연구논문

설문지의 길이가 응답의 질에 미치는 영향*

The Influence of the Questionnaire Length on Response Quality

김권현[a] · 유동주[b] · 김형준[c] · 김청택[d]
Kwonhyun Kim · Dongjoo Yoo · Hyungjoon Kim · Cheongtag Kim

본 연구에서는 설문조사의 비표집 오차의 원인 중 하나인 설문지의 길이가 응답에 미치는 영향을 조사하였다. 연구 1에서는 동일한 문항을 설문지의 전반부 혹은 후반부에 배치하는 실험계획을 통해 선행연구에서 밝혀진 설문지의 길이 효과를 재연하고 설문의 질을 나타낼 것으로 예상되는 통계치(개인별 엔트로피, 렘펠-지브 복잡도, 신뢰도, 문항반응이론 기반 개인우도 등)가 문항의 위치에 따라 어떻게 변하는지 조사하였다. 연구 2에서는 이 결과를 현장에서 실제로 얻은 설문 결과에 대해 적용하였다. 연구 1의 결과, 후반부 문항들의 개인별 표준편차, 개인별 엔트로피, 렘펠-지브 복잡도가 유의미하게 낮아졌다. 하지만 리커트 문항으로 이루어진 척도 점수에 대한 신뢰도는 큰 변화가 없었으며, 분기질문에 대한 “아니오” 응답의 비율은 낮아지는 경향을 보였지만 유의미한 차이를 보이지 않았다. 개방형 질문에 대한 응답의 길이는 후반부의 응답에서 짧아졌다. 연구 2에서는 현장에서 실제로 얻은 설문지 중 설문소요시간의 중앙값이 51분인 긴 설문지에서

* 이 논문은 한국리서치의 지원을 받아 연구되었음.

a) 서울대학교 협동과정 인지과학 박사과정
b) 서울대학교 심리학과
c) 서울대학교 협동과정 인지과학
d) 교신저자(corresponding author): 서울대학교 심리학과 교수 김청택.
E-mail: ctkim@snu.ac.kr.

만 개인별 엔트로피, 렘펠-지브 복잡도가 낮아지고 최대 한줄 응답 길이가 유의미하게 증가하였다.

주제어: 설문조사, 비표집오차, 설문지의 길이, 설문 응답

The present study investigated the effects of the questionnaire length upon response quality. In study 1, we manipulated the location of items in a questionnaire by putting the identical items either at the first part of the questionnaire or at the last. The qualities of responses to items at different locations were measured using several statistics (individual entropy, Lempel-Ziv complexity, reliability, respondent's likelihood based on IRT) which were expected to reflect the response quality. In study 2, we tested whether the item location effect is also observed in real data. The results of study 1 show that individual standard deviation, individual entropy, Lempel-Ziv complexity decrease as the same scales were placed towards the end of the questionnaire. But the reliability and response to filtering questions showed no significant difference. The length of answers to open questions went shorter as the questions were placed towards the end. In study 2, we could find that the individual entropy and the Lempel-Ziv complexity decreased and maximum length of straight response increased in the long survey in which the median duration was 51 minutes.

Key words: nonsampling error, questionnaire length, response quality

Ⅰ. 서론

설문조사에서 발생하는 오차는 비표집오차와 표집오차로 구분할 수 있다. 표집오차는 통계학적으로 추정할 수 있고, 표본의 크기를 증가시킴으로써 통제할 수 있지만, 비표집오차는 그 존재와 크기가 정확하게 추정되기 힘들고, 추정된다고 하더라도 이를 통제하는 것이 용이하지 않다. 이러한 점에서 조사의 질을 평가하는 데 있어서 비표집오차가 표집오차보다 더 중요할 수 있다(Weisberg 2009).

비표집오차의 주요 원인으로는 특정화 오류(specification error), 틀 오류(frame error), 무응답 오류(nonresponse error), 측정 오류(measurement error), 그리고 처리 오류(processing error) 등이 있다(Biemer 2010). 특정화 오류는 설문지의 질문들이 설문의 목적을 제대로 반영하지 못할 때 나타나고, 틀 오류는 표본을 추출하는 데 사용하는 틀을 잘못 구성하거나 사용하기 때문에 나타난다. 처리 오류는 코딩, 자료 편집, 분석 등에서 발생하며, 측정 오류는 인터뷰어, 질문 구성(question wording), 설문지(questionnaire)등에서 발생한다. 이 중 설문지의 길이는 측정 오류의 한 종류인데 본 연구에서는 설문지의 길이가 응답의 질에 미치는 영향을 조사하고자 한다.

설문지의 길이가 길어짐에 따라 응답자의 피로감은 증가하는 반면, 흥미는 감소하여 응답의 질이 떨어질 것을 예상할 수 있다. 그러나 설문조사를 시행하는 연구자/기관은 신뢰도 증가와 비용 감소, 편의성 증가 등의 이유로 긴 설문을 선호하는 경향이 있다. 설문지가 길어짐에 따라 응답의 질이 저하된다면 설문의 길이를 증가시키는 것이 이득만을 유발하는 것이 아니라 손실을 유발할 수도 있다. 따라서 설문지의 길이에 따라서 응답의 질이 변화하는지, 변화한다면 어떠한 형태로 변화하는지를 조사하는 것은 적절한 설문지의 길이를 설계하는 데 중요한 정보를 제공해 줄 것이다.

본 논문에서는 설문의 길이와 그에 따른 응답의 질에 대한 선행연구의 결과를 살펴보고, 응답의 질을 반영할 수 있을 것이라고 예상되는 통계치와 지수들을 제안할 것이다. 연구 1에서 동일한 내용의 문항을 설문지의 전반부 혹은 후반부에 배치한 실험설계를 통해 문항의 위치 혹은 설문지의 길이가 응답에 어떻게 영향을 미치는지를 살펴보았다. 연구 2에서는 연구 1에서 연구된 지수들을 현장에서 사용된 설문자료에 적용시켜 연구 1의 결과가 일반화될 수 있는지를 조사하였다.

Ⅱ. 질문지의 길이와 응답의 질

1. 설문지의 길이가 응답에 미치는 영향

설문지가 길어질수록 다양한 측면에서 응답이 변화한다는 연구들이 보고되어 있다. 설문지가 길어질수록 동일한 응답을 반복하는 비율과(Herzog & Bachman 1981) “모르겠다”는 응답의 비율이 증가하며(Deutskens et al. 2004), 개방형 질문에 대한 응답의 길이가 짧아지고, 무응답의 비율이 증가한다(Galesic & Bosnjak 2009). 분산에 대해서는 다소 상반되는 결과를 산출하기도 하였는데, Galesic & Bosnjak(2009)은 설문의 길이가 길어질수록 응답의 분산이 작아짐을 발견한 반면, Deutskens et al (2004)은 설문의 길이가 응답의 평균과 분산에 영향을 미치지 않음을 발견하였다.

Herzog & Bachman(1981)은 또한 설문지 길이 효과가 문항의 형태에 따라 달라짐을 발견하였다. 문항 세트(item set)의 경우에는 설문지에서 뒤에 위치할수록 평균의 차이가 증가하지만, 단일 문항의 경우에는 문항의 위치에 따라서 일관적인 평균의 차이를 내지 않았다. 이를 연속된 문항에 대해 동일한 응답을 하는 “한 줄로 응답하기” 때문이라고 설명하였다. 동일한 문항 세트 안에서 문항 응답들의 상관계수를 비교했을 때, 설문지의 뒤쪽에 위치할수록 상관계수가 커짐을 발견하였고, 이를 ‘한 줄로 응답하기’에 기인하는 것으로 해석하였다.

특정한 응답을 할 때에만 그 다음 질문이 제시되는 분기질문의 경우에도 문항위치의 효과가 관찰되었다. 분기질문이 뒤에 위치할수록 추가 질문을 회피하는 방향으로 응답하는 경향을 보였다(Jensen et al. 1999; Duan et al. 2007).

2. 응답의 질

앞에서 살펴본 여러 연구들은 설문지가 길어질수록 반응의 질이 저하된다는 가정을 하고 있다. 그 가정에 따르면 질문의 위치에 따라 차이가 나는 통계치는 반응의 질의 차이를 반영한다고 볼 수 있다. 그러나 각 연구들에서 사용되는 반응의 질이 다르게 정의되므로 이들의 결과를 통합하거나 비교하기가 쉽지 않다. 여기서는 응답의 질에 대한 정의를 개관하고 다양한 측면에서 반응의 질을 살펴보고자 한다.

Goetz et al.(1984)은 응답의 질을 완결성(completeness)과 정확성으로 구분했다. 완결성은 설문지의 질문들에 빠짐없이 응답하는 것을 말하며, 정확성은 응답자가 자신의 태도나 의견 등을 왜곡하지 않고 반응하는 것을 말한다. 완결성을 반영하는 통계치로는 개방형 질문에 대한 응답의 길이와 선다형 질문에 대한 무응답과 “모르겠다” 응답의 비율 등이 있다.

응답은 여러 가지 이유(예, 사회적인 바람직성, 자기방어)로 왜곡될 수 있기 때문에 정확성을 반영하는 통계치는 다양하고 복잡하다. Goetz et al.(1984)은 이를 측정하기 위하여 동일한 설문에 대한 동질한 집단의 응답의 분포를 비교하는 방법을 사용하였다. 만약 동질적인 두 집단이 동일한 설문에 대해 응답의 분포가 다르다면 한 집단 혹은 두 집단의 반응에 왜곡이 있는 것이다. 이때 응답의 분포는 평균과 분산으로 측정할 수 있다.

Nichols et al.(1989)은 응답의 정확성을 훼손하는 왜곡을 내용무관반응 왜곡과 내용반응성 왜곡으로 구분하였다. 내용무관반응 왜곡은 질문의 내용과 무관하게 반응하여 부정확하게 반응하는 현상이며, 내용반응성 왜곡은 반응이 설문의 내용에 의해 영향을 받으면서 부정확한 반응을 하는 현상이다. 예를 들면, 진단용 심리검사에서 진단을 회피할 목적으로 거짓 응답을 하거나,

사회적으로 바람직하게 보이려고 의식적 혹은 무의식적으로 거짓 응답을 할 수 있다. 내용무관반응과 내용유관반응의 특성을 조사하기 위해 연구자들은 실험 참가자에게 설문지를 보지 않고 응답을 하게 하거나, 특수한 목적(예, 사회적으로 바람직하게 보이기)을 부여하여 응답을 하게 한 다음 정상 응답과 비교한다(장은경 2010; Baer et al. 1997).

3. 응답의 질을 반영하는 통계치(지수)

본 연구에서는 응답의 질을 정확성, 일관성, 성실성의 세 측면으로 구분했다. 정확성은 응답에 왜곡이 존재하는지를 나타낸다. 동일한 설문에 대한 동일한 집단의 반응이 전반부와 후반부에서 일치하지 않는다면 어떤 식으로든 왜곡이 존재하는 것으로 판단할 수 있다.

일관성은 정확성의 하위 개념으로 유사한 문항들이 제시된 경우에 응답자의 반응이 얼마나 일관적인지를 나타낸다. 이 논문에서는 하나의 구성체(예컨대, 시험불안)을 측정하기 위하여 다수의 문항을 사용하는 문항세트를 다루고 있으므로 이를 정확성과 구분하였다. 일관성은 논리적 일관성, 집단 내 개인 일관성, 집단 일관성으로 구분해 볼 수 있다. “당신은 행복하십니까?”란 질문에 “5. 매우 그렇다”라고 답한 응답자가 “당신은 행복하지 않습니까?”란 응답에도 “5. 매우 그렇다”라고 대답한다면 논리적으로 일관적이지 않은 응답을 한 경우이다. 반면 “나는 감기에 자주 걸린다.”란 질문에 “5. 매우 그렇다”라고 답한 거의 모든 응답자가 “나는 감기를 예방하기 위해 노력한다.”란 질문에 대해서도 “5. 매우 그렇다”라고 답했을 때, 그와 반대로 답한 응답자는 집단 내 개인 일관성이 낮다고 말할 수 있다. 집단 내 개인 일관성은 논리적으로 모순된 경우는 아니지만 집단의 다른 사람들에 비추어 볼 때, 응답의 패턴이 비정상적으로 보이는 경우이다. “나는 감기에 자주 걸린다”란 질문과 “나는 감기를 예방하기 위해 노력한다.”란 질문에 대한 응답의 상관관계를 통해 일관성을 측정할 수 있다. 만약 두 응답의 상관관계가 높다면 집단의 일관성이 높은 것이고, 그렇지 않다면 집단의 일관성이 낮은 것이다. 어떤 척도가 하나의 구성체를 측정하기 위한 여러 문항으로 구성되었을 때, 집단 일관성

은 높을 것으로 예측할 수 있다.

본 연구에서 사용한 IRT 기반 개인 우도와 MRF 기반 개인 우도는 집단 내 개인 일관성을 측정하고, 고전 검사 이론의 신뢰도는 집단 수준의 일관성을 측정한다.

