한국심리학회지 : 일반
Korean Journal of Psychology: General
2017, Vol. 36, No. 3, 359-396
http://dx.doi.org/10.22257/kjp.2017.09.36.3.359

심리학 연구에서 재현성 위기의 현황과 원인 및 대안 모색에 대한 개관

김 빛 나 서울대학교 심리학과

최 준 원† 상명대학교 경영학부

고 현 석 공군사관학교 인문학과

최근 심리학계에서는 심리학 연구의 재현성에 대한 치열한 논란이 전개되어 왔다. 점화 효과 등 학계 및 일반 대중에게 큰 관심을 끌었던 유명 연구들이 반복검증에 실패하였고, 뒤따라 착수된 대규모 반복검증 프로젝트들에서의 반복검증 성공률도 만족스럽지 못한 것으로 나타나 과학으로서 심리학의 위상이 위협받는 계기가 되었다. 이 과정에서 통계적 추론절차, 연구 관행 그리고 출판편향 등이 재현성 위기의 원인들로서 조명되었으며, 그에 따라 심리학의 재현성을 증진하기 위한 해결책들(신뢰구간, 효과 크기의 사용, 메타분석, 베이지안 통계, 연구자료와 방법론의 투명성 증가)도 제안되고 있다. 앞으로 어떻게 해결되는가에 따라 작금의 재현성 위기는 향후 심리학이 발전해 나가는 기회가 될 가능성도 가지고 있다. 이에 본 논문에서는 심리학 내 재현성 위기가 전개되어온 현황을 소개하고 현재까지 도출된 원인과 잠정적 해결책들을 종합적으로 개관함으로써, 국내에서 관련 논의가 촉진될 수 있는 기초를 마련하고자 한다.

주요어 : 반복검증, 재현성, 출판편향, 방법론

† 교신저자: 최준원, 상명대학교 경영학부, (03016) 종로구 홍지문2길 20 상명대학교 경영대학
Tel: 02-781-7786, Email: joonwonchoi@smu.ac.kr

심리학적 연구와 재현성

과학으로서의 심리학

"정치, 경제, 조금 더 보편적으로 말해서 모든 사회과학을 떠받치고 있는 학문은 명백하게도 심리학이다. 심리학 원리로부터 사회과학의 법칙들을 이끌어낼 날이 언젠가 찾아올 것이다"(Pareto, 1906: Thaler, 2016에서 재인용).

독립적인 학문 분야로서 심리학의 공식적 출발점이 Wundt가 라이프치히 대학 내 심리학 실험실을 설치한 1879년이었다는 점을 고려한다면, 이로부터 불과 20여 년이 지난 시점에 이루어진 Pareto의 이러한 예측이 당시로서는 얼마나 획기적인 것이었는지 짐작할 수 있다. 그렇다면 Pareto의 예측 이후 100여 년이 지난 오늘날 심리학의 위상은 과연 어떠한가? 국내 상황을 살펴보면, 심리학은 한국연구재단 등재지 가운데 최근 영향력 지수가 가장 높았던 등재지 100종 중 4번째로 높은 순위를 차지하고 있다(한국연구재단, 2016)[1]. 또한 조사에서 1~3위를 차지한 교육학, 사회복지학, 경영학 연구들 또한 심리학의 이론과 변수를 활용하고 있다는 점을 고려하면, 심리학의 학문적 영향력은 표면적인 지수로 파악되는 것 이상이라고 할 수 있을 것이다. 더불어 최근의 베스트셀러 목록에 반영되듯 심리학과 관련 서적에 대한 일반 대중들의 관심도 크게 증가하고 있는 추세이다(교보문고, 2016).

한편 과학으로서 심리학에 대한 다소 회의적인 시각도 존재하고 있는 것으로 보인다. Lilienfeld(2012)는 이를 '심리학에 대한 대중의 회의주의(public skepticism of psychology)'라고 표현하고 그 원인을 진단한 바 있다. 즉, 심리학은 그저 상식적인 것일 뿐이라는 생각, 심리학은 과학적인 방법론을 적용하지 않는다는 생각, 개인들의 독특성을 고려할 때 심리학은 의미 있는 일반화가 불가능하다는 생각, 심리학 연구의 재현성과 예측력에 대한 회의적 시각, 심리학의 사회적 기여가 생물학, 화학, 물리학, 경제학 등의 분야와 비교하여 빈약하다는 회의적 시각 등으로 인해 대중들에게 심리학의 과학적 근간이 의심되고 있다는 것이다(Lilienfeld, 2012). 타 분야의 일부 학자들 또한 심리학이 엄격한 과학적 기준을 필요로 하는 분야에서 요구하는 사항 중 가장 기본적인 조건에 해당하는 명확하게 정의된 용어와 정량화 가능성에 대한 기준을 충족시키지 못하기 때문에 진정한 과학으로 볼 수 없다는 과격한 주장을 하기도 한다(Berezow, 2012).

이러한 현상은 부분적으로 심리학을 비롯한 사회과학이 지니는 연성 과학(soft science)으로서의 특성에서 기인한다고 볼 수 있다[2]. Wilson(2012)은 물리학이나 화학과 같은 경성 과학(hard science)은 현상에 대한 정확하고 섬세한 예측이 가능하지만, 심리학을 포함한 사회과학 분야에서 연구대상으로 삼는 사회에 대한 가설은 실험에 의해 입증되거나 반증되

1) '심리과학'이라는 중분류 기준이며, 한국연구재단 등재지 2,352종 가운데 최근 5년간(2011~2015) 영향력 지수가 가장 높았던 등재지 100종을 살펴보면, 심리과학이 7종으로 4위를 기록하였다.

2) 경성 과학과 연성 과학은 통상적으로 자연과학과 사회과학을 구분하기 위해 사용되는 표현이다(Hedges, 1987). 이러한 구분의 이면에는 방법론의 엄격함 정도에 따라 과학을 위계적으로 배열했을 때 물리학이 가장 높은 곳에, 생물학은 중간 수준에, 사회과학은 가장 낮은 곳에 위치하고 있다는 과학계의 인식이 반영되어 있다(Fanelli, 2010).

