가정의학회지 제23권 제4호 별책 2002
J Korean Acad Fam Med Vol. 23, No.4 Suppl April

■ 세 미 나

문항 분석

황 인 홍
한림의대 가정의학교실

초 록

시험은 평가의 한 방법으로 순환하는 교육과정의 일부분이다. 한 번 치른 시험은 그것으로 끝이 아니며 다음 해 시험의 기초 자료로도 활용된다. 따라서 시험을 치르고 나면 시험 문항은 그 기능을 다한 것이 아니며 시험을 치른 후 이에 대한 평가가 필수적이다.

실제로 시험을 치르고 난 후 시험이 쉬웠는지 어려웠는지를 평가하며 좋은 문제와 좋지 않은 문제를 지적하는 것과 같은 과정을 객관적인 방법으로 세밀히 수행하는 것이 문항 분석이다.

어떤 시험이 얼마나 적절하였는지를 평가하기 위해서는 각 문항에 대해서 오답분석을 하하고, 문항 난이도와 문항 분별도를 파악해야 한다. 여기에서는 그 방법들을 알아보고, 최근에 발전된 문항반응이론의 개념을 소개하고자 한다.

시험을 치르고 성적을 산출하면 시험에 사용된 문항은 그 기능을 다 마친 것으로 생각하기 쉽다. 그러나 시험이란 평가의 한 방법으로서 순환하는 교육 과정의 일부분이므로 시험 성적이 결코 교육의 최종 목적이 될 수는 없다. 대부분의 교육이 그러하듯이 시험도 반복되는 것이 보통으로 한 번 치른 시험이 다음 해의 시험의 기초 자료로 활용된다. 따라서 시험을 치른 후 이에 대한 평가가 필수적이다.

실제로 시험을 치르고 나면 응시자나 출제자 모두는 이번 시험이 쉬웠는지 어려웠는지를 평가하며 좋은 문제와 좋지 않은 문제를 지적하는 것을 흔히 볼 수 있다. 이러한 의견을 나누는 것은 호기심에 대한 답을 구하는 것도 되겠지만, 교육이라는 측면에서는 그 시험의 적절성을 평가하는 과정이 된다. 이와 같은 과정을 객관적인 방법으로 세밀히 수행하는 것이 문항평가이다. 즉 문항평가(item analysis)는 시험을 구성하는 문항(item)을 단위로 하여 시험이 수행된 결과를 가지고 적절성을 평가하는 것이다.

어떤 시험이 얼마나 적절하였는지를 평가하기 위해서는 개별 문항에 대해서 구체적으로 다음과 같은 것들을 생각하여야 한다. (1) 각각의 답지(response)를 선택한 사람들이 얼마나 되는가? (2) 이 문항은 몇 명이나 맞았는가? (3) 수험생들이 이 문항에서 얻은 점수가 나머지 다른 문항에서 얻은 점수와 상관관계가 있는가? 문항 평가는 이 세 가지 질문에 대한 답을 구하는 것이며, 이들을 각각 오답분석(distractor analysis), 문항 난이도(item difficulty), 문항 분별도(item discrimination)라고 부른다.

1. 난이도 (Item difficulty)

난이도는 말 그대로 어떤 문항이 얼마나 어려운가

하는 정도를 나타내는 용어로, 그 의미에 어려움은 전혀 없다. 그러나 그 내면을 살펴보면 뜻밖에 복잡한 면이 있음을 알게 된다.

우선 어떤 문항이 쉬운가 어려운가 하는 것은 수험생의 주관적인 요인이 강하게 작용한다. 이러한 점은 특히 암기형 문항에서 두드러지는데, 그 문항의 답을 암기하고 있는 사람에게는 쉽고 암기하고 있지 않은 사람에게는 어려운 문제가 될 뿐이다. 이 경우 난이도는 문항의 특성이라기 보다는 수험생의 특성을 반영한다고 할 수 있을 것이다. 물론 문제해결형 문항이나 수리 계산과 같은 경우에도 답을 맞추기 위한 기본 전제를 알고 있는가에 따라 똑같은 이론이 적용될 수 있다.

또 다음과 같은 예를 통해 문항의 난이도를 생각해 보기로 하자.

