설문 제작의 개요
- *설문 항목의 구성을 중심으로* -

성균관의대 강북삼성병원 가정의학과
신 호 철

각종 연구에서 자료 수집을 위해서 설문 조사나 면접 조사를 이용하는 것은 흔히 사용하는 조사 방법이다. 이런 상황에서 사용할 수 있는 연구 방법은 매우 다양하고 각각의 연구 방법은 모두 나름대로의 장점과 단점을 가지고 있지만 어떤 방법을 사용할 것인지는 연구 주제, 연구 대상, 그리고 연구를 수행하기 위한 재정 상태 등 여러 요인들을 고려하여 결정하게 된다. 하지만 여러 연구 방법들 중에서도 설문을 이용한 자료 수집 방법은 비교적 수행하기가 쉽고, 경제적이며, 지역에 관계없이 사용할 수 있는 등 여러 가지 장점이 있기 때문에 일차의료 연구에서 널리 사용되고 있다. 그렇지만 문제는 설문을 이용한 연구의 경우 사용되는 설문의 타당성 여부에 따라서 진행되는 연구의 성패가 결정되는 일이 많은데 비해서 정작 중요한 설문의 선택을 소홀히 하는 경우가 너무 많다는데 있다. 특히 기존 설문을 이용하지 않고 새로 제작하는 경우에는 많은 고려 사항이 있고 설문 제작에 대한 연구자의 경험이 매우 중요하기 때문에 이 기회에 이런 점들을 중심으로 간단히 알아보기로 하자.

설문 개발의 일반적인 순서

평가하고자 하는 내용에 따라서 달라질 수는 있지만 일반적으로 설문 제작의 순서는 다음과 같다.

⑴ 측정하고자 하는 내용을 분명하게 정의하고
⑵ 항목 집단(item pool)을 만들고
⑶ 평가 도구의 형식을 결정하고
⑷ 전문가의 의견을 구하고
⑸ 평가 도구의 항목을 구성하고
⑹ 예비 조사를 실시하고
⑺ 항목 분석(item analysis)을 실시하고
⑻ 1차 평가 도구의 완성
⑼ 평가 도구의 개정 작업을 실시한다

항목을 설정하기(writing items)

일단 측정하려는 내용 영역이 정해지면 각 개념에 맞는 항목들을 결정하게 되는데, 이 경우 폐쇄형 질문 형식(close-ended questions)을 사용하는 것이 장점이 많다. 대규모 연구의 경우 개방형 질문 형식(open-ended questions)은 자료 처리의 양이 많아지고 해석이 붙

가능한 응답이 생기는 경우가 종종 있다(특히 교육 수준이 낮은 응답자로부터). 게다가 개방형 질문들은 자료 등록시 시간이 많이 걸리고 그 코딩(부호화, coding)에 많은 경험을 필요로 한다. 따라서 특히 자기-기입식 설문조사(self-administered surveys)에서는 적당하지 않은 것이 보통이다. 하지만 예비 연구에서는 구성된 설문에서 누락되었을지도 모르는 내용들을 확인하기 위해서 개방형 질문을 병용하는 경우가 많다.

(1) 항목 내용(item content)

우선 가장 중요한 것은 '무엇을 알기 원하는가?'를 정확히 결정하는 일이다. 일단 알기를 원하는 내용이 확실하게 정해지면 그 개념을 측정하기 위한 항목의 내용을 결정하는데 빈도(frequency, 예를 들어 증상이 얼마나 자주 나타나는가? 전혀 없다 가끔 있다 자주 생긴다), 강도(intensity, 예를 들어 그 증상이 얼마나 심한가? 약하다 보통이다 심하다), 지속기간(duration, 예를 들어 그 증상이 얼마나 지속되는가? 수 분간 혹은 수 시간)을 묻을 것인가를 결정하는 것도 그 일부분이다. 기능에 어떤 제한이 있는지에 관한 질문도 건강 상태로 인한 기능 제한, 어떤 특별한 원인 혹은 원인에 관계없는 제한 사항이 있는지를 묻을 수 있다. 또 어떤 일을 하면서 문제가 있는지를 묻거나 혹은 도움을 필요로 하는지를 묻는 질문도 있을 수 있다. 하지만 항목을 결정할 때에는 그 내용이 원래 측정하고자 하는 개념의 정의와 일치해야 내용 타당도(content validity)가 높아지기 때문에 준비 단계에서 많은 생각을 해야 하는 것은 두말할 필요가 없다.