기에는 근원적인 한계가 있다고 주장하였다. 그런데 여기서 문제는 경성 과학의 입장에서 심리학을 비판하는 주장들은 과학을 매우 좁은 범위로 한정하고 있다는 점이다. 일례로 미국의 국립보건원(National Institutes of Health) 산하 행동 및 사회과학 조사국(Office of Behavioral and Social Sciences Research)에서는 과학에 대한 지나치게 좁은 정의로 인해 행동 및 사회과학 분야가 과학교육 체계에서 소외되는 상황을 비판하면서, 연방정부 차원에서 과학에 대한 인식을 변화시키는 작업에 나설 것을 촉구했다(Dunn, 2015). 이러한 흐름에 발맞추어 심리학의 연구 성과를 사회에 기여할 수 있는 제도와 정책의 근간으로 활용하고자 하는 시도가 미국과 영국 등지를 중심으로 확대되는 추세이다(Thaler, 2016). 그 일환으로 미국의 오바마 대통령은 2009년 규제정보국(Office of Information and Regulatory Affairs) 국장에 행동경제학자를 임명하였다. 그러나 2015년 9월 경 미국에서 행정명령에 의해 정부 정책에 심리학 기반 행동과학의 개입 여지가 확대되면서 주요 언론들 간에는 엇갈린 평가들이 양산되었다. 뉴욕타임스는 대체로 긍정적인 평가와 전망을 내놓은 반면, 월스트리트 저널은 심리학 실험을 신뢰하기 어렵다는 점에서 우려를 표명했다(한국금융신문, 2015. 11. 16). 이처럼 심리학에 대한 대중적 관심의 증가와 지속적인 학술적 성과의 축적에도 불구하고, 과학으로서의 심리학의 위상과 그 사회적 활용가능성에 대해서는 의문의 시각이 상존하고 있는 실정이다.

심리학 연구의 재현성 논란

이러한 상황에서 심리학이 과학으로서의 정체성을 확립할 수 있는 가장 기본적인 요건 중의 하나는 바로 연구결과의 재현성(reproducibility)일 것이다[3]. 재현성은 과학을 규정하는 속성 중 하나로 과학과 비과학을 구분하는 중요한 기준이 된다(Open Science Collaboration, 2012, 2015). 이는 단순한 전제에 기초하고 있다. 즉, 연구결과가 실재하는 것이고 견고한 것이라면, 동일한 절차에 따라 어느 연구자가 연구를 수행하더라도 같은 결과를 얻어야 한다는 것이다(Simons, 2014). 만약 심리학의 연구결과를 과학적으로 신뢰하고 사회의 제도 및 정책 수립 등에 적극적으로 활용하고자 한다면, 재현성의 담보가 중요한 선결과제가 된다고 할 수 있다.

문제는 과학으로서의 심리학의 정체성과 직결된 재현성 문제와 관련하여 최근 심리학계가 심각한 위기와 논란을 겪고 있다는 것이다

3) 재현성(reproducibility)이라는 용어와 반복가능성(replicability)이라는 용어는 유사한 개념으로 혼용되어 사용되는 경향이 있다(Bollen, Cacioppo, Kaplan, Krosnick, & Olds, 2015). 먼저 재현성이란 "과학적인 연구가 반복될 때 일관된 결과가 관찰되는 정도"를 의미한다(Open Science Collaboration, 2012). 이에 비해 반복가능성은 "새로운 자료를 수집해서 원 연구와 동일한 절차에 따라 연구결과를 반복할(duplicate) 수 있는 능력"으로 정의된다(Bollen et al., 2015). 한편 Open Science Collaboration(2015)에서는 반복검증(replication)을 "새로운 자료를 이용해서 연구결과의 재현성을 확립하는 수단"으로 규정하였다. 본고에서는 이러한 정의들을 종합하여 재현성을 "원 연구의 결과를 경험적으로 반복할 수 있는 능력"으로, 반복검증은 "재현성을 확립하기 위해 새로운 자료를 수집해서 동일한 절차에 따라 원 연구의 결과를 재확인하는 시도 및 과정"으로 간주하였다. 이에 따르면 재현성은 반복검증보다 광의의 개념이며 반복검증의 목적은 재현성을 확립하기 위한 것으로 이해할 수 있다.

(Lindsay, 2015; Pashler & Wagenmakers, 2012; Yong, 2012). 최근 연구결과의 재현성 위기는 비단 심리학 내의 일만이 아니며, 생물학과 의학, 경제학 전반에 걸쳐 전개되어 왔다(Nissen, Magidson, Gross, & Bergstrom, 2016). Ioannidis(2005)는 "왜 출판된 대부분의 연구결과들이 거짓인가(Why most published research findings are false)"라는 논쟁적인 제목의 논문을 발표하고, 베이지안 추론에 근거하여 의학 및 생명과학 분야에서의 재현성 문제를 지적한 바 있다. 역사적으로 볼 때 심리학 내에서의 재현성 위기가 처음은 아니지만(Gergen, 1973; Laws, 2016; Stroebe, 2016), 2011년 이후 이와 관련된 일련의 사건들이 벌어지며 본격적으로 대규모의 담론이 촉발된 측면이 있다.

논문의 개관범위와 방법

본 논문은 재현성 위기의 근원을 정확하게 진단하고 이를 개선하기 위한 다양한 관점과 대안들이 활발하게 논의되어야 한다는 문제의식에서 출발하였다. 더욱이 저자들이 아는 한도 내에서는 아직까지 국내에서 심리학 내의 재현성 이슈를 다룬 문헌이 드물며(박준석, 2015), 이와 관련한 본격적인 논의가 이루어지지 못한 것으로 보인다. 따라서 본 논문에서는 심리학 내에서 재현성 위기가 전개되어 온 현황을 살펴보고, 재현성 위기를 초래한 원인들을 다양한 각도에서 조망하고자 한다. 또한 심리학 연구의 재현성을 증진시키기 위한 개선 방법에 대한 논의와 제안을 하고자 한다. 이를 통해 향후 국내에서도 심리학 연구의 재현성을 높이기 위한 논의와 개선 방안 모색이 이루어지기 위한 기초적인 토대를 제공하려 한다.