성실성은 응답자가 얼마나 성실히 설문에 응답했느냐를 나타낸다. 이는 완결성과 밀접하게 관련되어 있는 개념으로, 본 연구에서는 무반응의 비율과 개방형 질문에 대한 응답의 길이가 성실성을 반영한다고 가정하였다. “시험을 치기 전에 불안을 느꼈던 경험에 대해 적어주세요.”와 같은 개방형 질문에 대해 간단하게 단답형으로 대답을 할 수도 있고, 그 상황을 자세히 서술할 수도 있다. 개방형 응답은 길이에 상관없이 동일한 내용을 나타낼 수도 있으므로 응답의 길이가 긴 것이 항상 정확하거나 자세한 응답이라고 말할 수는 없다. 하지만 그 길이를 통해 얼마나 설문에 성실히 응답하지를 정량화할 수 있다고 가정하였다. 무반응의 경우는 반드시 부정확한 응답이라고 분류할 수는 없지만, 여기에서는 무반응의 비율을 응답자의 성실성의 지수로 해석하였다.

1) 응답의 정확성을 검증하기 위해 사용한 통계치

본 연구에서는 정확성을 위하여 평균과 표준편차뿐 아니라 다양한 통계치를 통해 응답의 분포에 대한 좀 더 정밀한 분석을 시도하였다. 엔트로피, 렘펠-지브 복잡도는 정보이론적 관점에서 응답 분포의 특성을 나타내고, 최대한줄 응답 길이와 중간 반응의 비율은 기존 연구에서 길이 효과를 반영한다고 알려진 통계치이다. 그리고 문항별 반응의 통계치뿐만 아니라 개인별 반응의 통계치도 분석하였다. 분기질문에 대한 긍정 응답의 비율 역시 응답의 정확성을 검증하기 위한 통계치로 사용되었다.

엔트로피

엔트로피는 반응에 대한 변산성을 나타낸다. 주어진 선택지 중에서 하나의 선택지에만 모두 반응하면 엔트로피가 가장 낮으며, 모든 선택지에 같은 비율로 반응하면 엔트로피가 가장 높게 된다. 이 지수는 질문의 내용과 상관없이 동일한 선택을 하는 반응을 탐지할 있다. 표준편차와 유사한 특성을 측정

하지만, 중심(centrality)을 상정하지 않는다는 점에서 다르다.

렘펠-지브 복잡도

렘펠-지브 복잡도는 N개의 문항에 대하여 몇 가지 패턴으로 응답하였는지를 나타낸다. 예컨대 다섯 문항이라면 01010과 같은 숫자열로 표시될 수 있는데, 이때 사람들이 반응한 패턴이 몇 개 있는지를 헤아리는 것이다. 이때 단순히 서로 다른 패턴만을 헤아리지 않고 반복되는 패턴은 하나로 취급한다. 이 복잡도가 낮다는 것도 질문의 내용과 관계없이 동일한 패턴으로 대답한다는 것으로 정확성과 관련 있는 지수이다.

최대 한줄 응답 길이

연속된 리커트 척도의 문항에 대해 동일한 번호의 응답을 연속적으로 하는 것을 Herzog & Bachman(1981)은 직선 형태(straight-line pattern)의 응답이라고 불렀다. 본 논문에서는 이런 반응을 “한줄 응답”이라고 칭했으며, 주어진 척도에서 가장 긴 한줄 응답의 길이를 “최대 한줄 응답 길이”라고 칭하였다.

중간 반응의 비율

중간 반응이란 리커트 척도 문항에서 중간에 위치한 반응범주를 선택하는 반응을 의미한다. 본 연구에서 사용한 5점 리커트 척도 문항에서 “3. 보통이다.”를 선택하는 것은 “중간 반응”이다. Herzog & Bachman(1981)은 동기가 저하된 응답자는 좀 더 쉽게 응답할 수 있는 방법을 찾는데 많은 사람들이 선택하는 응답을 하거나 긍정도 부정도 아닌 응답을 한다고 보고하였다.

2) 응답의 일관성을 나타내는 지수

응답의 일관성은 크게 논리적 일관성, 집단 일관성, 집단 내 개인 일관성으로 나눌 수 있다. 본 연구에서는 집단 일관성과 집단 내 개인 일관성을 나타내는 고전 검사의 신뢰도, IRT 기반 개인 우도, MRF 기반 개인 우도를 사용하였다.

고전 검사의 신뢰도

고전 검사 이론에서 신뢰도는 측정된 검사점수의 전체 분산에서 진점수의 분산이 차지하는 비율이다. 본 연구에서는 여러 가지 신뢰도를 측정하는 방법 중에서 Cronbach(1951)가 제시한 Cronbach-α를 사용하였다. 신뢰도가 높다는 것은 동일한 구성체를 측정하는 문항들에 대한 반응들이 일관성이 있다는 것을 의미한다.

IRT 기반 반응자 우도

IRT는 척도나 검사에서 각 문항의 반응을 난이도, 변별도, 개인의 능력(태도)의 파라미터를 가지는 로지스틱 함수에 의하여 예언하는 통계적 모형이다. 의견이나 태도를 측정하는 경우의 난이도 파라미터는 얼마나 강한 태도를 지녀야지 "예"라고 대답하는지를 나타낸다. 반응자 우도는 개인의 반응을 IRT 모형이 잘 예측할 수 있는지를 나타낸다. l_z로 표시되는 이 통계치는 적절하지 않은(aberrant) 반응을 탐지하는 데 보편적으로 사용되고 있으며, 많은 특질에 관한 연구들에서 안정적이고, 일관된 결과를 보인다(Reise & Due 1991; Reise & Waller 1993). 그리고 l_z는 무작위 반응이나 추측 등 여러 형태의 반응 조건에서도 다른 지수들보다 더 잘 작동한다(Birenbaum 1985; Li & Olejnik 1997; Nering 1997). 일반적으로 l_z가 -2보다 낮은 사람들을 일탈(aberrant) 반응 패턴으로 간주한다(Ferrando & Chico 2001; Schmitt et al. 1999).

MRF(Markov Random Fields) 우도

MRF는 마코프체인을 일반화한 것으로, 각 응답은 그 전의 응답과 그 후의 응답에 의하여 확률적으로 예측된다고 가정하는 모형이다(자세한 모형은 부록을 참조하라). 이 모형에 의하여 개인별 우도를 내면, 이는 다른 사람의 응답에서 일탈된 반응 패턴인지에 대한 판단을 가능하게 한다. 이 모형은 IRT의 국지독립성(local independence)과 같은 통계적 가정을 하지 않기 때

문에, 독립성이 위배되는 한 줄로 응답하기와 같은 경우에도 처리가 가능한 모형이다. 김형준 등(2014)에 의해 이 방법이 MMPI 검사에서 내용무관반응을 잘 탐지할 수 있음이 발견되었다.

3) 응답의 성실성을 나타내는 통계치

앞에서 설명한 것처럼 무반응의 비율과 개방형 질문에 대한 응답의 길이가 응답의 성실성을 반영하는 통계치로 가정되었다.

Ⅲ. 연구 1: 문항의 위치에 따른 응답의 질의 차이

연구 1은 설문지의 길이에 따른 반응의 질을 평가하기 위하여, 동일한 내용의 문항을 설문지의 전반부 혹은 후반부에 배치하여 문항의 위치에 따른 응답의 질의 차이를 조사하였다. 질문지의 길이 효과는 질문지의 길이를 조작한 다음 반응의 질을 측정해야 되나, 본 연구에서는 동일한 길이의 설문지에서 제시 위치에 따른 응답의 질의 차이를 분석하여 설문지의 길이 효과를 측정하였다. 동일한 문항의 위치가 i 번째일 때와 i + j 번째일 때의 응답의 차이가 i 개의 문항으로 구성된 질문지와 i + j 개의 문항으로 구성된 질문지에서의 응답의 차이를 반영한다고 가정하였다.

연구 1은 두 개의 실험으로 구성되었는데, 실험 1의 설문지는 리커트 5점 척도의 응답 문항 74개, 개방형 문항 8개, 분기질문 6개로 구성되어 있고 예상 소요 시간은 30분이었다. 실험 2의 설문지는 리커트 5점 척도의 응답 문항 413개, 개방형 문항 20개, 분기질문 3개로 구성되었으며 예상 소요 시간은 1시간이었다.

1. 연구방법

1) 연구대상

피험자는 무급 피험자와 유급 피험자로 나눠 모집하였다. 무급 피험자는

○○대학교 심리학과 강의를 듣고 있는 수강생을 대상으로 모집하였으며, 유급 피험자는 ○○대학교를 제외한 대학들의 커뮤니티 사이트의 홍보를 통하여 모집하였으며, 정해진 참여비가 지급되었다.

2) 측정도구

본 연구의 목적이 '설문의 길이에 따른 응답의 질'의 변화이므로 설문의 내용은 크게 중요하지 않지만, 설문의 길이와 내용이 혼입(confounding)될 수 있기 때문에 실험 조건에 따른 설문의 내용이 크게 다르지 않도록 설계되었다. 동일한 구성체를 측정하는 심리 척도를 전반부와 후반부에 배치하였으며, 설문지 유형(A형/B형)에 따라 구체적인 심리척도의 배열순서가 달라졌다. 예를 들어, A형 설문지는 전반부에 Spielberg et al.(1980)의 시험불안 척도(Test Anxiety Inventory), 후반부에 황경렬(1997)의 시험불안 척도(K-TAI-K)가 배치되어 있고, B형 설문지는 전반부에 황경렬(1997)의 시험불안 척도, 후반부에 Spielberg et al.(1980)의 시험불안척도가 배치되어있다(표1 참조). 설문조사는 온라인으로 진행되었다.

실험 1에서 사용된 문항들은 <표 1>과 같다. 리커트 5점 척도 문항들은 Spielberg et al.(1980)의 시험불안척도, 황경렬(1997)의 시험불안 척도, 김지혜(1991)의 자기의식 척도, 김득란(1993)의 자기의식 검사, 김아영(2002)의 학업동기 척도, 그리고 이훈진(1997)의 자기개념 척도였다. 문항의 수를 통제하기 위해 일부 문항들은 제외되었다. 사용된 척도는 모두 보고된 신뢰도가 0.8 이상이었다. 개방형 질문은 대학 생활과 관련된 내용(시험, 아르바이트 등)이었다. 분기질문은 누구나 "예"라고 답할 수 있을 법한 질문들(예, "지난 3개월 동안 책을 산 적이 있습니까?")로 구성되었다. 만약 분기질문에 대해 "예"라고 답한다면, 그와 관련된 선다형 질문 5개에 대해 추가로 답해야 하였다. 피험자들은 무작위로 A형 또는 B형 설문조사 조건에 할당되었으며 A형과 B형의 두 설문조사에 포함된 문항은 동일하지만 <표 1>에서 제시된 바와 같이 순서만 달랐다.

실험 2에서 사용된 문항들은 <표 2>와 같다. 실험 2의 리커트 5점 척도 문항들은 Spielberg et al.(1980)의 시험불안척도, 황경렬(1997)의 시험불안 척도, 서봉연(1975)의 자아정체감 척도와 박아청(1996)의 자아정체감 척도, Beck et al.(1974)의 무망감 척도와 이영호(1993)의 무망감 척도, 학업동기 척도(김아영 2002), 인지욕구 척도(김완석 1994), 승인욕구 척도(금명자 1984), 사회적 문제해결능력 척도(김화자 1998), 부정적 평가에 대한 두려움 척도-단축형(최정훈・이정윤 1994), 이훈진(1997)의 자기개념 척도, 그리고 NEO-PI-R 인성검사 중 성실성 척도(전미현 2013)였다. 실험 2에서 사용된 척도 역시 보고된 신뢰도는 모두 0.8 이상이었다. 개방형 질문은 실험 1의 개방형 질문과 동일하였으며, 문항 수를 늘리기 위해 귀인양식 질문지(이영호 1993)의 12 문항이 추가되었다. 분기질문은 실험 1과 동일하였다.

<표 1> 실험 1의 설문조사 유형에 따른 문항의 구성

A형 설문조사의 구성 (문항의 개수)	B형 설문조사의 구성 (문항의 개수)
TA1(20)	TA2(20)
SC1(17)	SC2(17)
개방형질문 1(4)	개방형질문 2(4)
분기질문 1(3)	분기질문 2(3)
학업동기척도(44)	학업동기척도(30)
자기개념척도(30)	자기개념척도(44)
개방형질문 2(4)	개방형질문 1(4)
분기질문 2(3)	분기질문 1(3)
SC2(17)	SC1(17)
TA2(20)	TA1(20)

주. TA1 = Spielberg et al.(1980)의 시험불안척도(Test Anxiety Inventory); TA2 = 황경렬(1997)의 시험불안 척도(K-TAI-K); SC1 = 김지혜(1991)의 자기의식 척도; SC2 = 김득란(1993)의 자기의식 검사; 학업동기척도 = 김아영(2002)의 학업동기 척도; 자기개념척도 = 이훈진(1997)의 자기개념 척도

<표 2> 실험 2의 설문조사 유형에 따른 문항의 구성

A형 설문조사의 구성 (문항의 개수)	B형 설문조사의 구성 (문항의 개수)
TA1(20)	TA2(20)
SI1(48)	SI2(48)
HS1(20)	HS2(20)
개방형질문 1(10)	개방형질문 2(10)
분기질문 1(3)	분기질문 2(3)
SE(26)	SD(48)
FT(18)	N(12)
CG(25)	CC(30)
PO(30)	C(48)
C(48)	PO(30)
CC(30)	CG(25)
N(12)	FT(18)
SD(48)	SE(26)
개방형질문 2(10)	개방형질문 1(10)
분기질문 2(3)	분기질문 1(3)
HS2(20)	HS1(20)
SI2(48)	SI1(48)
TA2(20)	TA1(20)

주. TA1 = Spielberg et al.(1980)의 시험불안척도(Test Anxiety Inventory); TA2 = 황경렬(1997)의 시험불안 척도(K-TAI-K); SI1 = 서봉연(1975)의 자아정체감 척도; SI2 = 박아청(1996)의 자아정체감 척도; HS1 = Beck et al.(1974)의 무망감 척도; HS2 = 이영호(1993)의 무망감 척도; SE = 김아영(2002)의 학업동기 척도 중 자기효능감 척도; FT = 김아영(2002)의 학업동기 척도 중 실패내성 척도; SD = 승인욕구 척도(금명자 1984); CG = 인지욕구 척도(김완석 1994); N = 부정적 평가에 대한 두려움 척도-단축형(최정훈·이정윤 1994); PO = 사회적 문제해결능력 척도(김화자 1998); CC = 이훈진(1997)의 자기개념 척도; C = NEO-PI-R 인성검사 중 성실성 척도(전미현 2013)

3) 절차

실험은 개인별로 온라인 설문지 시스템에 의하여 행하여졌다. 피험자의 모

집 과정에서 연구의 목적을 “대학생활 문화조사”라고 지시문을 주었으며, 설문조사가 모두 끝난 후 연구의 목적을 알려주었다.