(2) 항목 간(item stems)

각 항목은 항목 간(item stem, 질문 내용)과 반응 옵션(response options, 응답 내용)으로 이루어진다. 항목 간은 응답자에게 묻고자 하는 내용이나 주제를 설명한 부분을 맡하는데 보통 짧고, 간결하며, 이해하기 쉽고, 한 가지 개념만을 묻는 내용이어야 한다. 그 의미를 분명하게 하기 위해서는 이중부정문이나 의미가 모호한 용어는 사용하지 않아야 한다.

(3) 항목 반응 옵션(응답 내용, item response options)

어떤 항목 반응 옵션을 사용할 것인가를 결정할 때에는 대략적인 간격 척도 수준의 정보를 제공할 수 있는 반응 옵션을 선택하는 것이 중요하다. 이런 목적을 위해서는 다음과 같은 3가지 중요한 특징이 고려되어야 한다.

첫째, 어떤 종류의 반응 간격(response intervals)을 사용할 것인가?
둘째, 중간 범주(middle 'neutral' category)를 적용할 것인가?
셋째, 몇 가지 반응 옵션을 사용할 것인가?

일반적으로 확인(endorsement), 빈도(frequency), 강도(intensity), 비교(comparison)와 같은 4 종류의 반응 옵션을 사용하게 된다(표 1).

예를 들어 확인(endorsement)의 경우 자신의 건강 상태에 대한 인식(예를 들어 '최근에 건강 상태가 나쁜 것을 느꼈다')을 평가할 때 사용될 수 있으며 빈도의 경우 여러 가지 주관적인 내용들을(예를 들어 활력이나 불안) 평가할 수 사용할 수 있다. 강도의 경우 통증과 같은 증상의 심한 정도를 평가하기 위해 사용할 수 있다. 강도의 경우 숫자로 된 반응 척도를 사용하기도 하는데, 예를 들어 환자에게 자신이 느끼는 통증의 정도를 평가할 때 1-20 사이의 숫자로 등급을 매기도록 하는 것이다. 이때 양쪽 끝의 숫자는 각각 '전혀 통증이 없음', '더 이상 상상할 수 없을 정도로 아픔' 등과 같은 의미를 부여한다. 가능하다면 다른 항목에서 사용된 반응 선택(response choices)과 비슷하게 구성함으로써 응답자들이 몇 가지 반응 선택에(limited set of choices) 친숙해지도록 하는 것이 바람직하다.

표 1. 확인, 빈도, 강도를 측정하는 반응 옵션들

확인(endorsement)	빈도(frequency)	강도(Intensity)
1. 확실히 그렇다	1. 항상	1. 없는
2. 그렇다	2. 대부분	2. 매우 약한
3. 잘 모르겠다	3. 자주	3. 약한
4. 아니다	4. 가끔	4. 보통
5. 확실히 아니다	5. 거의	5. 심한
	6. 전혀	6. 매우 심한

비록 많은 사람들이 다양한 반응들을 점점 증가하는 수준을 나타내는 용어들로 (최소한 순위 척도를 나타낼 수는 있도록) 표현할 수 있다는 것에 동의하고 있지만 그러한 부정확한 정량 방법들이 실제로 어떤 공통된 의미를 갖는지는 분명하지 않다. 오히려 Bradburn과 Sudman은 그 의미는 질문의 형식과 내용(context, 즉 어떤 종류의 질문이 주어지느냐에 따라서)에 따라 달라진다는 사실을 지적하고 있다. 따라서 반응자간의 해석의 차이를 줄이기 위해서는 표준화된 등록 절차(standardized administrative procedure)가 필요한 것이다. 만일 최종 측정이 단일 항목일 경우에는 반응 범주(response categories) 사이의 간격(interval)을 이해하는 것이 중요하다. 단일 항목의 경우 그 간격이 고르지 않을 가능성이 큰 습관적인 반응 선택(a 'custom' set of response choices)을 갖는 경향이 있기 때문이다. 예를 들어 건강 상태를 표현할 때 '더 이상 좋을 수가 없는(excellent)'과 '매우 좋은(very good)'의 차이는 '좋은(good)'과 '나쁘지 않은(매우 좋지도 않고 매우 나쁘지 않은, fair)'의 차이에 비해서 적을 수가 있기 때문에 이를 점수에 반영해야만 하는 것이다. 따라서 같은 개념을 긴 형태(long-form)로 측정할 수 있다면 각 반응 수준별 평균 점수(mean long-form scores)를 계산해서 각 범주 사이의 간격을 경험적으로 측정할 수도 있는 것이다. 확인(endorsemeet)을 하는 척도의 경우 중간(neutral) 범주(예를 들면, "잘 모르겠다"의 반응 선택이 특별한 질문에 대한 의견이 없는 응답자를 위한 배려이거나 부가적인 수준 측정을 위해서 주어지는 경우가 많다. 하지만 이 중간 범주에 대한 이견이 있기도 한데 일부 연구자들은(Converse and Presser, 1986) 응답자들이 의견을 표시하기 보다는 이 중간 범주의 반응 옵션을 선택하는 경향이 있다는 이유로 중간 범주의 반응 옵션을 주어서는 안된다는 주장을 하기도 한다.