이를 위하여 국내외 학술자료 검색 엔진(PubMed, PsycINFO, RISS)에서 '반복검증(replication)', '재현성(reproducibility)', '심리학(psychology)'라는 3가지 검색어로 2017년 1월 현재까지 출판된 문헌 약 500여 건을 1차로 탐색하였다. 또한 재현성 위기가 대두된 이후 발간된 학술지의 특별호(*Journal of Experimental Social Psychology* 66권, *Perspectives on Psychological Science* 7권 6호, 9권 1호, 11권 4·5호, *Social Cognition* 32권, *Social Psychology* 45권 3호)를 포함시켰다. 이 중에서 내용 상 주제 적합성이 맞지 않거나 서로 중복되는 문헌들을 제외하고 최종적으로 개관에 포함된 논문은 총 79개였다.

심리학 내 재현성 위기의 현황

재현성 위기를 촉발한 주요 사건들

심리학을 위기에 처하게 한 가장 분명한 사건은 역설적이게도 심리학의 주요 연구로부터 촉발되었다(Yong, 2012). Cornell 대학의 심리학 교수인 Darly Bem은 사회 및 성격심리학 분야의 권위 있는 학술지인 *Journal of Personality and Social Psychology*(이하 *JPSP*)에 논쟁적인 논문을 게재하였다. 그는 'Feeling the future'라는 논문에서 미래사건을 미리 예측하는 *psi* 현상을 9개의 실험연구를 통해 반복적으로 입증했다면서 초감각적인 지각(extrasensory perception)이 실재하는 현상이라고 주장하였고(Bem, 2011), 이 논문은 출판 직후부터 학계에 논란을 일으켰다. 당시 *JPSP*의 편집자들은 해당 논문이 엄격한 동료 심사과정을 통과하였으며, 편집자의 역할은 특정 가설을 승인하는 것이 아니라 과

학의 발전을 촉진시키는 것이기 때문에 연구결과와 자료가 사실이라는 저자의 주장을 받아들여 논문의 출판을 결정하게 되었다고 설명하였다. 또한 편집자들은 전통적 인과성의 이해를 뒤집는 결과를 보여준 이 논문으로 말미암아 사회인지 연구 분야에서 심도 있는 논의와 반복검증 시도가 이루어지기를 기대한다고 당부하였다(Judd & Gawronski, 2011). 그러나 편집자들의 기대와 달리, 그의 논문은 후에 심리학 연구결과의 신뢰성을 의심하게 만드는 결과를 초래하게 된다.

또 다른 결정적 사건으로는 같은 해에 발각된 Diederik Stapel의 연구부정 행위를 들 수 있다. 네덜란드 Tilburg 대학의 심리학 교수였던 Stapel은 사회심리학 분야에서 떠오르는 스타로 그가 발표한 연구결과들은 *Science*와 *JPSP*를 비롯한 저명 학술지는 물론 언론매체에도 소개되어 화제가 되었다. 하지만 놀랍게도 Stapel이 장기간 동안 논문의 데이터를 의도적으로 조작하는 연구부정 행위를 저질러 왔다는 것이 밝혀졌다. 연구부정 행위 조사위원회의 발표에 따르면 Stapel은 1996년부터 2011년까지 최소 34편의 논문에서 데이터를 체계적으로 조작하였다고 한다(Stapel Investigation, 2012: Stroebe, Postmes, & Spears, 2012, p. 671에서 재인용). 이 같은 사실은 학계에 커다란 충격을 주었으며, 주요 언론매체에서도 이 사건을 심리학 연구결과의 신뢰성 문제와 연결 지어 비중 있게 다루었다(Bhattacharjee, 2013). 더욱이 그의 연구 부정행위가 장기간 탐지되지 못했고, 결국 발각된 것도 동료 학자들의 검증에 의한 것이 아니라 내부 고발자의 제보에 의해서였다는 점에서 더 큰 문제가 되었다(Crocker & Copper, 2011; Yong, 2012). 이는 과학적 커뮤니티의 중요한 기능 중 하나인 자정 기능이 정상적으로 작동하지 못했다는 방증이기 때문이다(Pashler & Harris, 2012; Stroebe et al., 2012).

더욱이 학계뿐만 아니라 일반 대중에까지 소개되어 상당한 관심과 이목을 집중시켰던 심리학 연구결과들의 상당수가 일관되게 재현되지 않는다는 점이 드러나게 되었다. 앞서 소개한 Bem(2011)의 논문은 다수의 연구자들이 독립적으로 연구결과를 재현하고자 시도하였으나 원 저자의 주장과 달리 *psi* 현상은 재현되지 않았다(Galak, LeBoeuf, Nelson, & Simmons, 2012; Ritchie, Wiseman, & French, 2012; Robinson, 2011). 또한 학계에서 참신한 결과로 선풍적 인기를 끌었던 점화(priming) 효과도 대체로 재현되지 않는 것으로 나타났다. 대표적으로 Bargh, Chen과 Burrows(1996)는 노인과 관련된 단어(예: 늙은, 쓸쓸한, 무력한 등)에 무의식적으로 노출되는 것만으로도 사람들의 걷는 속도가 노인처럼 느려진다는 점화 효과를 밝혀 학계의 주목을 받았다(2017년

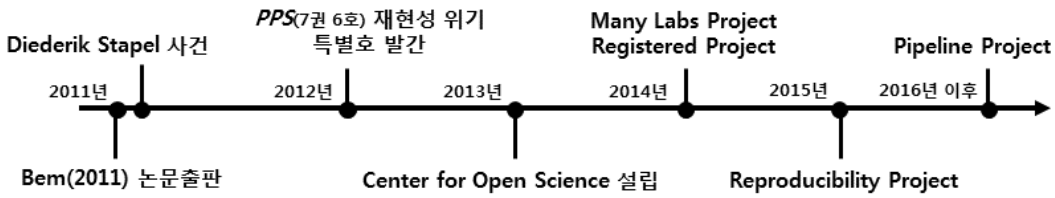

그림 1. 재현성 위기의 연대기적 현황

7월 현재 인용 횟수 4200회). 그러나 해당 논문을 동일한 절차대로 반복검증한 연구에서 원 논문에서 보여주었던 것과 같은 극적인 점화 효과는 발견되지 않는 것으로 나타났다(Doyen, Klein, Pichon & Cleeremans, 2012; Pashler, Harris, & Coburn, 2011). 이처럼 최근에 심리학계에서 일어난 일련의 사건들은 전례 없는 수준으로 신뢰성을 훼손함으로써, 과학으로서의 심리학의 지위를 위협하는 결과를 초래하였다(Lindsay, 2015; Pashler & Wagenmakers, 2012).