2. 연구결과

실험 1의 총 설문 참가자 수는 387명이었다. 이 중 유급 참가자는 277명, 무급 참가자는 110명이었다. 206명이 A형 설문지에 응답했으며, 181명이 B형 설문지에 응답했다. 실험 1에서 설문 소요시간의 중앙값은 21분이었다. 실험 2의 총 설문 참가자 수는 223명이었다. 이 중 유급 참가자는 167명, 무급 참가자는 56명이었다. 119명이 A형 설문지에 응답했으며, 104명이 B형 설문지에 응답했다. 설문 소요 시간의 중앙값은 49분이었다. 실험 1과 실험 2 모두에서 유급 참가자와 무급 참가자의 반응의 질에 있어서 그 차이가 크지 않았기 때문에 이 둘을 합쳐서 분석하였다. 실험 1과 실험 2의 설문지 유형별 참가자의 성별은 <표 3>과 같다. 두 실험에서 설문지 유형별 참가자의 성의 비율의 차이는 통계적으로 유의미하지 않았다.

<표 3> 실험 1, 2에서 설문지 유형별 참가자의 성

	실험 1			실험 2	
	A형(비율)	B형		A형(비율)	B형
남자	62(0.30)	61(0.33)	남자	62(0.52)	46(0.44)
여자	144(0.70)	120(0.67)	여자	57(0.48)	58(0.56)

1) 문항의 위치에 따른 정확성의 차이

개인별 평균

<표 4>는 동일한 척도가 전반부에 제시된 경우와 후반부에 제시된 경우에 개인별 평균이 어떻게 다른지를 보여준다. <표 4>와 이후의 표에서 척도의 제시순서는 A와 B의 설문지에서 제시된 위치가 차이가 많이 날수록 먼저 나타나도록 배열되었다. 예를 들어 실험 1에서 TA1은 A유형의 설문지에서

는 가장 앞 쪽에 위치해 있었지만 B 유형의 설문지에서는 가장 뒤 쪽에 위치해 있어 위치의 차이가 가장 큰 척도이므로 표에서 가장 먼저 배치되었다. SI1의 경우는 A유형에서는 두 번째 위치에 B유형에서는 끝에서 두 번째 위치에 배열되었으므로 TA1과 TA2 다음에 배치되었다. 실험 1에서 설문지의 중간에 위치한 학업동기 척도와 자기개념 척도는 설문지 유형에 따른 위치의 차이가 없었기 때문에 분석에서 제외하였다. 실험 2의 척도에 대한 배치도 동일한 방법을 사용하였다.

<표 4> 척도의 위치에 따른 개인별 평균

척도	전/후반부 척도 사이의 문항 개수	전반부 평균 (표준편차)	후반부 평균 (표준편차)	p-값	FDR p-값
실험 1					
TA1	L:128, 개:8, 분:6	2.79(0.72)	2.57(0.76)	0.005	0.021*
TA2	L:128, 개:8, 분:6	3.00(0.68)	3.09(0.75)	0.217	0.290
SC1	L:91, 개:8, 분:6	3.38(0.48)	3.39(0.46)	0.808	0.808
SC2	L:91, 개:8, 분:6	3.56(0.47)	3.49(0.48)	0.119	0.237
실험 2					
TA1	L:393, 개:20, 분:6	2.62(0.66)	2.59(0.78)	0.742	0.938
TA2	L:393, 개:20, 분:6	2.86(0.69)	2.82(0.74)	0.655	0.917
SI1	L:325, 개:20, 분:6	2.71(0.26)	2.75(0.27)	0.377	0.881
SI2	L:325, 개:20, 분:6	2.75(0.28)	2.79(0.37)	0.362	0.881
HS1	L:257, 개:20, 분:6	2.78(0.27)	2.83(0.27)	0.234	0.819
HS2	L:257, 개:20, 분:6	1.99(0.76)	2.14(0.78)	0.155	0.722
SE	L:211	3.12(0.32)	3.04(0.35)	0.078	0.549
SD	L:189	3.01(0.35)	3.04(0.36)	0.644	0.917
FT	L:167	3.08(0.50)	2.94(0.50)	0.033	0.462
CG	L:124	3.07(0.33)	3.06(0.33)	0.871	0.938
N	L:117	2.90(0.49)	2.90(0.45)	0.953	0.953
CC	L:87	3.03(0.23)	3.01(0.22)	0.463	0.913
PO	L:69	2.89(0.50)	2.88(0.47)	0.853	0.938
C	L:9	3.04(0.47)	3.00(0.42)	0.522	0.913

주. FDR = False Discovery Rate, L = 리커트 5점 척도 문항, 개 = 개방형 문항, 분 = 분기질문 문항.
* p<.05.

차이 검증 시 비교의 개수 증가에 따라서 1종 오류율이 급격히 증가하는 것을 보정하기 위해, Benjamini & Hochberg(1995)의 거짓 발견율(False Discovery Rate)을 사용하였다. 거짓발견율을 .05로 통제한다는 것은 기각된 영가설 중 잘못 기각된 영가설의 비율이 평균적으로 5%라는 의미이다. 분석의 결과, 개인별 평균은 척도의 위치에 따라 차이를 보이지 않았다. 유일하게 실험 1의 TA1에서 유의미한 차이를 발견할 수 있었다. 하지만 동일한 구성체를 측정하는 TA2에서는 유의미한 차이를 발견할 수 없었고, 실험 2에서도 TA1, TA2 모두 유의한 차이를 보이지 않았다.

개인별 표준편차

<표 5>는 각 척도에서 보인 개인 응답의 표준편차가 척도의 설문지 내의 위치에 따라 어떻게 다른지를 보여준다. 모든 척도에 대해 등분산을 가정하지 않는 t-검정(Welch 1938)을 실시하였으며, FDR p-값이 제시되어 있다. 실험 1에서는 전반부에 제시된 척도들에 비해 후반부의 척도들이 개인별 표준편차의 평균이 작아졌다. 실험 2에서도 위치 차이가 증가함에 따라 효과크기가 같이 증가하는 패턴을 보였다.

항상 양인 표준편차의 분포는 정규분포를 따르지 않기 때문에 표준편차의 평균 차이에 대한 웰치의 t-검정은 제한적일 수 있다. 그렇지만 실제 자료는 거의 모두 정규성을 어느 정도 어긋나게 마련이고(Micceri 1989), 정규성이 어느 정도 어긋난 자료의 경우에도 웰치의 t-검정이 1종 오류율을 정확하게 통제할 수 있다(Fagerland & Sandvik 2009; Kang et al. 2014). 따라서 중요한 것은 주어진 분포가 정규성에서 얼마나 벗어났는가 하는 것이다. 보통 왜도(skewness)의 절대값이 1보다 작으면 정규분포에서 크게 벗어나지 않는 것으로 간주한다(Osborne 2008). 이후에 제시될 모든 분석들에서 정규성 검정, 분위수-분위수 도표(quantile-quantile plot), 왜도값을 통해 정규분포에서 어긋난 정도를 확인하고, 정규 분포에 크게 어긋나지 않는 경우에만 웰치의 t-검정을 사용하였다.

<표 5> 척도의 위치에 따른 개인별 표준편차

척도	전반부 평균 (표준편차)	후반부 평균 (표준편차)	효과크기	p-값	FDR p-값
실험 1					
TA1	0.97(0.24)	0.87(0.31)	0.34	0.001	0.002**
TA2	1.04(0.27)	0.87(0.33)	0.56	0.000	0.000**
실험 2					
TA1	0.98(0.25)	0.86(0.32)	0.40	0.005	0.033*
TA2	1.03(0.29)	0.96(0.31)	0.24	0.082	0.335
SI1	1.21(0.26)	1.15(0.33)	0.21	0.127	0.356
SI2	1.11(0.26)	1.07(0.30)	0.15	0.279	0.424
HS1	1.29(0.37)	1.23(0.44)	0.15	0.261	0.424
HS2	0.66(0.35)	0.67(0.38)	0.02	0.870	0.937

주. FDR = False Discovery Rate. * p<.05. ** p<.01.

개인별 엔트로피

<표 6>은 척도의 위치에 따른 개인별 엔트로피의 평균의 변화를 보여준다. 개인별 엔트로피의 평균의 차이는 척도의 위치 차이가 클수록 증가하는 패턴을 보였다.

<표 6> 척도의 위치에 따른 개인별 엔트로피

척도	전반부 평균 (표준편차)	후반부 평균 (표준편차)	효과크기	p-값	FDR p-값
실험 1					
TA1	1.16(0.24)	0.98(0.35)	0.60	0.000	0.000**
TA2	1.21(0.25)	1.03(0.34)	0.60	0.000	0.000**
실험 2					
TA1	1.14(0.26)	0.89(0.39)	0.76	0.000	0.000**
TA2	1.21(0.27)	1.04(0.39)	0.49	0.000	0.001**
SI1	1.40(0.20)	1.22(0.35)	0.62	0.000	0.000**
SI2	1.32(0.19)	1.19(0.35)	0.46	0.000	0.001**
HS1	1.25(0.25)	1.10(0.34)	0.51	0.000	0.001**
HS2	0.78(0.41)	0.74(0.46)	0.09	0.490	0.686

주. FDR = False Discovery Rate. * p<.05. ** p<.01.

개인별 중간 반응의 비율

실험 1과 실험 2에 쓰인 모든 척도는 5점 척도의 리커트 응답 문항으로 이루어졌다. 이때 중간 반응은 “보통이다”를 의미한다. 개인별 중간 반응의 비율은 척도의 위치에 따라 크게 다르지 않았다. 중간의 세 반응(“다소 그렇지 않다”, “보통이다”, “약간 그렇다”)을 모두 합친 반응의 비율 역시 비슷한 결과를 보였다. 피험자들이 불성실한 반응을 할 때 단순히 “보통이다”에 반응하는 것은 아님을 시사해 주는 결과이다.

개인별 최대 한줄로 응답하기 길이와 분기질문 응답

개인별 최대 한줄 응답 길이는 무응답을 처리하는 방법에 따라 세 가지 방식으로 계산할 수 있다. 첫 번째 방법은 무응답이 하나라도 존재하는 응답자의 응답을 제외하는 것이다. 이 방법은 표본의 크기가 줄어든다는 단점이 있다. 두 번째 방법은 무응답을 제외하고 최대 한줄 응답 길이를 구하는 것이다. 예를 들면 응답이 (무응답, 무응답, 2, 2 ,2)인 경우 최대 한줄 응답 길이는 3이다. 이 방법은 최대 한줄 응답의 길이의 최대값이 응답자마다 다르다는 단점이 있다. 예를 들어 다섯 문항에 대한 응답 중 무응답이 2개 존재한다면, 최대 한줄 응답 길이의 최대값은 5가 아니라 3이 된다. 세 번째 방법은 무응답도 하나의 응답처럼 취급하여 최대 한줄 응답 길이 계산에 포함시키는 것이다. (무응답, 무응답, 무응답, 4, 4)의 응답의 경우 무응답이 세 번 연속되어 나왔으므로 최대 한줄 응답 길이는 3이 된다. 여기서는 세 번째 방법을 사용하여 분석하였다.

실험 결과 개인별 최대 한줄 응답 길이는 정규분포에서 많이 벗어나 있었다(실험 1에서 왜도의 평균 3.068, 첨도의 평균 13.386; 실험 2에서 왜도의 평균 2.86, 첨도의 평균 15.506). 따라서 음이항 분포를 가정하고 독립변수로 설문의 위치(전반부/후반부)를 사용한 일반화 선형 모형을 통해 설문의 위치와 개인 최대 한줄 응답 길이의 관계를 추정하였다(표 7).