항목 반응의 수와 관련되어 여러 연구의 결과는 5-7개 정도의 잘 선택된 반응 범주가 측정하고자 하는 내용을 적절하게 평가하기 위한 하한선(lower bound)을 제공한다는 것을 제시하고 있다. 또 다른 연구는 사람들은 감정(느낌, feeling)을 7가지 이상으로 분류할 수 없다는 것을 보여주고 있다. 결과적으로 7가지 이상의 분류는 불필요하다는 의견이다. 실제로 5-7가지의 반응 옵션으로 구성된 항목을 사용했을 때 그 이상의 많은 반응 옵션으로 구성된 항목을 사용했을 때와 강한 상관관계를 보이는 것으로 나타나 SF-36의 경우 주로 5-6가지의 반응 options을 사용했다.

(4) 다항목 설정(writing multiple items)

측정의 적절한 접근 방법인, 여러 항목을 한 가지 점수로 조합하는 방법은 여러 접근 방법을 여러 항목의 set로 이루어진 항목 내용으로 조합할 수 있게 한다. 만일 빈도가 여러 항목으로 이루어진 셑(set)의 반응 옵션으로 빈도가 선택되었다면 그 항목 간은 심한 정도(severity)의 범주(예를 들어, "우울한 생각이 든다", "자살하고 싶다")를 반영할 수 있을 것이다. 만일 강도가 반응 옵션으로 선택되었다면 보통 정도의 강도, 가장 나쁜 정도의 강도에 관련된 질문이 될 수 있을 것이다.

서로 다른 반응 옵션을(예를 들어, 빈도와 강도) 갖는 항목들을 조합할 경우의 이점은 방법 효과(method effect)가 감소된다는 것이다. 이 경우 서로 같은 수의 반응 옵션을 갖는 서로 동일한 분산(variances)을 갖게 될 가능성이 높고 조합된 점수에 동일하게 기여하게 된다는 점에서 도움이 된다. 반면에 서로 다른 반응 옵션을 갖는 항목들을 조합할 경우의 불리한 점은 전화상으로는 쉽게 사용할 수가 없다는 점이다. 전화 면담으로 조사하기 위해 심리적인 distress/well-being 항목을 선택할 경우에는 많은 항목 반응 셑(set)이 일정하게 되도록 수정을 해야 한다. 만일 서로 다른 형태의 반응 옵션을 그대로 두는 것이 더 중요한 경우에는 응답자들에게 면담 중에 다양한 옵션을 가진 카드가 읽혀질 것이라는 것을 미리 알려줄 수도 있다.

(5) 시간 설정(time frame)

대부분의 설문에 시간 설정을 해줄 필요가 있는데 특히 여러 단계의 강도나 빈도를 묻는 설문의 경우에 그렇다. 각 설문에서 묻는 사건에 대해서 충분히 정확하게 기억할 수 있는 정도의 짧은 시간, 그리고 주어진 시간 설정 안에 그 사건이 생길 가능성이 많고 사람마다의 다양성을 고려한 충분히 긴 시간의 설정이 필요하다.