대규모 반복검증 프로젝트들의 출현

이에 심리학자들은 이러한 실태를 정확히 진단하고, 해결책을 모색하기 위해 개인적, 제도적 노력을 시도해 왔다. 먼저 Virginia 대학의 Brian Nosek은 대규모의 재현성 프로젝트를 진행하기 위해 Open Science Framework라는 웹사이트를 개설하고 프로젝트에 동참할 학자들을 모집하였다(Open Science Collaboration, 2012; Spellman, 2012). San Diego 대학의 Hal Pashler 또한 누구든 자유롭게 반복검증 결과를 게시하고 공유할 수 있는 PsychFileDrawer라는 웹사이트를 개설하였다(PsychFileDrawer, 2012). 학회 차원의 제도적 노력으로 심리과학협회(Association for Psychological Science: 이하 APS)에서 발행하는 학술지인 *Perspectives on Psychological Science*에서는 심리과학에서의 재현성 문제를 다룬 특별호(7권 6호)를 출판하였다. 해당 호에서는 심리학에서의 검증가능성 문제의 실태와 원인을 분석하고 어떠한 노력을 통해 심리과학의 신뢰성을 증진시킬 수 있는지에 대한 심리학자들의 다양한 견해를 수록하였다. 2013년에는 과학적인 연구의 개방성, 진실성, 재현성을 증진시키기 위해 Center for Open Science라는 조직이 설립되어 심리학 연구의 신뢰성을 증진시키기 위한 프로젝트를 진행하게 되었다(Center for Open Science, 2013).

이와 같은 시도는 최근에 이루어진 4개의 대규모 반복검증 프로젝트의 착수로 이어졌다. 이들 프로젝트는 원 연구와 가급적 동일한 자료와 절차에 의거하여 결과를 재현하려고 하였으며, 지금까지 분절적이고 일회적으로 이루어져왔던 반복검증 문제를 심리학 전반에 걸쳐 촉발시키고 공론화시킨 도화선이 되었다. 지금부터는 이러한 대규모 반복검증 프로젝트들이 어떻게 진행되었는지 살펴보고, 그로부터 얻어진 결과들을 비교 분석하고자 한다(Laws, 2016; Stroebe, 2016).

첫째, Open Science Collaboration이 주도한 Many Labs 프로젝트에서는 12개국 36개 연구그룹이 6,300여명의 참여자를 대상으로 13개의 심리학 연구를 재검증하였다(Klein et al., 2014). 이들은 모든 반복검증을 표준화된 프로토콜에 의하여 컴퓨터로 실시 가능한 방식으로 변환하여 진행하였다. 그 결과 13개 중 10개의 연구(77%)는 효과 크기의 차이는 있을지라도 비교적 일관되게 반복검증되었고, 이는 지금까지 이루어진 대규모 반복검증 연구 프로젝트 중 가장 높은 수치이다. 반면 성조기에 노출되면 보수주의 성향이 증가한다는 성조기 효과(flag effect)와 돈에 노출되면 현재 사회제도를 수용하는 경향이 증가한다는 통화 효과(currency effect) 등 점화와 관련된 2개 연구(Carter, Ferguson, & Hassin, 2011; Caruso, Vohs, Baxter, & Waytz, 2013)는 반복검증에 실패하였으며, 외집단을 상상하는 것만으로도 편견이 감소한다는 상상 접촉 효과(imagined contact; Husnu & Crisp, 2010) 역시 반복검증 결과, 효

과 크기의 신뢰구간이 0에 가까운 것으로 나타났다.

둘째, 2014년 *Social Psychology* 특별호(45권 3호)에서 진행한 Registered 프로젝트에서는 사회심리학에서 중요한 연구결과들을 선별하여 재검증하고자 하였다(Nosek & Lakens, 2014). 원 연구의 저자들은 자료수집 전에 미리 등록된 연구 프로토콜에 따라 반복검증이 진행된다면, 연구결과가 어떻든지 간에 출판을 하는 것에 동의하였다. 여기에 포함된 연구들은 로미오와 줄리엣 효과(Driscoll, Davis, & Lipetz, 1972), 따뜻함 효과(Williams & Bargh, 2008), 어둠 효과(Banerjee, Chatterjee, & Sinha, 2012), 깨끗함 효과(Schnall, Benton, & Harvey, 2008) 등이었다. Many Labs 프로젝트에서의 높은 재검증 비율과는 대조적으로, Registered 프로젝트에서는 13개 중 10개의 연구가 반복검증에 실패한 것으로 나타났다. 구체적으로 부모의 간섭이 심할수록 연인 간에 낭만적인 사랑이 증가한다는 로미오와 줄리엣 효과(Sinclair, Hood, & Wright, 2014)와 신체적인 따뜻함에 잠시 노출되는 것만으로도 친사회적인 행동이 증가한다는 촉진효과(Lynott et al., 2014)에 대한 지지 증거는 발견되지 않았다. 또한 비윤리적인 행동을 회상하는 것이 방을 더 어둡게 지각하도록 만든다는 주장이나 깨끗함을 점화시키는 것이 오히려 도덕적인 판단의 엄격함을 감소시킨다는 정화 효과 역시 경험적으로 재현되지 않았다(Brandt, IJzerman, & Blanken, 2014; Johnson, Cheung, & Donnellan, 2014).