예)

1) 34 + 16 - 15 =
2) 27 + 253 - 9 + 48 =
3) 박찬호는 누구인가?
4) 장기려는 누구인가?
5) 의과대학장의 역할은 무엇인가?
6) 졸업 시험에 있어서 학생담당 학장보의 역할은 무엇인가?

우선 문항 1)과 2)를 비교하면 누구나 2)가 1) 보다 더 어렵다고 생각할 것이며, 그 이유로는 2)가 1) 보다 더 복잡하기 때문이라고 할 것이다. 이 예에서는 복잡성이 문항의 난이도를 결정하였다.

이번에는 문항 3)과 4)를 비교하기로 하자. 4)가 3)보다 어렵지만 더 복잡한 것 같지는 않다. 굳이 어려운 이유를 찾자면 전문적인 지식이 요구되기 때문이라고 해야 할 것이다. 그러므로 어떤 특정 집단에 대해서는 3)이 4)보다 더 어려울 수도 있다. 여기에서는 전문성이 난이도를 결정하고 있다.

마지막으로 문항 6)은 5)보다 훨씬 복잡하게 보이지만 더 쉽다. 문항 작성의 잘잘못을 떠나서 문항 5)의 정답을 구하는 것은 쉽지 않아 보인다. 이 예에서는 얼마나 명확한지가 난이도를 결정한 것으로 보인다.

문항의 난이도를 결정하는 요인은 여기에 열거한 몇 가지 예 이외에도 매우 많은 것이라는 것은 쉽게 짐작할 수 있을 것이다.

이처럼 문항의 난이도는 그 문항의 고유한 특성(intrinsic character)을 가지고는 설명하기 어렵기 때문에 임의로 정의하여 사용할 수밖에 없다. 이렇게 정의되는 난이도 중 가장 흔히 사용하는 것은 item's p-value라고 하는 것으로 이것은 '어떤 문항에 대해 올바른 답을 한 사람의 비율'을 의미한다.[1)]

이 식에 따라서 난이도는 0에서 1 사이의 값을 가지며, 정답자가 많을수록, 즉 쉬운 문제일수록 1에 가까운 값을 나타낸다. 따라서 난이도 0은 모든 사람이 틀린 것을 의미한다.

한편 주관식 문항은 그 문항의 점수가 다양하게 분포하므로 다음과 같은 식을 사용하여 난이도를 계산한다.

(여기에서 Sx_n은 번호 n인 수험생이 그 문항에서 얻은 점수, Tx는 x 번 문항의 배점-즉 그 문항의 만점, N 은 전체 수험생 수를 나타낸다.)

시험의 기본적인 목적은 우수한 수험생과 그렇지 못한 수험생을 가리는데 있으므로 난이도가 0 이나 1인 문항은 시험의 목적을 달성하는데 전혀 기여를 하지 못한 것이 된다. 한편 수험생의 우열을 가리는데 가장 유리한 문항은 응시자의 반 정도가 정답을 맞추는 문제가 된다. 이런 이론에 따라 난이도 0.5 - 0.6을 최적범위(desirable range)로 인정하며, 0.3 - 0.7 사이를 허용범위(acceptable range)로 정하고 있다.

그러나 일정한 수준의 내용을 알고 있는 지를 평가하여 합격과 불합격만을 판정하는 자격시험의 경우에는 이 보다 다소 높은 범위의 난이도를 가진 문항을 사용하는 것이 목적에 맞는다.[2)]

또 문항 전체로 볼 때는 모든 문제가 일정 범위의 난이도를 가지는 것보다는 높은 난이도와 낮은 난이도의 문제가 적당히 섞여 있는 것이 전체적인 분별도를 유지하는데 도움이 된다.

2. 분별도 (Item discrimination)

시험은 대개 여러 개의 문항으로 구성된다. 여기에서 각 문항들이 측정하고자 하는 바는 이상적으로는 시험 전체가 측정하고자 하는 것과 같은 것일 것이다. 만일 어떤 문항이 시험 전체가 측정하고자 하는 것을 잘 반영하고 있다면, 그 시험에서 성적이 좋은 사람

은 그 문항에 대해서도 좋은 성적을 보일 것이며, 시험 성적이 나쁜 사람은 그 문항에서도 좋은 점수를 얻지 못할 것이라고 기대할 수 있다. 이러한 관계를 다른 각도에서 본다면 잘 만들어진 좋은 문항은 그 시험 전체에서 좋은 성적을 거둔 사람과 나쁜 성적을 거둔 사람을 분별해 낼 수 있다는 말이 된다.

