고 했다면 응답의 변이가 대폭 축소되고 채점 신뢰도는 크게 높아졌을 것이다. 그렇다고 이 문항의 난이도가 크게 달라지지도 않을 것이다. 물론 두 단어 써넣기로 한다해도 What / Which concert? 이외 Whose concert? Which one? Who's playing? A concert? 등 여전히 세분된 채점기준은 필요할 것이다.

끝으로 좋은 시험 성공적인 시험을 위해서는 난이도, 변별도, 신뢰도, 내용 타당도 등 이외 채점 신뢰도가 최종적으로 중요한 요인이 됨을 간과해서는 안 된다. 각급 학교에서의 성취 시험이든 입학 및 취업 같은 경쟁 시험이든, 객관식 시험은 물론이고 특히 주관식 시험 문항의 경우 출제부터 채점에 이르기까지 언어 테스팅에 관한 전문적 지식과 경험이 필요하다. 그간 국가나 학교나 기관이 출제자로서 채점자로서 조심하고 꼼꼼하는데 소홀하지 않았나 싶다.

참고문헌

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Heaton, J. B. (1988). *Writing English language tests* (new ed.). London and New York: Longman.

Henning, G. (1987). *A guide to language testing: Development, evaluation, research.* Cambridge, MA: Newburry House.

Hughes, A. (1989). *Testing for language teachers.* Cambridge: Cambridge University Press.

Krzanowski, W. J., & Woods, A. J. (1984). Statistical aspects of reliability in language testing. *Language Testing, 1,* 1-20.