셋째, 현재까지 가장 대규모로 이루어진 반복검증 연구로는 Open Science Collaboration (2015)에서 진행하였던 Reproducibility 프로젝트가 있다. 여기에서는 270명의 협력 연구자들이 2008년 한 해 동안 3개의 주요 학술지(*Psychological Science, JPSP, Journal of Experimental Psychology: Learning, Memory, and Cognition*)에 게재된 논문 중에서 선택 편향(selection bias)을 최소화하면서 동시에 대표성을 가질 수 있는 100개의 연구를 선정해서 이를 반복검증하고자 하였다. Registered 프로젝트와 마찬가지로 Reproducibility 프로젝트 또한 원 저자에 의해 제공된 자료를 사용하였음에도 불구하고 반복검증 연구의 많은 부분(예: p 값, 효과 크기, 재검 연구팀의 주관적 평가, 효과 크기의 메타분석 등)에서 원 연구에 비해 약화된 결과가 나타났다. 즉, 상당수의 연구가 재현되지 않는 것으로 나타났는데, 원 연구에서는 이 중 97개 연구결과가 통계적으로 유의하였던 것에 비하여, 반복검증 연구에서는 36%만이 유의하였으며, 반복검증에서의 평균 효과 크기는 원 연구에서 보고된 것의 1/2 수준이었다. 또한 Reproducibility 프로젝트는 반복검증 연구의 성공 여부가 재검 연구를 수행한 팀의 특성(예: 경험과 전문성) 보다는 원 연구의 강도에 의해 보다 신뢰롭게 예측된다는 사실과 선택된 연구들 중에서는 사회심리학에 비해 인지심리학의 재현성이 더 높다는 점도 추가적으로 밝혀냈다.

넷째, 가장 최근에 이루어진 Pipeline 프로젝트에서는 다국적의 25개 연구 그룹을 모집하여 아직 출판되지 않은 도덕적 판단 실험 10개를 전향적으로 반복검증하였다(Schweinsberg et al., 2016). Pipeline 프로젝트에서는 10개의 실험 중 6개가 반복검증 기준을 통과하였다. 반면에 자선단체에 기부하는 금액보다도 기부사실을 홍보하는데 더 많은 돈을 쓰는 회사는 전혀 기부를 하지 않는 회사보다 오히려 부정적인 평가를 받는다는 도덕적 전도 효과(moral inversion)의 경우 반복검증 되기는 하였으나 원

연구보다 효과 크기가 작았다. 또한 지폐로 적은 팁을 남기는 손님보다도 더 많은 금액이기는 하나 팁 전체를 동전으로만 남기는 손님을 더 부정적으로 평가한다는 나쁜 손님 효과(bad tipper effect)는 원 문화권 내에서는 일관되게 재검증되었지만 타 문화권에서는 반복검증되지 않았으며, 나머지 2개의 연구는 반복검증에 실패하였다.

종합해 보면 반복검증 비율은 프로젝트마다 상당히 다르게 나타났다(Laws, 2016; Stroebe, 2016) - Many Labs 프로젝트(77%), Pipeline 프로젝트(60%), Reproducibility 프로젝트(36%), Registered 프로젝트(23%)[4]. 프로젝트마다 다른 반복검증 비율이 나타난 원인이 무엇인가는 몇 가지 설명이 가능해 보인다(Laws, 2016). 전반적으로는 하나의 연구 그룹이 각 연구를 개별적으로 1회씩 반복검증한 2개 프로젝트(Registered와 Reproducibility 프로젝트)보다 여러 연구 그룹에서 동일한 연구들을 여러 차례 반복검증하여 그 효과를 각 연구에 대한 전반적 효과 크기로 계산한 2개 프로젝트(Many Labs와 Pipeline 프로젝트)에서 재현성이 높게 나타났다. 다음으로 어떠한 연구들이 반복검증 후보로 선택되었는지도 성공률에 영향을 미친 것으로 생각된다. 일례로 Many Labs 프로젝트에서는 이미 재검증된 적이 있는 연구들이 일부 포함되어 있었다.

4) 현재까지 진행된 4개의 대규모 반복검증 프로젝트 결과에 대한 종합적인 개관은 Laws(2016)와 Stroebe(2016)를 참고할 수 있다. 다만 두 논문에서 반복검증에 성공하였다고 판단한 연구의 개수에서 약간의 차이가 있는데, 이에 대해서는 원 연구에 대해 약한 지지(weak support)를 보인 결과를 반복검증에 성공한 것으로 포함시킬 수 있는지의 여부에 대해 학자들 간의 판단이 다르기 때문인 것으로 생각된다.

4개의 대규모 반복검증 프로젝트는 중요한 공헌을 하였지만 근본적인 한계점도 남겼는데, 반복검증 대상이 된 대부분의 연구가 경제적인 기준을 고려하여 선정됨으로써 고비용의 연구들을 포함시키지 못하였다는 점이다. 예를 들어 오랜 기간을 필요로 하는 연구, 고가의 장비를 사용해야 하는 연구(예: fMRI, 안구추적), 역사적 사건에 의존적인 연구, 접근하기 힘든 표본(예: 원숭이, 자폐증) 등을 활용한 연구는 반복검증을 시도하는데 현실적인 어려움이 있었기 때문에 분석대상에서 제외되었다(Laws, 2016). 자료 수집이 간편하고 용이한 분야(대표적으로 사회심리학과 인지심리학)에만 영역 특정적으로 집중하는 것은 중요한 발견에 대한 진위여부를 확인할 기회를 제한한다는 점에서 또 다른 문제가 될 수 있다. 반복검증 연구를 선택적으로 시행하는 것은 연구결과의 재현성에 대한 전체적인 통찰을 저해한다는 점에서 향후 논의가 필요한 부분이다.