<표 7> 척도의 위치에 따른 개인별 최대 한줄 응답 길이

척도	전반부 평균 (표준편차)	후반부 평균 (표준편차)	추정계수 (표준편차)	p-값	FDR p-값
실험 1					
TA1	4.45(2.63)	6.04(3.94)	0.31(0.06)	0.000	0.000**
TA2	3.76(2.36)	5.05(3.99)	0.29(0.06)	0.000	0.000**
실험 2					
TA1	4.62(2.95)	7.14(4.63)	0.44(0.08)	0.000	0.000**
TA2	4.13(2.93)	5.63(4.70)	0.31(0.09)	0.002	0.002**
SI1	4.45(4.40)	7.44(9.82)	0.51(0.10)	0.000	0.000**
SI2	4.03(1.98)	6.50(9.30)	0.48(0.10)	0.000	0.000**
HS1	2.93(1.91)	4.08(4.05)	0.33(0.09)	0.000	0.000**
HS2	7.89(5.68)	8.13(6.18)	0.03(0.09)	0.755	0.755

주. FDR = False Discovery Rate. * p<.05. ** p<.01.

<표 7>에서 추정 계수는 설문지가 후반부에 위치했을 때 최대 한줄 응답 길이가 늘어나는 정도를 log-척도상에서 제시한 것이다(예, $0.31 \simeq \log 6.04 - \log 4.45$). 위치의 차이가 가장 작은 HS2를 제외하고는 설문지의 후반부가 전반부보다 최대 한줄 응답 길이가 통계적으로 유의하게 높았다.

분기질문에 대한 “아니오” 응답은 위치에 따라 큰 차이 없는 것으로 나타났다.

개인별 렘펠-지브 복잡도

<표 8>은 척도의 위치에 따른 개인별 렘펠-지브 복잡도의 차이를 보여준다. 후반부의 복잡도가 전반부보다 감소하였으며, 위치 차이가 증가할수록 효과크기가 일관적으로 증가하였다.

<표 8> 척도의 위치에 따른 개인별 렘펠-지브 복잡도

척도	전반부 평균 (표준편차)	후반부 평균 (표준편차)	효과크기	p-값	FDR p-값
실험 1					
TA1	0.86(0.09)	0.80(0.13)	0.59	0.000	0.000**
TA2	0.88(0.09)	0.82(0.13)	0.55	0.000	0.000**
실험 2					
TA1	0.85(0.10)	0.77(0.14)	0.67	0.000	0.000**
TA2	0.87(0.10)	0.82(0.14)	0.45	0.001	0.003**
SI1	0.98(0.08)	0.91(0.15)	0.59	0.000	0.000**
SI2	0.95(0.07)	0.91(0.15)	0.36	0.005	0.016*
HS1	0.87(0.10)	0.83(0.13)	0.38	0.006	0.016*
HS2	0.73(0.15)	0.71(0.17)	0.11	0.420	0.588

주. FDR = False Discovery Rate. * p<.05. ** p<.01.

문항별 평균, 표준편차, 엔트로피

문항별로 평균, 표준편차, 엔트로피의 평균의 차이를 검증한 결과 개인별 응답 분포와는 달리, 일부 문항에서 통계적으로 유의미한 차이가 있었으나 그 차이가 극히 일부 문항에 국한되었고, 차이의 방향도 일관적이지 않았다.

2) 문항의 위치에 따른 일관성의 차이

검사 점수의 신뢰도는 예상과 다르게 후반부에서 점수의 신뢰도가 다소 높아지는 경향을 보였다. 예를 들어 실험 1에서 TA2의 신뢰도는 전반부에서 0.90, 후반부에서 0.94로 0.04(p=0.003) 증가하였으며, 실험 2에서도 TA2의 신뢰도는 0.90에서 0.93으로 0.03(p=0.08) 증가하였다. 대체적으로 위치의 차이가 클수록 신뢰도의 증가폭이 증가하는 경향을 보였다. IRT 기반의 반응자 우도는 척도의 위치 차이에 따라 일관적인 결과가 나오지 않았다. 이런 결과는 라쉬 모형과 2-모수 모형에서 거의 동일하였다. MRF 모형 기반의 개인 우도 역시 척도의 위치에 따른 차이가 거의 없었다. 이는 문항의 위치효과가 이 논문에서 사용된 일관성 척도에 의하면 차이가 나지 않음을 나타낸다.

3) 문항의 위치에 따른 성실성 차이

개인별 결측치 비율

개인별 결측치 개수의 분포는 정규분포에서 상당히 벗어났다. 개인별 결측치 개수의 분포는 0이 다수이며, 그 다음에 1, 2이고 극단적으로 높은 값들이 소수 보이는 분포였다. 이 분포는 전통적으로 가산 자료의 분포로 제시된 이항분포, 음이항분포, 포와송 분포와도 상당히 다른 분포이다. 여기에서는 0의 개수가 월등히 많은 가산 자료를 분석하는 데 적합한 영과잉 가산모형(Zero-inflated count models)이 사용되었다. 영과잉 가산모형은 두 분포의 혼합분포이다. 첫번째 분포는 결측치가 하나도 없거나 하나 이상인 경우를 확률적으로 결정하고, 두번째 분포는 결측치가 하나 이상일 경우 그 개수를 포와송 분포로 결정한다(Mullahy(1986)와 Lambert(1992)를 참조).

분석결과 결측치가 하나도 없는 응답의 비율에는 유의미한 차이가 없었지만, 결측치가 하나 이상일 때, 결측치의 개수는 전반부에 비해 후반부에서 더 많았다(<표 9>).

<표 9> 척도의 위치에 따른 개인별 결측치

척도	전반부 완결한 응답비율	후반부 완결한 응답비율	p-값	전반부 결측치 평균	후반부 결측치 평균	p-값
실험 1						
TA1	0.93	0.91	0.439	1.36	2.50	0.008**
TA2	0.89	0.92	0.996	1.30	1.19	0.521
실험 2						
TA1	0.93	0.89	0.369	1.12	5.75	0.001**
TA2	0.95	0.85	0.211	1.60	6.72	0.001**
SI1	0.89	0.83	0.273	1.15	9.22	0.000**
SI2	0.89	0.84	0.849	1.55	13.89	0.000**
HS1	0.97	0.92	0.867	1.25	8.12	0.003**
HS2	0.96	0.85	0.949	1.00	6.33	0.000**

주. 완결한 응답비율 = 결측치가 없는 응답의 비율; 결측치 평균 = 결측치가 하나 이상인 응답에 대해 결측치의 갯수의 평균. ** $p<.01$.

개방형 질문에 대한 응답 길이

<표 10>은 위치에 따른 개방형 질문에 대한 응답의 길이(문자 수)를 보여준다. 후반부의 질문에 대한 응답이 일관적으로 짧아지는 결과를 볼 수 있다. 응답의 길이를 단어 수나 문장 수로 분석해도 동일한 결과를 얻을 수 있었다. 하지만 문장당 단어 수는 차이가 나지 않았다.

<표 10> 개방형 질문의 위치에 따른 길이

문항번호	전반부	후반부	효과크기	p-값	FDR p-값
실험 1					
O1	64.19(60.46)	44.58(44.20)	0.37	0.000	0.000**
O2	52.81(39.24)	44.07(39.27)	0.22	0.029	0.035**
O3	67.76(47.34)	52.71(34.17)	0.36	0.000	0.000**
O4	74.05(51.25)	61.57(43.09)	0.26	0.010	0.027**
O5	53.46(32.39)	46.20(29.67)	0.23	0.023	0.035**
O6	56.11(46.60)	47.06(33.59)	0.23	0.031	0.035**
O7	50.61(32.93)	42.81(32.61)	0.24	0.020	0.035**
O8	70.28(58.03)	60.10(40.64)	0.21	0.049	0.049**
실험 2					
O1	71.71(69.16)	51.29(80.93)	0.27	0.046	0.052†
O2	58.16(51.49)	39.13(33.94)	0.43	0.001	0.003**
O3	62.30(49.09)	47.15(34.19)	0.35	0.008	0.013*
O4	69.78(48.16)	55.63(46.76)	0.30	0.027	0.036*
O5	58.73(70.80)	43.71(38.46)	0.27	0.056	0.056†
O6	66.53(46.67)	48.97(37.49)	0.42	0.002	0.004**
O7	56.13(42.66)	37.98(24.60)	0.53	0.000	0.000**
O8	55.04(44.12)	37.95(23.62)	0.49	0.001	0.003**

주. FDR = False Discovery Rate.

† p<.10. * p<.05. ** p<.01

3. 논 의

<표 11>은 실험 1, 2의 결과를 종합적으로 보여준다. 응답의 질의 세 측면(정확성, 일관성, 성실성) 중 정확성과 성실성에서만 설문지의 길이 효과가 발견되었다. 다양한 통계치에서 전반부의 응답의 분포와 후반부의 응답의 분포가 차이가 났다. 그러나 문항 수준의 통계치들에서는 일관적인 차이를 발견할 수 없었고, 개인 수준의 통계치(표준편차, 엔트로피, 최대 한줄 응답 길이, 렘펠-지브 복잡도)에서만 일관적인 차이를 관찰하였다. 설문의 위치의 차이에 비례하여 효과 크기가 커지는 설문 길이의 효과를 발견하였다. 이러한 결과는 응답의 질을 평가하는 지수로서 문항별 통계치는 효율적이지 않음을 보여준다.

<표 11> 설문의 위치에 따른 여러 통계치/지수의 차이

응답의 질	세부 특징	통계치/지수	전반부 대비 후반부에서 응답
정확성	순서 무관	개인별 평균	(통계적으로 유의미한) 차이 없음
		개인별 표준편차	작아짐
		개인별 엔트로피	작아짐
		개인별 중간 반응의 비율	차이 없음
		개인별 분기질문에 대한 응답	차이 없음
		문항별 평균, 표준편차, 엔트로피	일관적인 차이 없음
	순서 유관	최대 한줄 응답 길이	길어짐
		렘펠-지브 복잡도	작아짐
일관성	집단 일관성	고전 검사 이론의 신뢰도	차이 없음
	집단내 개인 일관성	IRT 개인 우도	차이 없음
		MRF 개인 우도	차이 없음
성실성	선다형 문항	결측치의 비율	많아짐
	개방형 문항	개방형 응답의 길이	작아짐

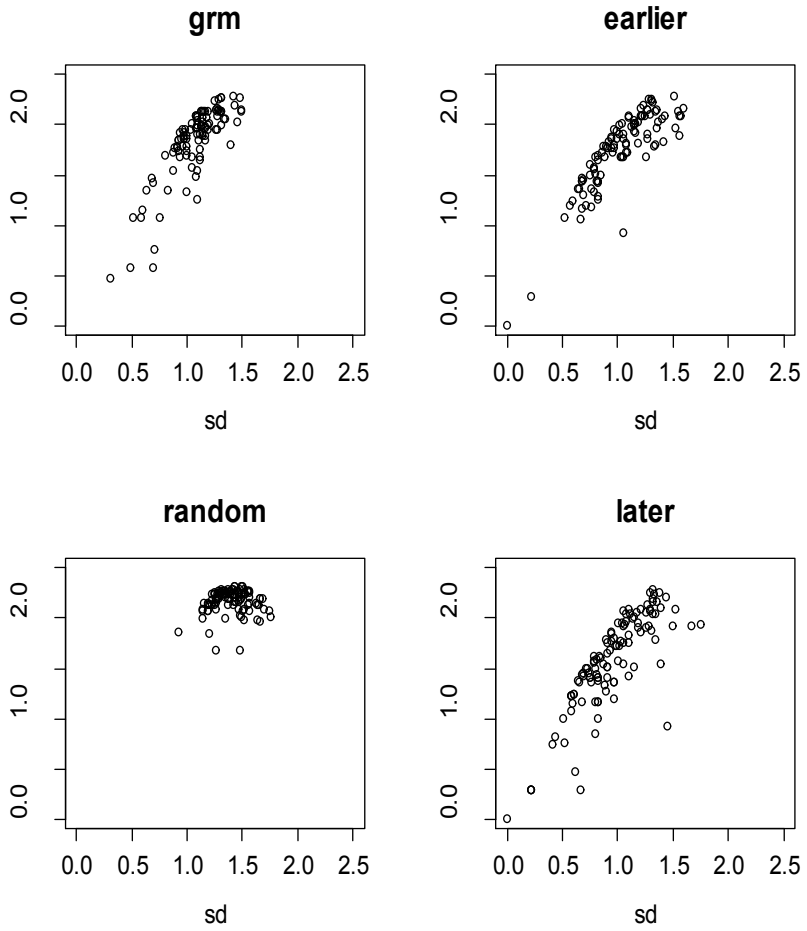
응답의 일관성을 반영하는 신뢰도, IRT 개인 우도, MRF 개인 우도는 설문지의 위치에 따른 체계적인 차이를 보이지 않았다. IRT 개인 우도는 일관적이진 않지만 다소 증가하는 경향이 있었다. 성실성을 반영하는 결측치의 개수와 개방형 응답의 길이를 살펴보면, 결측치의 개수는 후반부에서 증가하고, 개방형 응답의 길이는 후반부에서 감소했다. 이 결과는 후반부에서 성실성이 감소함을 시사한다.