대부분의 경우 설문들이 지난 4주간의 상황을 묻는 경우가 많은데 그 이유는 4주간의 기간은 대부분의 사람들이 건강과 관련된 사건(health events)을 쉽게 기억할 수 있고 그 사건에 대한 합리적이고 안정적인 표본을 얻을 수 있기 때문이다. 이 시간 설정은 매일 매일의 변화(dailly fluctuation)보다는 다양한 건강 상태의 평균적인 상태를 평가하는데 도움이 되기 때문이다. 이러한 시간 간격은 정신적인 distress/well-being을 측정할 때 가장 중요한데 작은 시간 간격의 설정은 오랜 기간 동안의 distress/well-being의 평균적인 수준보다는 "그날의 기분(mood of the day)"을 평가하게 될 가능성이 높기 때문이다. 잘 알려진 일부 설문

평가 도구의 경우 (예를 들어 MOS) "지난 한 달간"이라는 설명을 피했는데 이 경우 응답자들은 단순히 지난달(달력상의)로 생각할 가능성이 높기 때문이다. 마찬가지 이유로 "지난 30일간"이라는 설명도 피했다. 한 가지 예외가 있다면 정신적인 distress/well-being 항목의 경우 과거의 연구와 비교하기 위해서 "지난 한 달간"이라는 설명을 사용하였다.

항목의 pretesting

새로운 항목을 개발하거나 기존의 평가 도구를 개정할 경우 그 도구가 실제로 작용하는지를 확인하기 위한 사전 점검(pretest) 혹은 예비 조사(pilot test)가 필요하다. 어떤 연구자는 "사실상 모든 설문은 응답자의 편의를 위해, 연구자의 목적에 맞게 어떤 방법으로든지 변화시킬 수 있다."라고 말하기도 했다(Fowler, 1984). 전반적인 사용의 문제를 해결하기 위한 preliminary pretest는 십여 명의 사람을 대상으로 간단하게 시행할 수도 있는데 지침의 명확성(clarity of instructions), 질문이 공감을 형성하는지(make sense), 응답자의 부담(burden) 등을 평가하게 된다. 그리고 본격적인 예비조사(full-scale pilot study)는 본 연구와 비슷하게 50-200명 정도의 사람들을 대상으로 이루어지는데 일부 연구자는 적절한 pretesting에서 얻을 수 있는 점들과 몇 가지 지침(guidelines)을 제시하고 있다(Converse and Presser, 1986).

Preliminary pretest에서 주요한 정보의 자원은 대상으로부터 얻어지는 "debriefing"이라고 할 수 있는데 어떤 질문이 어렵고, 혼란스러우며(confusing), 불분명한지를 확인하고 대상들이 설문의 지침이나 생략 형태(skip pattern)를 잘 이해하는지를 알아보게 된다. 대상들에게는 그 설문 조사가 pretest라는 사실과 그 조사를 통해 설문의 문제점을 파악하려는 것이라는 사실을 알려주는데 이것을 소위 "participating pretest"라고 한다 (Converse and Presser, 1986). 본격적인 예비조사 연구에서는 각 항목의 통계적 분석을 통해서 정보를 얻기도 한다. 예비조사 연구의 목적은 도구마다 다르지만 일반적으로 사용상의 문제(administrative issues)와 많은 항목의 집단에서 가장 좋은 항목들을 경험적으로 선택할 수 있는지를 조사한다. 50-100명 정도의 표본 크기를 통해서 다양성(variability)이 떨어지는 항목과 반응을 하지 않는 비율이 높은 항목을 나쁜 항목으로 선정하여 그 항목을 개정하거나 제외하게 되고 지침이 불분명한 항목은 분명하게 수정하며, 항목간의 상관관계를 분석하여 같은 개념을 평가하기 위한 항목들과 강한 상관관계를 보이지 않는(not converge) 항목들을 확인하고, 반대로 분명히 다른 개념을 평가하기 위한 항목임에도 불구하고 그 항목과 너무 강한 상관관계를 보이는 항목들을 확인하게 된다.