반복검증의 절차와 방법에 대한 논쟁

야심차게 추진되었던 대규모 반복검증 프로젝트 내의 상반된 결과들은 연구를 반복검증하는 절차와 방법에 대해 심리학자들 간에 상당한 논쟁을 야기하였다. 먼저 자신의 연구가 재현되지 않았던 Cambridge 대학의 Simone Schnall(2014)은 Registered 프로젝트의 절차상의 문제점을 지적하고 그로 인해 자신이 받게 된 피해를 토로하였다. 흥미로운 연구를 하는 유능한 연구자가 연구결과가 재현되지 않았다는 이유로 일순간에 부도덕한 연구자로 의심을 받고, 연구비를 지원받지 못하는 심각한 부작용이 나타났다는 것이다. Harvard 대학의 Daniel Gilbert(2014) 또한 연구결과가 재현되지

않았다는 이유로 원 저자들이 비난을 받게 되는 분위기를 집단 괴롭힘(bullying)이라며 신랄하게 비판하였다. 반복검증 방식에 반대하는 학자들은 이상적인 연구를 이상적으로 반복검증하는 시도조차도 충분히 실패할 여지가 있기 때문에 원 논문에서 보고된 결과가 재현되지 않았다는 것이 필연적으로 원 연구가 결점이 있거나 타당성이 부족하다는 것을 입증하지는 않는다고 주장한다(Maxwell, Lau, & Howard, 2015; Stroebe, 2016; Stroebe & Strack, 2014). 원 논문의 결과가 재현되지 않는 데에는 많은 요인들이 관여될 수 있으므로, 반복검증에 실패했다는 사실을 의문에 대한 최종적인 답을 의미하는 것으로 받아들여서는 안 되며, 더 많은 의문을 제기하는 것으로 이해해야 한다는 입장인 것이다(Law, 2016).

한편 논쟁이 격화되는 것에 대한 대안으로 노벨상을 수상한 심리학자 Daniel Kahneman (2014)은 원 저자와의 논의 없이 일방적으로 진행되는 반복검증 연구는 제한할 필요가 있다고 제안하였다. 그는 인간의 행동이 일견 관련이 없을 것 같은 요인들(예: 검사지의 글자 모양, 검사시행 요일 등)에 의해서도 쉽게 영향을 받을 수 있기 때문에, 논문의 방법론 섹션에 기술된 내용만으로는 다른 연구자에게 정확한 처방을 제공하기에는 모호함이 있다고 주장하였다. 또한 연구를 수행하는데 있어 원 저자와 재검 연구자가 연구에 투입하는 노력(예: 시간 투자, 연구비 지원 등)의 규모와 동기 자체가 다르며, 과연 누가 평가할 것인가의 문제도 존재한다(Bissell, 2013). 그렇기 때문에 Cesario(2014)는 반복검증의 중요성은 인정하지만 연구결과의 신뢰성을 확인하는 최선의 방법은 다른 연구자들에서가 아닌 원 저자들에 의한 자기검증의 차원에서 이루어져야 한다고 주장하였다. 이러한 입장에서는 다른 연구자들에 의해 무차별적으로 이루어지는 반복검증 방식이 재현성 위기를 근본적으로 해결하기보다는 오히려 예기치 않은 부작용을 가져올 수도 있음을 지적한다(Bissell, 2013).

그러나 이와 대척점에 있는 상당수의 학자들은 반복검증의 가치와 필요성을 지지하고 있는데, 가장 직접적인 이유는 반복검증이야말로 과학을 지탱하는 가장 중요한 초석이라고 생각하기 때문이다(Brandt et al., 2014; Crocker & Cooper, 2011; Jasny, Chin, Chong, & Vignieri, 2011; Klein et al., 2014; Nosek & Lakens, 2014). 먼저 반복검증은 경험적인 발견의 일반화 가능성과 더불어 실제 존재하는 효과의 크기를 정확하게 추정할 수 있도록 해준다(APS, 2013; Asendorpf et al., 2013; Brandt et al., 2014; Klein et al., 2014; Makel & Plucker, 2014; Nosek & Lakens, 2014; Simons, 2014). 연구 과정에서의 오류는 필연적이며, 선택적으로 정보를 탐색하는 확증편향도 자신의 가설을 지지하는 결과가 나올 경우 기저 오류를 적극적으로 탐색하지 않게끔 하는데 기여할 수 있기 때문이다(Bollen et al., 2015; Simmons, Nelson, & Simonsohn, 2011). 이에 연구결과를 일반화하고 일련의 주장을 경험 과학 내의 사실로 증명해 가는데 있어 반복검증이 중요한 수단이 된다는 점은 일찍부터 여러 학자들에 의하여 주장되어 왔다(Cohen, 1994; Schmidt, 2009).

또한 반복검증 연구는 과학계에서 발생할 수 있는 사기나 데이터 조작과 같은 연구부정행위를 저지할 수 있는 1차 방어선과 같은 기능도 할 수 있다(Crocker & Cooper, 2011). 만약 특정 결과에 대해 다양한 연구자들이 수행한 연구에서 원 논문의 결과가 일관되게 재현되

지 않는다는 보고가 축적되면, 이는 원 결과의 신뢰성을 의심해 볼 수 있는 경고신호로 간주할 수 있다. 앞서 언급된 Stapel의 사례에서도 알 수 있듯이 논문에 대한 심사과정만으로는 부정행위를 적발해 내기 어려우며, 그렇기 때문에 일부 학자들은 반복검증 연구만이 심리과학을 비롯한 과학 전반에서 나타날 수 있는 연구부정을 차단할 수 있는 유일한 수단이라고 주장한다(Crocker & Cooper, 2011; Makel & Plucker, 2014; Roediger, 2012; Simons, 2014). 다른 연구자들도 반복검증 연구가 기적의 치료제는 아니지만 연구결과의 진실성과 재현성에 대한 우려를 최소화하는데 중요한 도움을 줄 수 있다는 점에 동의하고 있다(Makel, Plucker, & Hegarty, 2012).