이 결과를 종합해 보면, 설문의 길이가 길어짐에 따라 응답의 정확성이 낮아지고 성실성은 감소하지만, 일관성은 큰 차이를 보이지 않는다. 일관성에서 위치에 따른 차이가 발견되지 않은 것은 예상되지 않았던 결과였다. 이에 대한 한 가지 설명은 실험 설계상에서 일관성을 측정하기 위하여 유사한 질문들로 구성된 척도를 사용하였기 때문일 수 있다. 응답자들이 문항을 읽는다면 유사한 문항들이 계속하여 제시되고 있다는 것을 지각할 수 있을 것이고, 이때 주의를 기울여서 반응할 가능성이 있다. 이러한 경우에는 일관성이 떨어지지 않을 것이다. 이에 대한 추후의 연구가 필요할 것으로 보인다.

표준편차의 감소, 엔트로피의 감소, 최대 한줄 응답 길이의 증가, 렘펠-지브 복잡도의 감소가 응답의 질의 저하를 나타내는 지수로 해석하는 데에는 다소 주의가 필요하다. 만약 모든 응답을 무작위로 생성하면 이 응답의 엔트로피와 표준편차는 정상적인 응답들보다 크다. 하지만 이 응답의 질이 좋다고 말할 수 없다. <그림 1>은 시뮬레이션과 실제 반응에 대한 표준편차와 엔트로피 그래프를 보여준다. 첫 번째(위/왼쪽) 그래프는 실험 2의 TA2에 대해서 전반부의 응답을 등급반응모형(graded response model, Samejima 1969)을 기반으로 모수를 추정한 후, 다시 추정된 모형을 기반으로 반응을 생성했을 때 개인별 표준편차와 엔트로피 그래프이고, 두 번째(아래/왼쪽) 그래프는 무작위로 반응을 생성했을 때 개인별 표준편차와 엔트로피 그래프이다. 세 번째, 네 번째 그래프는 실제 실험 2의 전반부 TA2, 후반부 TA2 응답에서 구한 개인별 표준편차와 엔트로피의 그래프이다.

여기에서 무작위로 생성된 자료는 엔트로피의 값은 높지만, 모형에 기반하여 생성된 자료나 실제자료와는 다른 패턴을 가지고 있는 것을 발견할 수 있다. 이는 표준편차와 엔트로피가 응답의 질을 판단하는 개별 지수로 사용하

는 데에는 한계가 있으며, 다른 지수들과 함께 사용될 때 보다 정확하게 사용될 수 있음을 시사한다.



〈그림 1〉 응답의 표준편차/엔트로피 그래프

왼쪽/위 = 등급반응모형에서 생성한 반응,
왼쪽/아래 = 무작위 생성 반응,
오른쪽/위 = 실험 2 전반부 TA2의 실제 응답,
오른쪽/아래 = 실험 2 후반부 TA2의 실제 응답

마지막으로 후반부에서 검사 점수의 신뢰도가 하락하지 않은 것은 검사의 타당도 지수를 해석하는 데 시사점을 제공한다. 심리검사를 사용하는 문헌에서는 높은 신뢰도를 확보하는 것을 필수적으로 여긴다. 검사의 타당도가 가

장 중요하기는 하지만 신뢰도는 타당도의 상한을 결정하기 때문에 높은 신뢰도를 유지하는 것도 중요하다. 그렇지만 본 연구의 결과는 너무 높은 신뢰도는 내적인 일관성은 있지만 응답자의 심리적 특성을 충실히 반영하지 못할 가능성이 있기 때문에, 오히려 타당도를 저하시킬 수도 있음을 보여준다 (Loevinger 1954).

Ⅳ. 연구 2: 실제 설문 자료에서 나타난 설문 길이에 따른 응답의 질의 차이

연구 1에서는 실험설계에 의하여 동일한 문항을 설문지의 앞부분이나 뒷부분에 위치시켰을 때의 반응의 질이 어떻게 변화하는지를 실험법을 사용하여 조사하였다. 연구 2에서는 연구 1에서 사용된 방법들이 실제 조사 자료에 적용할 수 있는지를 예증하고자 하였다. 특히 실제 자료는 실험실에서 생성된 자료와는 달리 여러 가지 분석의 제약이 존재하게 되는데 이때 어떻게 분석할 수 있는지를 시연하고자 한다.

1. 연구방법

1) 분석자료

한국 리서치에서 2013년 웹을 통해 실시한 두 개의 설문조사 자료가 분석되었다. 첫 번째 자료는 기업 이미지에 대한 조사였고, 두 번째 자료는 상품에 대한 이용실태에 대한 조사였다. 첫 번째 자료는 설문소요시간의 중앙값이 21분이었고, 참가 인원은 1,300명이었다. 두 번째 자료는 설문소요시간의 중앙값이 51분이었고, 참가 인원은 5,000명이었다. 이들 자료에서 리커트 척도의 응답만을 선별하여 분석하였다. 첫 번째 자료에는 5점 척도의 리커트 척도 문항이 268개 있었고, 두 번째 자료에는 7점 척도의 리커트 척도 문항이 249개 있었다. 두 자료 모두 결측치는 없었다. 두 번째 자료는 분기질문이 있었다(예를 들어 “pda를 사용하십니까?”란 질문에 예라고 답한 사람들에게

만 해당하는 질문이 있었다). 첫 번째 자료는 “자료 1”, 두 번째 자료는 “자료 2”로 부르겠다. 개별반응에서 결측치의 유무에 따라 문항의 위치가 다르게 정의될 수 있는데, 이 분석에서는 설문지에서 몇 번째로 제시되었는지를 문항의 위치로 삼았다.

2) 분석방법

이 연구에서 사용한 자료는 문항의 내용을 통제한 실험적 자료가 아니기 때문에 여러 통계치의 변화가 문항의 위치에 의한 것인지, 아니면 문항의 내용에 의한 것인지 구분하기 힘들다. 이러한 상황에서 문항의 위치에 대한 효과를 분석하기 위하여 다음의 두 가지 방식을 사용하였다.

첫 번째 분석은 자료에서 일정한 크기의 이동창(moving window)으로 문항을 선택하여 개인별 평균, 개인별 표준편차, 개인별 엔트로피, 개인별 중간반응 비율, 최대 한줄 응답 길이, 렘펠-지브 복잡도를 구하고 그래프로 그려서 설문이 진행됨에 따라 각 통계치가 어떻게 변하는지 검토하였다.

두 번째 분석은 전체 응답을 부분으로 나눈 후 변화시점탐지기법을 활용하여 각 통계치의 변화시점과 변화된 값을 구하였다. 변화시점(change point)을 탐지하는 문제는 잘 연구된 주제이며, 생물정보학, 금융, 기상학, 의학 등에서 응용되고 있다(Johnson et al. 2013). 변화시점 탐지는 크게 모수적 변화시점 탐지와 비모수적 변화시점 탐지로 구분된다. 모수적 변화시점 탐지는 관측치가 특정한 모수를 가지는 동일한 분포에서 독립적으로 추출된다는 가정을 가지고, 모수가 변화되는 시점을 추정한다. 여기서 모수는 다시 특정한 사전 분포를 따른다는 가정을 한다. 이런 가정 하에서는 우도비 검정(likelihood ratio test)을 할 수 있다는 장점이 있다(Killick et al. 2012; Johnson et al. 2013에서 재인용). 이때 변화의 시점은 하나일 수도 있고, 다수일 수도 있다. 변화의 시점이 여럿일 때 효율적인 추정방법은 Killick et al.(2012)에 의해 제안되었다.

본 연구에서는 Killick et al.(2013)이 구현한 변화시점 탐지 기법을 활용하여 변화시점이 하나 혹은 여럿일 때 관찰값의 분포가 정규분포를 따른다는 가정하에서 평균의 변화시점을 분석하였다. 여기서 관찰값은 여러 가지 통계

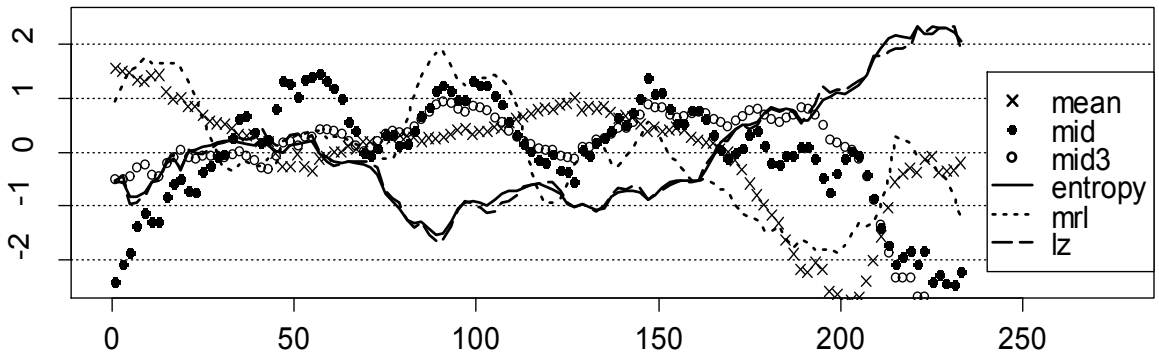
치(개인별 평균, 개인별 표준편차, 개인별 엔트로피, 개인별 중간반응, 최대 한줄 응답 길이, 렘펠-지브 복잡도)의 평균이다. 변화시점 탐지 기법이 관찰값이 모두 독립이라고 가정하기 때문에 구간의 크기를 여러 통계치에서 자기 상관이 유의미하지 않게 조정하였다. 보통 구간의 크기가 10 이상이면 여러 통계치의 자기 상관이 통계적으로 유의미하지 않았다.

2. 연구결과

<그림 2>와 <그림 3은> 자료 1과 자료 2에 대해 이동창(크기: 35)의 시작 위치에 따른 평균, 중간반응의 비율, 중간 세반응의 비율, 개인별 엔트로피, 개인별 최대 한줄응답 길이, 렘펠-지브 복잡도의 변화를 보여준다. <그림 2>를 살펴보면 엔트로피와 렘펠지브 복잡도가 비슷한 형태로 움직이고, 중간반응과 중간 세반응의 비율이 비슷한 움직임을 보이고, 최대 한줄 응답 길이와 평균은 다소 독립적으로 움직이고 있음을 볼 수 있다. 연구 1과는 달리 평균과 중간 반응의 비율이 변하고 있으므로 문항의 특성이 변하고 있다고 추측할 수 있다.

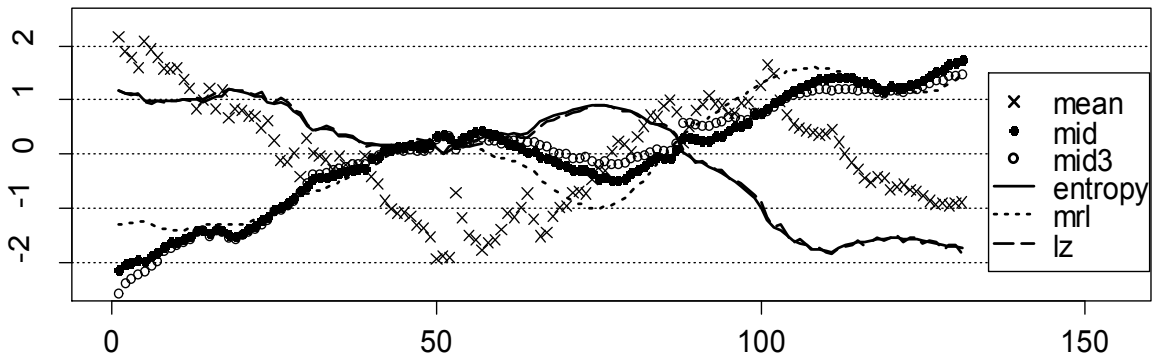
<그림 3>의 경우는 한줄 응답 길이, 중간 반응의 비율, 중간 세 반응의 비율이 한 집단을 이루고, 개인별 엔트로피와 렘펠-지브 복잡도가 다른 집단을 이루고 있으며, 평균은 이들과 독립적으로 움직이고 있음을 확인할 수 있다. 역시 평균과 중간 반응의 비율이 변하고 있으므로 문항의 특성이 달라지고 있다고 추측할 수 있다.

전체 응답을 블록으로 나눈 후 부분에 대해 개인별 평균, 개인별 표준편차, 개인별 중간 반응의 비율, 개인별 중간 세 반응의 비율, 개인별 최대 한줄 응답 길이, 개인별 엔트로피, 그리고 개인별 렘펠-지브 복잡도의 평균을 구하여 변화시점 탐지기법을 적용한 결과 자료 2의 개인별 최대 한줄 응답 길이의 평균에 대해서만 변화시점이 탐지되었다. 변화시점 이전의 블록에서 평균은 4.23, 변화시점 이후의 블록에서 평균은 6.84였다(그림 4).