이와 같이 어떤 문항이 그 시험 전체에서 측정하고자 하는 바를 얼마나 반영하고 있는가 하는 것을 그 문항의 분별도(item discrimination)라고 한다. 이와 같은 분별도를 표시해주는 방법으로는 분별도 지수(discrimination index)와 ITC(item-total correlation), inter-item correlation 등이 있다.

(1) 분별도 지수(discrimination index, D)

분별도 지수를 산출하는데는 여러 가지 방법이 있으나, 수험생의 수가 많은 경우에는 양극집단법(extreme groups method)을 사용한다. 이것은 시험의 결과 상위 성적을 얻은 수험생 집단의 난이도와 하위 성적을 얻은 수험생 집단의 난이도의 차이를 말하는 것으로 계산이 용이하다는 장점 때문에 가장 널리 쓰이고 있다.

이 방법으로 분별도 지수를 계산하기 위해서는 먼저 상위 집단과 하위 집단의 크기를 결정해야 하는데, 전통적으로 응시자의 27 퍼센트를 상위 집단 혹은 하위 집단으로 사용한다. 그러나 어떤 경우에는 33 퍼센트를 사용하기도 하며[3], 25 퍼센트와 35 퍼센트 사이의 어떤 값을 선택해도 비슷한 결과가 나온다고 알려져 있다.

분별도 지수는 이렇게 가려진 상위군의 정답자 비율에서 하위군의 정답자 비율을 뺀 값이다. 즉

여기에서 U는 상위 집단에 속한 수험생 중 정답자의 수, L은 하위 집단에 속한 수험생 중 정답자의 수, nU는 상위집단 수험생의 수, nL은 하위 집단 수험생의 수를 나타낸다. 일반적인 경우 상위 집단의 수와 하위 집단의 수는 같으므로 식 (3)은

와 같이 표시된다. 여기에서 n은 상위 집단의 수 혹은 하위 집단의 수이다.

이 식을 보면 상위 집단의 정답자 보다 하위 집단의 정답자가 많은 경우에는 분별도지수는 음수가 나올 수 있다는 것을 알 수 있다. 이런 이유로 해서 분별도 지수는 최고 1에서 최하 -1 사이의 범위를 가지는데, 0.35 이상이면 가장 좋은 것으로 평가되어 우수군(excellent)으로 분류한다. 또 0.25 - 0.35 는 양호(good), 0.15 - 0.25 는 경계(marginal)로 분류되며, 0.15 이하는 불량한 것으로 간주한다.

그러나 앞서 난이도 부분에서 언급한 바와 같이 의사 국가시험이나 전문의 자격시험과 같은 자격시험에 사용되는 문항은 이상적인 난이도를 가지고 있는 것이 아니므로 이에 따라 분별도도 다소 낮아지는 경향이 있고, 0에 가깝거나 음의 분별도 지수를 나타내는 경우도 적지 않다(그렇다고 그것이 허용될 수 있는 것은 아니다).

주관식 문항의 분별도는 (식 3)을 응용하여 구할 수 있다.

(여기에서 Dx는 x 번 문항의 분별도, N_U는 상위집단 수험생의 수, N_L은 하위 집단 수험생의 수이며 N_P는 상위진단과 하위 집단의 수가 같을 경우 $N_P = N_U = N_L$이다.)

(2) Item-total correlation

ITC는 그 문항에서 얻은 점수(대개 1 아니면 0)와 총점과의 상관관계를 말한다. 이것은 일반 통계에서 사용되는 상관관계와 동일한 것으로 사용되는 지표도 상관계수 r을 그대로 사용한다.

이 방법은 분별도 지수에서 상위 집단 혹은 하위 집단 내에 속한 모든 사람의 점수를 동일한 것으로 취급한 문제점을 보완할 수 있다는 장점이 있다. 또 상관계수가 우리에게 익숙한 것이므로 해석에 용이하다는 장점도 있다. 즉 어떤 문항의 Item-total correlation coefficient가 0.3이라면 이 문항이 총점의 9 퍼센트($r^2=0.09$)를 기여하였다는 것을 쉽게 알 수 있다는 것이다. 그리고 Item-total correlation은 그 시험의 신뢰도와 직접적인 관계가 있다는 장점도 있다.