항목들을 척도화하는 방법들(techniques for conbining items into scales)

여러 항목들이 한 가지 점수로 조합될 수 있는 것은 처음에 가설을 세워야 한다. 그 가설은 그 내용이 서로 같은 개념을 측정하는 것처럼 보이는 항목들의 논리적인 조합에 기초해야 한다. 일반적으로 항목의 조합을 이루는 가설을 평가하기 위해서 multitrait scaling을 사용하게 된다. 사실 어떤 개념들을 측정하기 위한 도구의 셑(set)을 개발하기 위해서 관련된 개념들의 셑(set)의 구조를 평가하는데 가장 흔히 사용되는 방법은 시험적(exploratory) 인자 분석이라고 할 수 있다. 이 방법은 항목들이 속해있는 차원(underlying dimension)을 이해하기 위한 초기 단계에서 항목들의 집합 속에서 어떤 관계들을 찾아내는데 적합한데 예

비에서 간혹 사용된다. 그러나 이 방법은 몇 가지 제한점을 갖고 있는데 우선 결과로 나타난 항목의 구조는 인자 모형(principal components 혹은 common factor analysis)과 관련된 선택, 적절한 인자의 수, 선택된 회전 방법(rotation method), 그리고 분석에 포함된 다른 항목들에 의해서 영향을 받는다. 즉 매번의 선택마다 이루어지는 결정이 인자 분석의 결과와 그 결과의 분석에 영향을 미치는 것이다. 게다가 이 방법에서는 변수들 사이의 관계가 명확하지 않으며 자료를 이루는 또 다른 이론적인 구조(alternative theoretical structures)를 점검하기가 불가능하다. 만일 개념에 대한 이론적인 진전이 시험적인(exploratory) 개발 수준 이상이고 하부 구조(underlying structure)에 관한 가설이 매우 좋다면 confirmatory analysis에서 더 많은 정보를 얻을 수 있다.

multitrait scaling

이 방법은 점수를 합산하는(summating rating) 전통적인 Likert 방법에 기초한다. 만일 여러 문항에 대한 반응들이 한 가지 척도 점수로 합산이 된다면 이것은 일반적으로 summated 혹은 "Likert-style" 척도라고 부른다. 이런 합산 척도(summated scale)는 가설로 이루어진 각 척도(hypothesized scale)의 항목들을 합산함으로써 이루어지는데 이때 각 항목에는 같은 비중이 주어진다. 간단한 예를 들어 보자. "지난 4주간 힘(에너지)이 넘치는 것을 느낀 적이 얼마나 있었습니까?"와 "지난 4주간 피곤함을 느낀 적이 얼마나 있었습니까?"라는 두 가지 항목이 있고 이 항목들은 "모두 전혀 없었다", "약간 있었다", "그저 그랬다", "자주 있었다", "항상 그랬다"라는 반응 옵션을 갖고 있고 이 반응 옵션에 대한 점수는 차례로 1, 2, 3, 4, 5점이라고 하자. 이 두 가지 항목 모두가 높은 점수가 에너지와 관련되게 하기 위해서는 두 번째 항목의 반응 옵션에 대한 점수를 반대로 한 후에 두 항목의 점수를 합산하면 되는 것이다. 가장 낮은 점수는 2점, 가장 높은 점수는 10점까지 나올 수 있을 것이다. 그렇다면 이러한 항목들의 셑(set)가 summated rating scale로 적절하게 조합될 수 있는지의 여부를 결정하는 여러 가지 분석이 이루어져야 할 것이다. Multitrait scaling에는 Likert 척도화에 필요한 기준 외에도 여러 가지 척도화의 기준이 추가되는데 일반적으로 다음과 같은 5가지의 단계가 필요하다.

(1) 가정된 항목의 집단안의 각 항목은 그 집단의 다른 항목들로 이루어진 전체 점수와 직선적으로 비례해야 한다(전통적으로 convergence의 기준은 보통 내적 일치도로 표현된다).
(2) 각 항목은 다른 어떤 개념(construct)보다도 그 항목이 측정하는 것으로 가정된 개념(construct)과 강한 상관관계를 가져야 한다(항목-판별 기준).
(3) 처음에 가정되지 않은 항목 집단이 자료에서 확인되지 않아야 한다(factor analytic test).
(4) 같은 척도를 구성하고 있는 항목들은 서로 그 개념(construct)에 대한 정보를 같은 정도로 제공해야 한다(test for approximately equal item-total correlation).
(5) 같은 개념(construct)을 측정하는 항목들은 서로 같은 분산(variances)을 가져야 하며 따라서 같은 척도로 조합되기 전에 표준화할 필요가 없어야 한다(equal variances criterion).