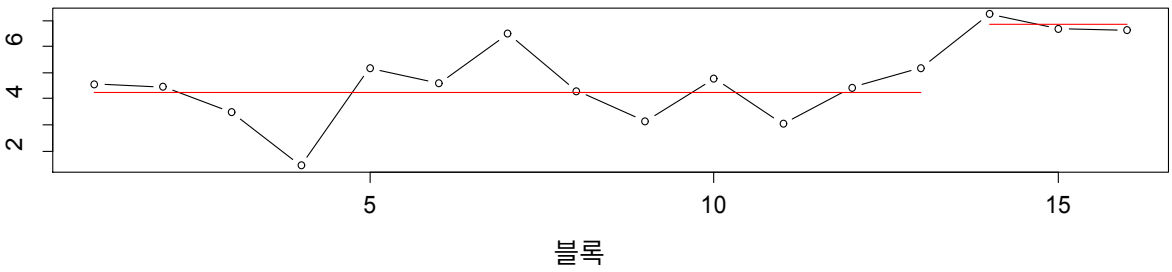
〈그림 2〉 〈자료 1〉에서 이동창 시작 위치에 따른 이동창(크기:35) 표준화값의 변화

mean=평균, mid=중간반응의 비율, mid3=중간 세 반응의 비율,
entropy=개인별 엔트로피, mrl=개인별 최대 한줄 응답 길이,
lz=렘펠-지브 복잡도



〈그림 3〉 〈자료 2〉에서 이동창 시작 위치에 따른 이동창(크기:35) 표준화값의 변화

mean=평균, mid=중간반응의 비율, mid3=중간 세 반응의 비율,
entropy=개인별 엔트로피, mrl=개인별 최대 한줄 응답 길이,
lz=렘펠-지브 복잡도



〈그림 4〉 자료 2에서 블록(크기:15)에 대한 최대 한줄 응답 길이의 평균

3. 논 의

연구 2에서는 연구 1의 통계치들을 실제 자료에 적용하였다. 이동창을 사용하여 엔트로피, 최대 한줄 응답 길이, 렘펠-지브 복잡도를 구한 결과, 짧은 설문에서는 후반부에서 엔트로피와 렘펠-지브 복잡도가 증가하고, 최대 한줄 응답 길이는 짧아지는 현상을 관찰하였다. 반면 긴 설문에서는 엔트로피와 렘펠-지브 복잡도가 감소하고, 최대 한줄 응답 길이는 길어지는 현상을 관찰하였다. 이러한 결과는 21분 정도 소요되는 짧은 설문지에서는 설문지의 뒤에 있는 문항이 앞에 있는 문항에 비하여 반응의 질이 떨어지지 않고 오히려 좋아지는 경향이 있는 반면, 51분 정도 소요되는 긴 설문지에서는 반응의 질이 떨어지는 경향이 있음을 의미한다. 따라서 21분 정도의 설문지는 반응의 질을 떨어뜨릴 만큼 긴 설문지는 아니며, 51분은 긴 설문지라고 판단할 수 있다. 또한 흥미 있는 결과는 설문지의 초기에 하는 반응의 질이 20분 지난 후에 하는 반응의 질보다 더 낮다는 것이다. 이는 초기에 설문에 응답하는 준비가 되지 않았기 때문일 가능성이 있다.

전체 응답을 부분으로 나눈 후 여러 가지 통계치에 대한 평균을 구해 변화시점 탐지기법을 적용한 결과에서는 자료 2의 후반부에서 최대 한줄 응답 길이의 변화가 탐지되었다. 두 분석은 실제 자료를 통해 여러 가지 통계치가

실제로 어떻게 변화하는지 살펴보고, 그것을 응답의 질과 관련하여 어떤 해석을 할 수 있는지를 보여주었다는 데 의의가 있다 .

Ⅴ. 전체 논의

이 논문에서는 다양한 통계치와 지수를 활용하여 질문지의 길이가 응답의 질에 미치는 영향을 탐구하였다. 이 연구는 다음과 같은 의의를 가진다. 첫째, 설문지의 길이가 반응의 질에 미치는 효과를 검증하였다. 선행연구의 방법을 활용하여 설문의 길이 또는 문항의 위치에 따른 응답의 질의 차이를 조사한 결과, 평균은 문항의 위치 차이에 따라 일관적인 차이를 보이지 않았으며, 표준편차는 문항 위치의 차이가 클수록 큰 차이를 보였고, 결측치의 비율은 후반부에 증가하는 패턴을 보였다. 개방형 질문의 경우 후반부에서 응답의 길이가 짧아졌다. 그리고 선행연구와 다르게 분기문항은 문항의 위치에 따라 큰 차이를 보이지 않았다. 이러한 결과를 통해 설문지의 길이가 증가함에 따라 응답의 정확성과 성실성이 낮아졌다고 결론을 내릴 수는 있었다.

둘째, 문항에 대한 반응의 질을 반영할 수 있는 새로운 통계치와 지수를 제안하였다. 새롭게 제안된 통계치와 지수는 개인별 엔트로피, 최대 한줄 응답 길이, 렘펠-지브 복잡도, 고전검사이론의 신뢰도, 문항반응이론 기반 반응자 우도, MRF 모형 기반 개인도였다. 앞에서 논의한 내용에 근거하여 설문의 길이가 응답의 질을 감소시킨다는 결론을 받아들인다면 개인별 엔트로피, 최대 한줄 응답 길이, 렘펠-지브 복잡도는 문항의 위치에 따른 응답의 질의 변화를 뚜렷하게 반영하였으나, 고전검사이론의 신뢰도, 문항반응이론 기반 반응자 우도, MRF 모형 기반 개인도는 그렇지 못했다. 이는 이러한 지수들이 반응의 질을 측정하기에 적합하지 않았든지 혹은 이 지수들이 측정하는 일관성이 변화되지 않았기 때문일 수 있다. 이에 대한 체계적인 연구가 필요할 것이다.

일부 지수들이 효과적이지 않은 결과는 Karabatsos(2003)의 결과와 비교해볼 수 있다. Karabatsos(2003)는 일탈 반응(aberrant response)을 탐지하기 위해 36개의 개인 적합도 통계치를 비교하였는데, 일탈 반응을 탐지하는 데 상대적으로 더 효과적이었던 통계치는 모형에 기반하지 않은 비모수 통계치였다. 물론 Karabatsos(2003)의 연구는 시험 상황에서의 일탈 반응(부정행위, 추측해서 맞추기, 무작위 반응 등)을 탐지하는 개인 적합도 통계치에 대한 연구였기 때문에 설문의 길이에 따른 응답의 질의 저하에 직접 적용할 수는 없지만, 반응의 질의 측면에 따라서 다른 통계치가 적용될 수 있음을 시사한다.

본 연구의 결과는 김형준 등(2014)과 장은경(2010)의 결과와도 비교해 볼 수 있다. 김형준 등(2014)과 장은경(2010)은 각각 MRF와 문항반응이론을 바탕으로 무작위적 반응 패턴들을 그렇지 않은 반응 패턴들로부터 구분할 수 있음을 보였다. 하지만 그들이 연구한 무작위 반응은 응답자가 질문을 읽지도 않고 설문지만 받은 상태에서 답을 한 반응이었다. 따라서 본 연구에서 질문의 내용을 읽을 수 있는 상황과는 다르다. 김형준 등(2014)과 장은경(2010)의 연구 결과에 비추어 볼 때, 설문의 길이에 따른 질의 저하는 응답자가 질문 내용을 읽지도 않고 대답하는 무작위 반응과는 다르다는 것을 시사한다.

일관성 지표가 반응의 질의 차이를 잘 반응하지 못한 데 대한 하나의 해석을 김형준 등(2014)의 연구 결과에서 찾아 볼 수 있다. 김형준 등(2014)은 전체 문항에 대하여 일탈 반응을 보인 것 외에 모의 실험을 통해 정상 반응 패턴들 중에서 일부 문항만을 무작위 반응으로 조작하여 MRF와 IRT 기반 개인 우도의 판별 정확성을 알아보았다. 그들의 연구 결과에 따르면, 전체 문항 중 30% 이하의 문항에 무작위 반응을 보인 경우에는 MRF와 IRT 모형이 일탈 반응을 잘 탐지하지 못하였다. 따라서, 설문의 후반부에서 일부 문항만을 무작위적으로 반응할 경우 MRF와 IRT 모형은 일탈 반응을 잘 탐지하지 못할 것이다. 반면, 개인별 최대 한줄 응답 길이의 경우는 부분적인 일탈 반응 패턴에 의해서 통계치가 결정되기 때문에 일부 문항에 대해서만 일탈 반응(한줄로 응답하기)을 하더라도 응답의 차이(질의 저하)가 잘 드러난다. 하지만 본 연구는 설문의 질 저하에 따른 응답이 구체적으로 어떻게 바뀌는지

에 대해서는 어떤 가정도 하지 않았다. 그것은 Karabatsos(2003)가 시험 상황에서 일탈 반응에 대한 구체적인 모형을 상정하고, 시뮬레이션한 것과 차이가 나는 점이다. 따라서 좀 더 구체적으로 설문의 길이에 따른 응답의 질 저하를 연구하기 위해서는 후반부 응답에 대한 구체적인 모형을 제안하고 연구할 필요가 있다.

셋째, 여기서 제안된 지수들을 실제 자료에 적용하는 방법에 대하여 예증하였다. 연구 2에서 실무에서 얻은 자료인 자료 1과 자료 2에 앞에서 제안한 여러 가지 지수를 적용하였다. 실제 자료의 특성상 문항이 설문지의 초기 문항들과 후기 문항들이 동일하지 않기 때문에 이를 가능한 한 통제하는 방법을 사용하여 연구 1에서 얻은 결론과 비슷한 결과를 얻을 수 있었다.

위의 연구에 대한 실용적인 함의는 다음과 같다. 설문지의 길이가 증가함에 따라 그 반응의 질이 떨어지는 경향이 관찰되었다. 실제 자료에서 약 20분 정도의 길이에서는 이러한 경향이 뚜렷하지 않았지만, 50분 정도의 길이의 설문지에서는 후반부로 감에 따라 반응의 질이 떨어지는 경향이 발견되었다. 이 결과는 추후 조사표를 구성하는 방식에 대한 시사점을 제공할 것이다. 설문지가 길어질 경우에는 조사에서 중요한 질문은 앞부분에 위치하는 것이 바람직할 것이다.

설문지의 길이의 효과는 당연히 조사자의 성실성이나 능력 등과 피조사자의 성실성과 태도 등에 의하여 달라질 수 있다. 특정한 조건하에서 어느 정도가 적절한 질문지의 길이인지에 대해서는 수집된 자료를 이용하여 이 논문에서 제시되는 기법들을 사용하여 분석하여야 확인할 수 있을 것이다.

이 연구는 또한 질문지의 길이의 효과를 분석하는 실험설계에 대한 함의도 제공할 수 있다. 예컨대 문항의 위치에 따른 응답의 질 저하를 확인하기 위해서는 역문항을 사용하거나, 사전 검사에서 얻어진 응답 패턴을 통해 피험자가 성실하게 응답했을 때 렘펠-지브 복잡도가 최대가 나오도록 문항의 순서를 조정할 수 있을 것이다.

참고문헌

금명자. 1984. ≪내담자의 승인욕구와 상담자의 자기-공개가 내담자의 자기-공개에 미치는 효과≫. 서울대학교 석사학위 논문.

김득란. 1993. ≪양성적 남녀의 성역할 반응양식과 관계변인의 분석≫. 성균관대학교 박사학위 논문.

김아영. 2002. “학업동기 척도 표준화 연구.” ≪교육평가연구≫ 15(1): 157-184.

김완석. 1994. “한국형 인지욕구 척도 개발연구.” ≪한국심리학회지: 산업 및 조직≫ 7: 87-100.

김지혜. 1991. ≪자기초점화 주의가 불안에 미치는 영향≫. 고려대학교 박사학위 논문.

김화자. 1998. ≪대학생의 자아 정체감 수준에 따른 역기능적 가족구조와 사회적 문제해결능력의 차이 연구≫. 연세대학교 석사학위 논문.

김형준・김청택・장병탁. 2014. “마코프랜덤필드를 이용한 설문지에서의 비정상 반응 탐지.” ≪한국정보과학회 동계학술발표회 논문집≫ 41: 577-579.

박아청. 1996. “한국형 자아정체감검사 개발에 관한 연구.” ≪한국심리학회지: 상담 및 심리치료≫ 15(1): 140-162.

서봉연. 1975. ≪자아정체감 형성에 관한 심리학적 연구≫. 경북대학교 박사학위 논문.

이영호. 1993. ≪귀인양식, 생활사건, 사건귀인 및 무망감과 우울의 관계: 공변량 구조모형을 통한 분석≫. 서울대학교 박사학위 논문.

이훈진. 1997. ≪편집증과 자기개념 및 귀인양식≫. 서울대학교 박사학위 논문.

장은경. 2010. ≪반응자 적합도와 마코프 체인을 이용한 한국판 MMPI-2의 반응왜곡 탐지≫. 서울대학교 석사학위 논문.

전미현. 2013. ≪완벽주의, 자기효능감, 성실성이 지연행동에 미치는 영향≫. 대구대학교 석사학위 논문.

최정훈・이정윤. 1994. “사회적 불안에서의 비합리적 신념과 상황 요인.” ≪한국심리학회지: 상담과 심리치료≫ 6(1): 21-47.

황경렬. 1997. “행동적, 인지적, 인지-행동 혼합적 시험불안 감소 훈련의 효과 비교.” ≪한국심리학회지: 상담과 심리치료≫ 9(1): 57-80.

Aboy, M., R. Hornero, D.Abásolo, and D. Álvarez. 2006. “Interpretation of the Lempel-Ziv Complexity Measure in the Context of Biomedical Signal Analysis.” *IEEE Transactions on Biomedical Engineering* 53(11): 2282-2288.