이 방법의 유일한 단점인 계산이 복잡하다는 문제인데 최근에는 컴퓨터를 쉽게 이용할 수 있기 때문에 별 어려움이 없다.[4]

(3) Inter-item correlation

이것은 시험에 사용된 모든 문항간의 상관계수를 측정하는 것을 말한다. 이 분석을 해 보면 item-total

correlation이 높은 문항과 낮은 문항이 확연히 구별되는 것을 알 수 있다. 즉 item-total correlation이 높은 문항은 대부분의 문항과 높은 상관계수를 보인다는 것이다.

3. 오답 분석 (distractor analysis)

객관식 문항에 있어서 좋은 문항이라면 다음과 같은 속성을 갖추고 있어야 한다. 첫째 그 문항에서 묻고자 하는 것을 알고 있는 수험생은 항상 정답을 고를 수 있어야 한다. 둘째 모르는 수험생은 모든 답지(option) 중에서 무작위로 선택하는 형태가 되어야 한다. 오답 분석은 각각의 답지를 선택한 사람이 얼마나 되는지를 통하여 그 문항이 위의 두 가지 조건을 어느 정도 충족하는지를 평가한다.

예를 들어 문항의 오답 중에서 아무도 선택하지 않거나 극소수의 수험생만이 선택한 오답이 있다면 잘못 만들어진 문항이라고 할 수 있다. 이러한 오답은 문항을 쉽게 만들며 역할을 하며 시험 전체의 신뢰도를 저하시킨다. 반대로 너무 많은 수의 수험생이 선택한 오답도 좋은 것이 아니다. 이런 경우는 다음과 같은 두 가지 가능성에서 비롯된다. 첫째는 질문의 정답을 고르는데 필요한 내용 중 일부의 지식만을 가진 수험생이 많은 경우이다. 이런 경우라면 이 오답은 비교적 좋은 오답이라고 할 수 있다. 둘째로는 그 문항 자체가 교묘한 함정을 가진 문항일 경우이다. 이런 경우에는 문항 자체가 좋지 않다고 평가하여야 하며 문항을 다시 작성하는 것이 바람직하다(물론 예상문제집에 정답이 잘못된 경우도 있을 수 있다).

이와 같은 분석을 간단하게 표시해 줄 수 있는 지표는 개발된 것이 없다. 보통은 각 문항에 대해 답지별로 응답한 수험생의 수를 표시하는 문항반응분포(item response distribution)를 파악하여 오답 분석을 한다.

예) 문항 1)에서 4)까지의 반응 분포(* 정답)

문항	답지 ㄱ	답지 ㄴ	답지 ㄷ	답지 ㄹ	답지 ㅁ	합계
1)	12	58*	11	10	9	100
2)	73	0	12*	5	10	100
3)	96*	1	3	0	0	100
4)	14	18	13	31	34*	100
5)						

이 예에서 문항 1)은 난이도 0.58로 적절한 문제라고 생각되며, 오답자도 각 오답에 고루 분포하고 있음을 보여준다. 문항 2)에서 보면 답지 ㄴ은 아무도 선택하지 않아 이 문제는 실제적으로는 답지가 5개가 아닌 4개였음을 알 수 있다. 또 답지 ㄱ에 지나치게 많은 수험생이 응답을 하고 있음도 주목하여 그 원인을 찾아야 할 필요가 있는 문항이다.

문항 3)은 난이도 0.96으로 너무나 쉬운 문제가 되어 오답에 대한 분석을 할 여지가 없을 정도이다. 문항 4)는 다소 낮은 난이도를 보이면서 답지 ㄹ에 예상보다 많은 수가 응답을 보인 경우이다. 이 문항은 답지 ㄹ을 고친다면 적절한 난이도를 보일 것이라고 쉽게 예상할 수 있다.

4. 문항반응이론 (Item response theory)

고전적인 문항 분석 과정은 난이도와 분별도 등에 대해 좋은 정보를 제공해준다. 그러나 이것들은 시험의 점수에 초점을 맞추어 분석되는 것으로 문항의 속성이라고 하기보다는 응시자의 속성이라고 하는 것이 더 타당할 것이다. 이와 같은 자료를 통해 개인이 어떤 문항에 접하였을 때 어떤 반응을 보일지를 추정하는 것은 매우 어려운 일이다.