심리학적 연구 재현성 위기의 원인

그렇다면 왜 일부 또는 상당수 심리학적 연구들은 연구결과가 반복적으로 재현되지 않는가? 연구 재현성 위기가 초래되는 중요한 원인으로 심리학적 연구방법론 및 연구관행 상의 본질적인 한계점과 더불어 출판 과정 내에 존재하는 구조적인 문제들이 지적되어 왔다. 이는 의도적으로 결과를 조작하는 비윤리성과는 구분되어야 할 측면으로서, 다수의 저자들은 원 연구결과가 재현되지 않더라도 이것이 원 연구자들의 도덕적 진실성을 의심할 직접적인 단서로 해석될 수는 없다는 점을 지적해 왔다(Francis, 2012b; Laws, 2016; Open Science Collaboration, 2015). 이러한 측면들은 현재 관행 내에 존재하는 한계점 및 맹점들로 인하여 발생하는 것으로 간주하는 것이 보다 타당하며, 그렇기에 앞으로 변화를 통한 개선 노력이 요구되는 측면들이라고 할 수 있겠다. 다음에서는 재현성 위기의 원인을 크게 원 연구결과의 도출 과정의 문제(영가설검증의 한계, 연구관행의 문제)와 학계 내의 구조적 문제(출판편향)로 구분하여 고찰하도록 하겠다.

영가설검증의 한계

우선 현행 통계적 추론절차의 한계로 인하여 심리학적 연구결과들의 재현성이 저하된다는 문제 제기가 오래 전부터 이어져 왔다(Cohen, 1994; Cumming, 2014; Krueger, 2001). 전통적으로 심리학을 비롯한 사회과학 분야에서는 빈도주의 전통에 입각한 영가설검증(NHST; null hypothesis statistical testing)의 논리에 근거해서 연구의 결론을 도출해 왔다. 영가설검증의 논리를 간략히 설명하면 다음과 같다. 연구자는 자신이 증명하고자 하는 연구가설과는 반대로 '두 집단 간에 차이가 없다'거나, '두 변수 간에 관련성이 없다'와 같이 연구자가 애초에 입증하기를 원하지 않는 가설, 즉 영가설을 설정한다. 다음으로는 그렇게 설정한 영가설이 참이라는 가정 하에서 반복표집 시 실제로 관찰된 자료처럼 극단적인 자료가 관찰될 확률을 계산한다. 이 확률이 바로 p 값(value)이다. 이렇게 계산된 확률이 낮을수록 영가설이 거짓일 가능성은 더 높아지므로 영가설을 기각하고, 연구가설을 채택할 수 있다는 것이 가설검증의 논리이다(Nuzzo, 2014). 여기에서 '확률이 아주 낮다'라는 기술은 임의적인 것으로 통계학적으로 이 기준을 분명하게 규정한 이론은 없다. 다만 관습적으로 .05 수준을 알파 값(α)이라고 하며, 여러 과학 영역에서 낮은 확률로 받아들이고 있다(김청택, 2011).

영가설검증은 심리학의 역사에서 가장 광범위하게 사용되어온 방법이지만(Krueger, 2001), 문제는 그 결과가 자주 오해석 및 오용되고 있다는 점이다(Greenland et al., 2016; Wasserstein & Lazar, 2016). 영가설검증에서 대표적으로 오해석 되는 것이 바로 p 값의 의미이다. 많은 연구자들에게 p 값은 연구결과의 강도를 의미하는 지표로 사용되는 경향이 있으며(Nuzzo, 2014), 연구결과를 채택하는 결정적인 기준으로 받아들여지고 있다(American Statistical Association[ASA], 2016; Nuzzo, 2014; Wasserstein & Lazar, 2016). 일례로 많은 과학 논문에서 p 값이 .001인 경우 '매우 유의'하다고 강조하여 해석하는 경우가 대표적이다(김청택, 2011). 하지만 p 값은 가설 확률이 아니며, 특히 연구자가 궁극적으로 알고 싶어 하는 연구가설에 대해서는 답을 해 줄 수 없다(김청택, 2011; Greenland et al., 2016; Head, Holman, Lanfear, Kahn, & Jennions, 2015). 여기서 p 값이 의미하는 확률이 일종의 빈도 확률(frequency probability)이라는 점을 강조할 필요가 있다(김청택, 2013; Greenland et al., 2016). 즉, p 값은 무수히 많은 시행의 반복표집을 가정할 때 "통계적 모델과 그 가정들이 참이라면 관찰되길 기대하는 바와 주어진 자료가 얼마나 양립 가능(compatible)한지를 나타내는 연속적인(continuous) 요약치"이다(Greenland et al., 2016). 따라서 작은 p 값은 (영가설을 포함한 통계적 모델의 모든 가정이 참이라면) 반복표집의 연속에서 주어진 자료가 관찰되는 일이 드물다는 것을 의미할 뿐 영가설 또는 연구가설 자체가 참인지에 대해서는 직접적으로 알려주지 않으며, .05를 기준으로 이분법적 해석에 의존하는 것은 임의적이라 할 수 있겠다(박준석, 2015; Greenland et al., 2016). 만약 연구자가 단지 p 값이 유의하다는 것에만 근거하여 결과를 해석하고 유의한 결과를 선택적으로 보고한다면, 그럴 듯하지만 거짓된 과학적 지식을 양산하는 결과를 가져올 수 있다. 더욱이 영가설검증에서의 p 값은 표본크기에 민감한 특성을 지니고 있다(김청택, 2011). 즉, 표본의 크기가 증가할수록 p 값은 작아지게 되어 있기 때문에 표본의 수를 대폭 증가시켜서 자료를 수집할 경우 실제로는 매우 미미한 차이도 통계적으로는 유의한 결과를 만들어 낼 수 있으며, 반대로 표본 크기가 작은 경우 중요한 효과조차도 유의하지 않은 것으로 나타날 우려가 있다(Levine, Weber, Hullett, Park, & Lindsey, 2008; Wasserstein & Lazar, 2016).