Baer, R.A., J. Ballenger, D.T. Berry, & M.W. Wetter. 1997. “Detection of Random Responding on the MMPI-A.” *Journal of Personality Assessment* 68(1): 139-151.

Beck, A.T., A. Weissman, D. Lester, and L. Trexler. 1974. “The Measurement of Pessimism: The Hopelessness Scale.” *Journal of Consulting and Clinical Psychology* 42(6): 861-865.

Benjamini, Y. and Y. Hochberg. 1995. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society. Series B*(Methodological) 57(1): 289-300.

Biemer, P.P. 2010. “Overview of Design Issues: Total Survey Error.” in Marsden, P.V. and J.D. Wright(eds.). *Handbook of survey research*. Emerald Group Publishing, 27-57.

Birenbaum, M. 1985. “Comparing the Effectiveness of Several IRT Based Appropriateness Measures in Detecting Unusual Response Patterns.” *Educational and Psychological Measurement* 45(3): 523-534.

Bishop, C. 2006. *Pattern Recognition and Machine Learning*. Springer.

Cronbach, L.J. 1951. “Coefficient Alpha and the Internal Structure of Tests.” *Psychometrika* 16(3): 297 - 334.

Deutskens, E. De Ruyter, K.M. Wetzels, and P. Oosterveld. 2004. “Response Rate and Response Quality of Internet-based Surveys: An Experimental Study.” *Marketing Letters* 15(1): 21-36.

Duan, N., M. Alegria, G. Canino, T.G. McGuire, and D. Takeuchi, 2007. “Survey Conditioning in Self Reported Mental Health Service Use:

Randomized Comparison of Alternative Instrument Formats." *Health Services Research* 42(2): 890-907.

Drasgow, F., M.V. Levine, and E.A. Williams. 1985. "Appropriateness Measurement with Polychotomous Item Response Models and Standardized Indices." *British Journal of Mathematical and Statistical Psychology* 38(1): 67-86.

Fagerland, M.W. and L. Sandvik. 2009. "Performance of Five Two-sample Location Tests for Skewed Distributions with Unequal Variances." *Contemporary Clinical Trials* 30(5): 490-496.

Ferrando, P. & E. Chico. 2001. "Detecting Dissimulation in Personality Test Scores: A Comparison between Person-fit Indices and Detection Scales." *Educational and Psychological Measurement* 61(6): 997-1012.

Galesic, M. and M. Bosnjak. 2009. "Effects of Questionnaire Length on Participation and Indicators of Response Quality in a Web Survey." *Public Opinion Quarterly* 73(2): 349-360.

Goetz, E.G., T.R. Tyler, and F.L. Cook. 1984. "Promised Incentives in Media Research: A Look at Data Quality, Sample Representativeness, and Response Rate." *Journal of Marketing Research* 21(2): 148-154.

Hammersley, J.M. and P. Clifford. 1971. "Markov Fields on Finite Graphs and Lattices." Unpublished paper.

Herzog, A. and J. Bachman. 1981. "Effects of Questionnaire Length on Response Quality." *Public Opinion Quarterly* 45(4): 49-59.

Jensen, P.S., H.K. Watanabe, and J.E. Richters. 1999. "Who's Up First? Testing for Order Effects in Structured Interviews Using a Counterbalanced Experimental Design." *Journal of Abnormal Child Psychology* 27(6): 439-436.

Johnson, O., D. Sejdinovic, J. Cruise, R. Piechocki, and A. Ganesh. 2013. "Non-Parametric Change-Point Estimation using String Matching Algorithms." *Methodology and Computing in Applied Probability* 1-22.

Kang, Y., J.R. Harring, and M. Li. 2014. “Reexamining the Impact of Nonnormality in Two-Group Comparison Procedures.” *The Journal of Experimental Education*(ahead-of-print), 1-28.

Karabatsos, G. 2003. “Comparing the Aberrant Response Detection Performance of Thirty-six Person-fit Statistics.” *Applied Measurement in Education* 16(4): 277-298.

Killick, R., P. Fearnhead, and I.A. Eckley. 2012. “Optimal Detection of Changepoints with a Linear Computational Cost.” *Journal of the American Statistical Association* 107(500): 1590-1598.

Killick, R. and I. Eckley. 2013. *Changepoint: An R package for changepoint analysis*. R package version, 1.

Lambert, D. 1992. “Zero-inflated Poisson Regression, With an Application to Defects in Manufacturing.” *Technometrics* 34(1): 1-14.

Levine, M.V. and D.B. Rubin. 1979. “Measuring the Appropriateness of Multiple Choice Test Scores.” *Journal of Educational Statistics* 4(4): 269-290.

Li, M.F. and Olejnik, S. 1997. “The Power of Rasch Person-fit Statistics in Detecting Unusual Response Patterns.” *Applied Psychological Measurement* 21(3): 215-231.

Loevinger, Jane. 1954. “The attenuation paradox in test theory.” *Psychological Bulletin* 51(5): 493.

Micceri, T. 1989. “The Unicorn, the Normal Curve, and Other Improbable Creatures.” *Psychological Bulletin* 105(1): 156.

Miller, G.A. 1955. “Note on the Bias of Information Estimates.” *Information Theory in Psychology: Problems and Methods* 2: 95-100.

Mullahy, J. 1986. “Specification and Testing of Some Modified Count Data Models.” *Journal of Econometrics* 33(3): 341-365.

Murphy, K. 2012. *Machine Learning: A Probabilistic Perspective*. MIT Press.

Nering, M.L. 1997. "The Distribution of Indexes of Person Fit Within the Computerized Adaptive Testing Environment." *Applied Psychological Measurement* 21(2): 115-127.

Nichols, D.S., R.L. Greene, and P. Schmolck. 1989. "Criteria for Assessing Inconsistent Patterns of Item Endorsement on the MMPI: Rationale, Development, and Empirical Trials." *Journal of Clinical Psychology* 45(2): 239-250.

Osborne, J.W. 2008. "Best Practices in Data Transformation" In Osborne, J.W.(ed.). *Best Practices in Quantitative Methods*(pp. 197-204). Sage.

Reise, S.P. and A. Due. 1991. "The Influence of Test Characteristics on the Detection of Aberrant Response Patterns." *Applied Psychological Measurement* 15(3): 217-226.

Reise, S.P. and N.G. Waller. 1993. "Traitedness and the Assessment of Response Pattern Scalability." *Journal of Personality and Social Psychology* 65(1): 143-151.

Samejima, F. 1969. "Estimation of Latent Trait Ability Using a Response Pattern of Graded Scores." *Psychometrika Monograph* 34(Suppl.): 100-114.

Schmitt, N., D. Chan, J.M. Sacco, L.A. McFarland, & D. Jennigs, 1999. "Correlates of Person Fit and Effect of Person Fit on Test Validity." *Applied Psychological Measurement* 23(1): 41-53.

Shannon, C.E. and Warren Weaver. 1949. *The Mathematical Theory of Communication*. Univ of Illinois Press.

Spielberg, C.D., H.P. Gonzalez, C.J. Taylor, W.D. Anton, B. Algaze, G.R. Ross, and L.G. Westberry, 1980. *Preliminary Manual for the Test Anxiety Inventory*. California: Consulting Psychologist Press.

Urbina, S. 2004. *Essentials of Psychological Testing*. Hoboken, NJ: John Wiley.

Weisberg, H.F. 2009. *The Total Survey Error Approach: A Guide to the New Science of Survey Research*. University of Chicago Press.

Welch, B.L. 1937. "The Significance of the Difference between Two Means

When the Population Variances are Unequal.” *Biometrika* 29(3/4): 350-62.

Ziv, J. and A. Lempel. 1978. “Compression of Individual Sequences via Variable-rate Coding.” *IEEE Transactions on Information Theory* 24(5): 530.

<접수 2014/10/29 수정 2014/12/21 게재확정 2015/1/16>

부록 A. 반응의 질 척도

엔트로피

엔트로피는 분포의 불확실성의 정도를 나타낸다. 만약 실험에서 가능한 사건이 n개 있고, 그들의 확률을 $p_i(1 \leq i \leq n)$로 나타낸다면, 이 확률 실험의 엔트로피는 다음과 같다.

$$E = \sum_{i=1}^{n} p_i \log \frac{1}{p_i}$$

이 논문에서는 최대우도법을 사용해 엔트로피를 추정하였다(Miller 1955 등).

$$\hat{E} = \sum_{i=1}^{n} \hat{p_i} \log \frac{1}{\hat{p_i}}$$

($\hat{p_i} = \frac{n_i}{N}$, N:총 시행 수, n_i : 사건 i의 발생 횟수)

렘펠-지브 복잡도

렘펠-지브 복잡도는 데이터 압축법의 하나인 렘펠-지브 압축법(Ziv & Lempel 1978)을 기반으로 정의되는데 수열(number sequence)의 무작위(randomness) 정도를 나타낸다. 어떤 수열의 렘펠-지브 복잡도는 수열의 처음부터 마지막까지의 숫자를 코딩하면서 새로운 부분수열이 등장하는 횟수를 수열의 길이에 따라 표준화한 값이다(그 자세한 알고리즘은 Aboy et al.(2006)을 참조하라). 만약 어떤 수열의 길이가 n이고 발견된 부분수열의 수가 c_n이라면 렘펠-지브 복잡도는 다음과 같이 정의된다.

$$LZ = \frac{c_n}{b_n}$$

(c_n : 새로운 부분수열의 개수, $b_n = \frac{n}{\log_2 n}$)

패턴의 수 c_n을 구하는 방법을 자세히 설명하면 다음과 같다. 만약 1-2-1-2-1-2-1과 같은 수열이 있을 때, 첫 번째 수 1은 새로운 부분수열이다. 두 번째 수 2는 첫 번째 부분수열과 다르므로 역시 새로운 부분수열이다. 세 번째 수 1은 첫 번째 부분수열과 동일하므로 새로운 부분수열이 아니다. 만약 어떤 위치까지의 부분수열이 새로운 부분수열이 아니라면 다음 위치의 수를 합친 부분수열에 대해 새로운 부분수열인지 확인한다. 따라서 세 번째 수 1과 네 번째 수 2를 합친 1-2가 새로운 부분수열인지 확인한다. 1-2는 첫 번째 부분수열(1)과 두 번째 부분수열(2)과 다르다. 따라서 세 번째 부분수열은 1-2가 된다. 다섯 번째 수 1은 첫 번째 부분수열이고, 다섯 번째 수와 여섯 번째 수를 합친 1-2는 세 번째 부분수열이다. 다섯 번째 수, 여섯 번째 수, 일곱 번째 수를 합친 1-2-1은 새로운 부분수열이다. 따라서 1-2-1-2-1-2-1에서 발견한 새로운 부분수열의 개수 c_5는 4가 된다. 이 수열은 규칙적으로 변화하고 있기 때문에 수열을 쉽게 예측할 수 있고 렘펠 복잡도는 낮다. 반면 1-2-3-1-4-2-3과 같은 수열은 새로운 부분수열을 다섯 개(1, 2, 3, 1-4, 2-3) 찾을 수 있다.

IRT 기반 반응자 우도

문항반응이론(IRT)은 문항에 반응하는 반응자의 패턴을 모형화한다(Urbina 2004). 특정한 구성체(예, 시험불안 등)를 측정하는 문항들은 문항 모수(변별도(α_j)와 난이도(β_j))가 있고 반응자는 시험 불안의 정도를 나타내는 모수(θ_i)가 있어서 반응자의 반응은 아래와 같이 확률적으로 나타낼 수 있다.

$$P_{ij} = P(X_{ij} = 1|\theta_i) = \frac{1}{1+\exp(-\alpha_j(\theta_i - \beta_j))}$$

$i-$번째 응답자의 $j-$번째 문항에 대한 응답을 나타내는 X_{ij}는 응답자의 구성체(예, 시험불안)를 나타내는 θ_i(응답자 모수)와 문항의 모수(α_j, β_j)에 의해서 확률적으로 결정된다. 만약 반응자 모수(θ_i)와 문항 모수(α_j, β_j)가 주어지면, 반응자의 특정한 반응 패턴 $\vec{x_i} = (x_{i1}, x_{i2}, ..., x_{in})$($n$: 문항의 개수)에 대한 로그 우도는 식(1)과 같이 계산할 수 있다. 이 우도(l_0)를 활용하여 반응자의 반응 패턴이 반응자 모수와 문항모수에 비추어 적절한지 판단할 수 있다 (Levine & Rubin 1979).

$$l_0(\vec{x_i} = (x_{i1}, x_{i2}, ..., x_{in})|\theta_i) = \sum_{j=1}^{n}[x_{ij}\log P_{ij} + (1-x_{ij})\log(1-P_{ij})] \quad (1)$$

우도가 낮은 반응을 일탈(aberrant) 반응이라고 부르는데, 문항 모수 추정치가 특정한 집단 속에서 추정된다는 점을 감안하면 그것은 개인의 반응 패턴이 집단의 반응 패턴에서 일탈되어 있다는 의미로 해석할 수 있다. Drasgow et al.(1985)는 모수 추정치를 통해 구한 l_0를 표준화한 l_z를 제안하였다(식(2)).