문항반응 이론은 경향이나 속성에 따른 개인 차이가 어떤 특정 문항을 접했을 때 개인의 반응에 어떻게 영향을 미치는지를 정확하게 설명하기 위하여 개발되었다. 이것은 개인의 능력과 그 사람이 문항을 맞출 가능성 사이의 수학적인 관계에 대한 몇 가지 가정에서 출발한다.[5]

(1) 문항 특성곡선

시험은 어떤 속성을 측정하기 위해 실시하는 것이며, 일반적으로 측정하고자 하는 속성을 많이 가지고 있는 사람일수록 그 문항을 맞출 가능성이 높다. 이것을 그림으로 표현하자면 그림 1과 유사한 형태가 될 것이다. 여기에서 X 축은 수험생의 능력을 나타내며 Y 축은 그 문항을 맞출 확률을 표시한다. 이 곡선을 보면 수험생의 능력이 가장 낮은 경우에는 문제를 맞출 확률도 가장 낮으며, 능력이 높아짐에 따라 문

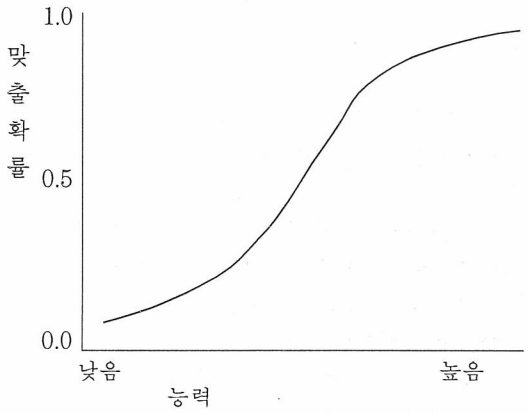

그림 1. 문항특성곡선의 예

제를 맞출 확률이 급격히 증가하다가 어느 정도 지나면 서서히 증가하는 모양을 보인다.

이와 같은 곡선을 문항특성 곡선이라고 한다. 이 곡선은 문항에 따라 그 모양이 달라질 것이며, 곡선의 모양을 다르게 만드는 것이 그 문항의 특성이라고 해석한다. 이 곡선의 일반형을 가정하고, 그에 대해 수학적으로 접근한 것이 문항반응 이론이다.

문항반응 이론에서는 문항은 수험생 집단의 특성에 따라 변하지 않는, 고유한 난이도와 분별도를 가진다. 또 이에 따라 수험생의 고유한 능력을 추정할 수 있다고 본다.

(2) 난이도

문항반응 이론에 따른 난이도는, 문항특성 곡선에서 그 문항을 맞출 확률이 0.5가 되는 수험생의 능력으로 정의한다. 그림 1의 곡선에서 Y 축의 값이 0.5에 해당하는 지점을 찾아 아래로 수직선을 그어 X 축의 값을 읽으면 그 값이 난이도가 된다.

문항특성 곡선에서 수험생의 능력은 평균(0)를 기준으로 한 z 값(평균과의 차이를 표준편차로 나눈 값)으로 표현되므로 난이도는 대략 -3 - +3 사이의 값으로 나타난다.

(3) 분별도

문항반응 이론에서 문항 분별도는 문항특성 곡선에서 난이도를 나타내는 점의 기울기로 정의 된다. 수험생의 능력 차이에 따라 맞출 확률이 급격하게 변한다면 문항특성 곡선은 가파른 모양을 보일 것이며, 기울기도 커질 것이다. 즉 분별도가 높은 값으로 표현되며, 최대값은 무한대이다.

만일 수험생의 능력과 맞출 능력이 음의 상관관계를 보인다면 분별도는 음수가 될 수 있다.

참 고 문 헌

1. KR Murphy. Psychological testing. 4th edition. New Jersey: Prentice-Hall; 1998.
2. Lord FM. The relation of the reliability of multiple-choice tests to the distribution of item difficulties. Psychometrika, 1952;17, 181-94.
3. Cureton EE. The upper and lower twenty-seven percent rule. Psychometrika, 1957;22, 293-6.
4. 황인홍. 분별도 지표로서 문항 총점 상관계수의 활용. 한국의학교육 2000;12(1):45-51.
5. RL McKinley. An introduction to item response theory. Measurement and evaluation in counseling and development. 1989;22:37-57.