만일 가정된 집단안의 각 항목들이 위의 기준을 모두 만족시킨다면 한 가지 척도 점수를 구하기 위해서 각 항목들의 점수를 단순히 합산(혹은 평균)하는 것이 적절한 방법이 된다. 하지만 첫 번째, 두 번째 기준이 만족되지 않는다면 항목을 새로 구성해야 하며, 인자 분석

에서 애초에 가정되지 않았던 항목 집단이 나타난다면 다른 4가지 기준에 따라서 평가를 해 보아야 한다. 4번째 기준의 경우 각 항목이 실질적으로 전체 점수에 기여하는 한 철저히 지켜지지 않는 경향이 있지만 엄격한 기준이 적용되고 이 기준을 만족시키지 않을 때에는 각각의 항목마다 다른 가중치가 부여될 수 있다. 항목들은 그 분산이 유의하게 다를 경우에는 조합되기 이전에 표준화를 시켜야 한다. 분산의 동일성은 multiple range tests를 이용해 평가할 수 있다(Levy, 1975).

Multitrait scaling은 각 항목의 빈도, 평균, 표준 편차, 항목-척도간 상관관계(item-scale correlation, corrected for overlap), 척도의 내적 일치도 신뢰도 평가, 그리고 각 척도간의 상관관계 등을 조사한다. Multitrait scaling은 각 척도에 걸쳐 항목 판별을 조사함으로서 내적일치도를 검사하는 전통적인 방법 이상이라고 할 수 있다. 즉 항목들이 다른 개념(construct)에 비해서 어느 특정한(그 항목이 측정하는 것으로 가정된) 개념을 얼마나 잘 나타내고 있는지가 평가되는 것이다. Multitrait scaling 분석을 실행하기 전에 항목의 분산을 조사할 필요가 있다. 상호 비교 가능한 항목 분산을 찾고자 하는데 대략적으로 각 항목의 분포가 좌우 대칭적이고 표준 편차가 1.0에 가까우면 된다. 만일 이러한 조건들이 충족되면 가중치를 더할 필요가 없이 각 항목들을 바로 조합할 수 있다. 물론 이러한 기준에도 예외가 있는데, 만일 항목이 흔하지는 않지만 중요한 상태를 반영하고 있다면 항목 반응의 분포가 나빠도 건강 상태의 전 영역을 나타내기 위해서(represent fully a range) 그대로 두는 것이 바람직한 경우도 있다. 만일 항목들의 반응 옵션의 수가 많이 다르다면 각 항목들에게 같은 비중을 부여하기 위해서 각 항목을 조합하기 이전에 표준화를 시행하였다. 그러나 반응 옵션의 수가 같거나 하나 정도의 차이라면 그 분산이 심각하게 차이가 나지 않는다면 표준화를 시키지 않고 조합하였다.

항목-척도간 상관관계는 multitrait scaling의 기본적인 요소로 multitrait scaling 분석의 첫 번째 단계는 각 항목과 각 척도간의 상관관계를 조사하는 일이다. 모든 항목은 행(row)에 배치하고 모든 척도는 열(column)에 배치하여 서로의 상관관계를 조사하는데 매 행은 한 항목과 모든 척도간의 상관관계를 나타낸다. 물론 그 항목이 측정한다고 가정된 척도도 포함된다. 마찬가지로 매 열은 한 척도와 모든 항목간의 상관관계를 나타낸다. 그 항목이 포함되는 척도와의 상관관계에서는 지나치게 과장되는 피하기 위해 중복(overlap)을 수정한다. 만일 항목과 그 항목이 측정하는 것으로 가정된 척도와의 상관관계(수정된 상관 관계, corrected correlation) 0.3 이상인 경우에는 item convergence가 있다고 판단한다. 하지만 기존에 개발된 척도이거나 새로 개발하기 보다는 척도를 더 다듬기 위한 경우에는 보다 엄격한 기준인 0.4를 기준으로 하기도 한다. 일반적으로 위의 기준에 맞지 않는 항목은 모두 제거한다. Multitrait scaling의 두 번째 기준은 항목과 그 항목이 측정하는 것으로 가정된 척도와의 상관관계가 그 항목과 다른 척도들과의 상관관계보다 유의하게 높을 경우에 만족된다. 즉 어떤 항목과 그 항목이 측정하는 것으로 가정된 척도와의 상관관계가 같은 행의 다른 상관관계에 비해서 실질적으로 높을 경우에 항목 판별(item discrimination)은 인정되고 척도화가 성공적으로 진행된다. 성공적인("definite" scaling success) 척도의 경우는 항목과 그 항목이 측정하는 것으로 가정된 척도와의 상관관계가 같은 행의 다른 상관관계보다 두 표준오차(two standard errors) 이상 클 경우라고 정의할 수 있다. 만일 항목과 그 항목이 측정하는 것으로 가정된 척도와의 상관관계가 같은 행의 다른 상관관계에 비해서 낮을 경우에 분