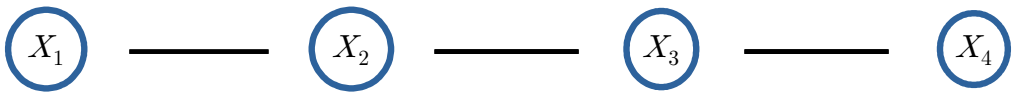
$$l_z = \frac{l_0 - E(l_0)}{[Var(l_0)]^{1/2}} \quad (2)$$

$$E(l_0) = \sum_{j=1}^{n}[P_{ij}\log P_{ij} + (1-P_{ij})\log(1-P_{ij})]$$

$$Var(l_0) = \sum_{j=1}^{n}[P_{ij}(1-P_{ij})\log[(P_{ij}/(1-P_{ij})]^2$$

MRF(Markov Random Fields) 우도

MRF는 마코프체인을 일반화한 것으로, 컴퓨터 시각 분야에서 많이 사용되는 기법이다(Bishop 2006; Murphy 2006). 본 연구에서는 j-번째 반응은 $(j-1)$-번째 반응과 $(j+1)$-번째 반응에 의해서만 영향을 받고 그 외의 반응과는 독립이라고 가정하는 선형 연쇄 모형(linear-chain model)을 사용하였고 오류함수는 이징 모형(ising model)의 오류 함수를 변형하였다. <그림 1>은 4개의 문항에 대한 응답 X_1, X_2, X_3, X_4에 대한 선형 연쇄 모형을 보여준다.



〈그림 1〉 이징 모형(문항들의 반응에 대한 모형)

반응 $X_1, X_2, ..., X_n$의 결합 확률은 다음과 같이 나타낼 수 있다.

$$P(X_1, X_2, ..., X_n) = P(X_j \mid \bigcap_{k, k \neq j} X_k) P(\bigcap_{k, k \neq j} X_k) \quad (3)$$

$$= P(X_j \mid X_{j-1}, X_{j+1}) P(\bigcap_{k, k \neq j} X_k) \quad \text{(독립성 가정에 의해)}$$

$$(X_0 = 1, X_{n+1} = 1, P(X_0 = 1) = 1, P(X_{n+1} = 1) = 1)$$

해머즐리-클리포드 정리(Hammersley-Clifford theorem)에 의하면 식(3)은 다음과 같이 나타낼 수 있다(Hammersly & Clifford 1971)

$$P(X_1, X_2, \cdots, X_n) = \frac{1}{Z}\psi_1(X_1)\psi_n(X_n)\prod_{j=1}^{n-1}\psi_j(X_j)\psi_{j,j+1}(X_j, X_{j+1}) \quad (4)$$

$$Z = \sum_{X_1}\sum_{X_2}\cdots\sum_{X_n}\left\{\psi_1(X_1)\psi_n(X_n)\prod_{j=1}^{n-1}\psi_j(X_j)\psi_{j,j+1}(X_j, X_{j+1})\right\}$$

이때, 식 4의 $\psi_j(X_j)$와 $\psi_{j,j+1}(X_j, X_{j+1})$ 는 이징 모형에서처럼 식 5와 식 6과 같이 정의된다.

$$\psi_j(X_j = l) = \left(e^{\theta_j}, e^0\right)_{l+1} \quad (5)$$

$$\psi_{j,j+1}(X_j = l, X_{j+1} = m) = \begin{pmatrix} e^0 & e^{\lambda_{j,j+1}^{(k)}} \\ e^{\lambda_{j,j+1}^{(k)}} & e^0 \end{pmatrix}_{l+1, m+1} \quad (6)$$

$$\begin{cases} \mathrm{k} = 1 & \mathrm{if} \quad (j, j+1) - \text{정이행 관계} \\ k = 2 & \mathrm{if} \quad (j, j+1) - \text{역이행 관계} \end{cases}$$

본 연구에서는 $\lambda_{j,j+1}^{(1)}$과 $\lambda_{j,j+1}^{(2)}$가 j에 상관없이 동일하다고 설정하였다. 따라서 이 모형의 모수는 $\Theta = (\theta_1, \theta_2, ..., \theta_n)$와 $\Lambda = (\lambda^{(1)}, \lambda^{(2)})$이다. 이 모수의 의미는 다음과 같이 해석할 수 있다.

θ_j는 $\frac{\psi_j(X_j = 1)}{\psi_j(X_j = 0)} = \frac{e^{\theta_j}}{e^0} = e^{\theta_j}$에서 $j-$번째 문항에 대해 '그렇다(1)'을 선택할 확률과 '아니다(0)'을 선택할 확률의 비를 대략적으로 나타낸다고 볼 수 있다 (하지만 정확한 확률값은 아니다. 왜냐하면 $\psi_{j-1,j}\psi_{j,j+1}$는 고려되지 않았기 때문이다). 따라서 고전 검사 이론의 난이도와 유사한 개념으로 해석할 수 있다.

보통 설문지를 만들 때, 설문의 집중도를 높이고 한줄 응답과 같은 반응 패턴을 감소시키기 위해 역문항을 사용한다. 예를 들어 '반사회성'을 측정하는 다음의 네 문항 설문지를 보자(<표 1>).

<표 1> 설문지 구성 예시

문항번호(특징)	문항 내용
1 번(정문항)	다른 사람들의 욕구나 감정에 둔감하다
2 번(역문항)	호감이 느껴지고 괜찮은 인상을 준다
3 번(정문항)	공격적으로 행동한다
4 번(정문항)	자기중심적이다

반사회성을 가진 사람들은 보통 1번, 3번, 4번 문항에서 '그렇다(1)'로 응답하고 2번 문항에서는 '아니다(0)'로 응답한다. 따라서, 1번, 3번, 4번 문항은 '정문항'이라 정의하고, 2번 문항은 '역문항'이라 정의한다. 여기서 정문항과 정문항 혹은 역문항과 역문항이 이어질 경우를 '정이행 관계'라고 정의하고, 정문항과 역문항 혹은 역문항과 정문항이 이어질 경우를 '역이행 관계'라고 정의하자.

$\lambda_{j,j+1}^{(1)}$은 '정이행 관계'에서 $\psi_{j,j+1}$를 결정하고, $\lambda_{j,j+1}^{(2)}$는 '역이행 관계'에서 $\psi_{j,j+1}$를 결정한다. 응답자의 반응에 대해 $j, j+1-$문항에 대한 반응이 $(0,0)$ 또는 $(1,1)$일 때 '정이행 반응', $(1,0)$ 또는 $(0,1)$일 때 '역이행 반응'이라고 정의하자. $\lambda_{j,j+1}^{(1)}$는 대략적으로 $(j, j+1)$-문항이 '정이행 관계'일 때 응답자의 반응이 '역이행 반응'일 확률과 '정이행 반응'일 확률의 비를 결정하고, $\lambda_{j,j+1}^{(2)}$는 $(j, j+1)$-문항이 '역이행 관계'일 때 동일한 확률의 비를 결정한다고 생각할 수 있다. 전체 반응자들의 반응 패턴을 통해 모수인 (Λ, Θ)를 추정하면 그것을 활용하여 개인의 반응패턴에 대한 개인 우도를 구할 수 있다.

부록 B. 표준편차와 엔트로피의 차이

표준편차는 주어진 자료가 평균에서 떨어져 있는 정도를 나타낸다면, 엔트로피는 특정한 중심점을 가정하지 않고 자료가 흩어져 있는 정도를 나타낸다고 볼 수 있다. 표준편차와 엔트로피의 차이를 좀 더 확연히 이해할 수 있도록 설문 응답 자료를 생각해보자(<표 1>).

<표 1> 예시 설문 응답 자료

피험자	문항									
	1	2	3	4	5	6	7	8	9	10
A	3	3	4	3	4	4	3	4	3	4
B	2	2	5	5	5	2	5	2	2	5

<표 1> 에서 피험자 A는 총 10 문항에 대해 3으로 응답한 경우가 5번, 4로 응답한 경우가 5번이었다. 따라서 평균은 3.5이고, 표준편차는

$$\sqrt{\frac{5(3-3.5)^2+5(4-3.5)^2}{10-1}} \approx 0.527$$

이다. 피험자 A의 응답에 대해 경험적 엔트로피(empirical entropy)를 구하면

$$-(\frac{1}{2}log\frac{1}{2}+\frac{1}{2}log\frac{1}{2})=1$$

이다.

피험자 B의 응답은 2가 5번, 5가 5번이었다. 평균은 피험자 A와 같이 3.5이지만, 표준편차는

$$\sqrt{\frac{5(2-3.5)^2+5(5-3.5)^2}{10-1}} \approx 1.5811$$

이다. 피험자 B의 응답이 피험자 A의 응답에 비해 평균에서 멀리 떨어진 만큼 표준편차가 증가하였음을 알 수 있다. 하지만 피험자 B의 응답에 대해 경험적 엔트로피를 구해보면

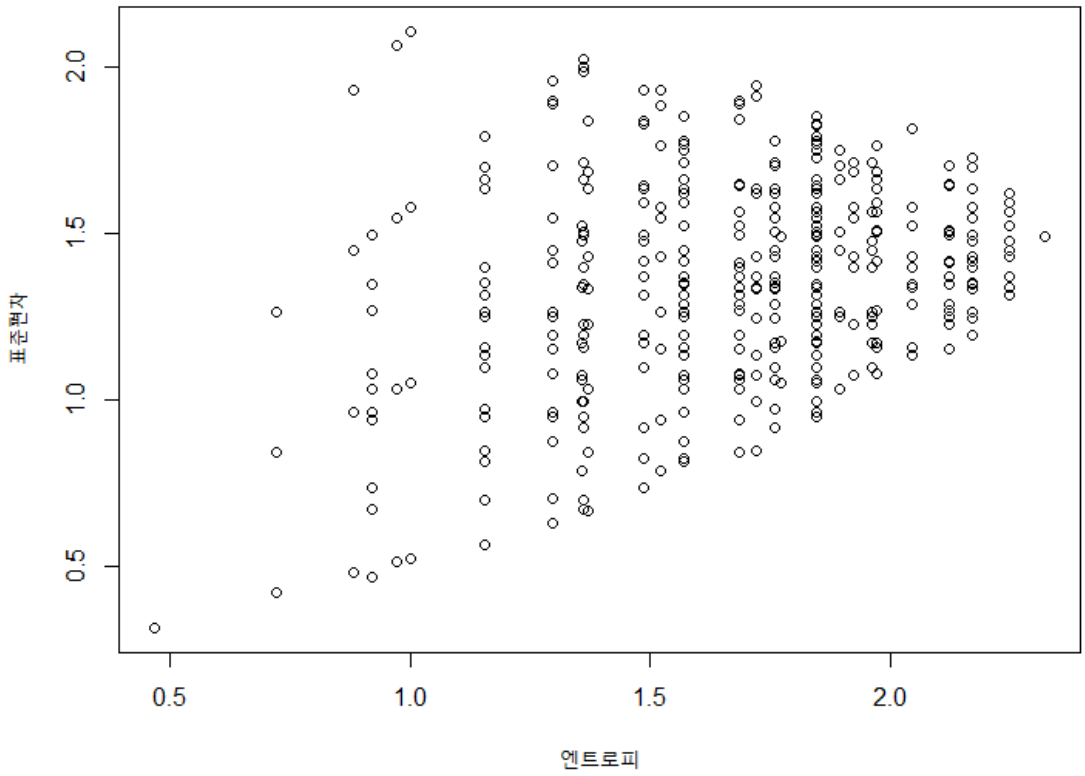
$$-(\frac{1}{2}log\frac{1}{2}+\frac{1}{2}log\frac{1}{2})=1$$

로 피험자 A의 경우와 동일하다. 이것은 엔트로피를 구할 때에는 특정한 중심점을 가정하지 않기 때문이다.

위에서 예로 든 5점 척도의 리커트 문항 10문항에 대한 응답에서 엔트로피와 표준편차의 최소값과 최대값을 알아보자. 엔트로피와 표준편차의 최소값은 0이다. 만약 모든 응답이 같다면 엔트로피와 표준편차는 모두 0이 된다. 엔트로피와 표준편차의 최대값은

$$-\sum_{i=1}^{5}\frac{1}{5}log\frac{1}{5}\simeq 2.322 \text{와} \sqrt{\frac{5(1-3)^2+5(5-3)^2}{10-1}} \approx 2.108$$

이 된다. 엔트로피가 최대값이 되는 경우는 모든 응답의 확률이 동일한 경우(1에서 5번까지의 응답을 각각 2번씩 한 경우)이고, 표준편차가 최대값이 되는 경우는 모든 응답이 평균에서 가장 멀리 떨어진 경우(1번 응답을 5번, 5번 응답을 5번 한 경우)이다. <그림 1>은 리커트 5점 척도 10개 문항에 대한 응답을 무작위로 생성하여 표준편차와 엔트로피를 그래프에 나타낸 것이다. 표준편차나 엔트로피 모두 최저/최대값에서는 서로를 심하게 제약하지만, 중간값에서는 다소 자유로운 관계가 있음을 알 수 있다. 두 값은 상관관계는 0.3 정도이다.



〈그림 1〉 리커트 5점 척도 10 문항에서 엔트로피와 표준편차의 관계