명한 오류("definite" scaling error)라고 할 수 있으며, 만일 항목과 같은 행의 다른 척도와의 상관관계가 그 항목이 측정하는 것으로 가정된 척도와의 상관관계의 두 표준오차 이내에 있다면 가능한 오류("probable" scaling error)라고 할 수 있다. 계속적으로 오류를 항목은 척도의 구성에서 제외시켜야 한다. 그렇지만 가능한 오류의 경우에는 그 항목을 제외시킬지 혹은 포함시킬지는 분석 대상의 수, 척도를 구성하는 항목수, 내적 일치도 신뢰도, 포함된 개념간의 상관관계의 강도 등 여러 가지 요인에 따라서 달라진다. 만일 이론적으로 서로 관련이 있는 것으로 알려진 척도들이라면(예를 들어, 우울증과 불안증) 이 두 가지 척도내의 항목들이 보이는 가능한 오류는 적어도 척도를 개발하는 초기 단계에서는 참을 만하다. 그러나 척도 개발이 어느 정도 진행된 단계에서는 이 기능한 오류도 받아들일 수 가 없다.

단일항목 척도과 다항목 척도

많은 경우에 있어서 원하는 정보를 얻기 위해서는 한 가지 질문만으로도 충분한 경우가 보통이지만, 예를 들어 사람의 나이, 성별, 체중 등을 알기 위해서는 한 가지 질문으로 충분한 것처럼, 건강 및 건강 관련 문제에 관한 연구에서는 한 가지 질문으로 정의가 되지 않는 경우 그 개념은 복잡하고 어려워진다. 이런 경우 자연스럽게 평가하려는 내용을 '여러 각도에서 비추어 보기' 위한 여러 항목의 질문들이 필요하게 되고, 예를 들어 '우울증'의 정의는 우울해지고, 기분이 저조해지고, 희망이 없어지는 것과 같은 다양한 성분(내용)들을 포함하는 것처럼, 결과적으로 그 정의의 모든 성분(내용)들을 나타낼 수 있는 여러 가지 질문들이 필요한 것이다.

이런 다항목 측정(multi-item measures)은 단일항목 측정(single-item measures)에 비해 여러 가지 장점들을 갖는다. 첫째, 각 변수를 정의하기 위해 필요한 최종 점수의 숫자를 줄일 수 있다. 둘째, 항목들이 서로 공통적으로 갖고 있는 정보를 공유하고 모아놓음으로써 점수의 신뢰도(score reliability)를 높인다. 셋째, 점수의 다양성(score variabilty)을 증가시킴으로써 민감도(sensitivity)를 증가시킨다. 넷째, 측정하고자 하는 개념에 관한 정보를 보다 잘 나타냄으로써 타당도(validity)을 증가시킨다. 다섯째, (긍정적인 항목과 부정적인 항목들이 섞여 있을 때에는) 내용에 관계없이 각 항목에 긍정적으로 혹은 부정적으로 반응하는 개인의 성향 때문에 생기는 왜곡을 최소화할 수 있다. 여섯째, 만일 결손 항목(missing response)이 있을 경우 다른 항목들을 이용하여 점수를 추정할 수 있어 결과적으로 결손 점수(missing score)를 줄일 수 있다.

참고문헌

Stewart AL, Hays RD, Ware JE Jr.: Mehods of constructing health measures. In: Stewart AL, Ware JE Jr, eds. *Measuring functioning and well-being: The medical outcomes study approach*. Durham, NC: Duke University Press, 1992:67.

Boynton PM, Greenhalgh T. Selecting, designing, and developing your questionnaire. *BMJ* 2004, 328: 1312-15.