이러한 이유로 이미 오래 전부터 많은 연구자들은 영가설검증이 통계적인 추론방법으로 적합하지 않다고 주장하였으며(Balluerka, Gómez, & Hidalgo, 2005; Schmidt, 1996), 일부 학자들은 영가설검증의 기계적이면서 반복적인 절차를 '영가설 의식(null ritual)'이라는 표현으로 비판하기도 하였다(Gigerenzer, 2004; Gigerenzer, Krauss, & Vitouch, 2004). 특히 알파 값이라는 임의적인 절단점을 기준으로 '통계적 유의성'의 이분법적 판단을 하는 것의 오류와 잠재적 해악에 대하여 많은 연구자들이 비판해 왔으며, 작금의 재현성 위기의 맥락에서 영가설검증에 내재되어 있는 문제들이 과학적 연구의 신뢰도와 재현성을 저해하고 있다는 점이 공통적으로 지적되어 왔다(박준석, 2015; Cumming, 2014; Nuzzo, 2014; Wasserstein & Lazar, 2016). 이에 최근 미국통계학회(ASA)에서는 재현성 위기에 대한 심각한 우려를 표명하며, p 값에 대한 올바른 해석 지침을 제시한 바 있다 (표 1). 아울러 이 공식 성명의 마지막 문장이 "어떠한 단일 지표도 과학적 추

표 1. 통계적 유의성과 p 값에 대한 미국통계학회(ASA)의 공식 성명 (ASA, 2016에서 발췌 인용)

(1) p 값은 특정 통계적 모델과 자료가 얼마나 불합치한지를 알려줄 수 있다.
(2) p 값은 연구가설이 참일 확률 또는 자료가 우연에 의해 산출되었을 확률을 측정하는 것이 아니다.
(3) 과학적 결론과 산업 및 정책적 결정은 p 값이 특정 역치를 넘었는지 여부에만 근거해서는 안 된다.
(4) 타당한 통계적 추론은 완전한[5] 보고와 투명성을 요한다.
(5) p 값 또는 통계적 유의성은 결과의 효과 크기나 중요성을 측정하는 것이 아니다.
(6) 그 자체로 p 값은 모델이나 가설에 대한 좋은 측정치가 되지 못한다.

론(scientific reasoning)을 대체할 수는 없다"라는 점은 자못 의미심장하다.

연구관행의 문제

영가설검증과 더불어 심리학자들이 연구를 수행하고 자료를 분석하는 과정에서 결과의 타당도를 저해하는 요인으로 비판받아온 것이 '의심스러운 연구관행(questionable research practices: 이하 QRP)'이다(Lindsay, 2015; Open Science Collaboration, 2015; Pashler & Wagenmakers, 2012; Simmons et al., 2011). QRP는 가설을 지지할 증거를 발견할 목적으로 사후적인 기준에 의해 자료수집이나 분석을 결정하는 관행을 의미한다. QRP는 심리학계 내에 광범위하게 확산되어 있는 것으로 보인다. 2천명에 달하는 심리학자를 대상으로 한 대규모 연구(John, Loewenstein, & Prelec, 2012)에서 다수의 연구자들이 다음과 같은 연구관행을 적용한 경험이 있는 것으로 확인되었다 - 논문에서 종속변인 측정치를 제외하는 것, 결과의 유의성을 기준으로 자료의 추가 수집 여부를 결정하는 것, 결과를 얻기 위해 요구되는 중요한 정보들을 충분히 명시하지 않는 것, 본래 찾고자 했던 결과가 나타나면 계획했던 자료의 수집을 중단하는 것, 자료가 결과에 미치는 영향을 본 후에 자료의 제외를 결정하는 것, 유의했던 결과만 논문에 선택적으로 보고하는 것, 논문에서 예기치 못한 발견을 하게 되면 처음부터 예상을 했던 것처럼 보고하는 것 등.

문제는 이러한 관행들이 영가설이 참임에도 불구하고 영가설을 기각하는 1종 오류인 위양성(false-positive) 결과의 발생비율을 급격하게 증가시킨다는데 있다. Simmons 등(2011)은 일련의 시뮬레이션을 통하여 특정 가설과 일치하는 증거를 찾는 것이 얼마나 손쉬운지를 예증하였다. 이들의 분석에 따르면 현재 위양성의 허용 기준은 최대 5%(p<.05) 내로 설정되어 있지만, 자료수집 및 분석과정에서 '연구자의 자유도(researcher degree of freedom)'로 지칭되는 융통성을 발휘하는 경우 가설이 잘못되었음에도 불구하고 통계적으로 유의한 증거를 손쉽게 찾아낼 수 있었다. 만약 이러한 주장이 사실이라면 심리학 학술지에 출판된 상당수의 논문이 QRP에 의한 위양성 결과를 내포하고 있을 가능성을 시사하는 것이다.

이러한 관행 중에서도 특히 통계적으로 유의한 결과를 얻을 때까지 자료를 수집 또는 제외하거나 사전에 계획되지 않은 여러 분석을 무분별하게 실시하는 p-hacking이 문제시되고 있다(Simonsohn, Nelson, & Simmons, 2014).

5) '완전한(full)' 보고는 유의한 결과만을 선택적으로 보고하는 것과 반대되는 의미이다.

P-hacking이란 실제 효과보다 연구결과를 부풀려서 보고하는 일종의 인플레이션 편향이다(Head et al., 2015). 연구자들이 의도적으로 *p*-hacking을 하는 경우, p 값은 통계적 유의성 기준인 .05 보다 조금 낮은 값에 집중되기 때문에 기존 문헌들에서 얻은 p 값의 분포를 나타내는 *p*-curve의 모양은 .05의 바로 아래에서 기대되는 빈도에 비해 p 값이 과잉발생(overabundance)하는 부적 편포된 형태를 보이게 된다(Head et al., 2015). 실제로 연구자들은 일련의 분석을 통해 학술지에 게재된 상당수 논문들의 p 값이 .05 주위에 모여 있다는 사실을 발견하였다(Head et al., 2015; Simonsohn et al., 2014). 이처럼 QRP나 *p*-hacking과 같은 잘못된 연구관행을 통해 통계적으로 유의한 발견만을 선별적으로 보고하는 것은 학술지에 신뢰할 수 없는 결과가 양산되는 문제를 초래한다(Wasserstein & Lazar, 2016).

그렇다면 연구자들은 왜 QRP 또는 *p*-hacking을 하게 되는가? Simmons 등(2011)은 QRP를 하는 이유가 악의적이라기보다는 연구자들이 연구수행 과정에서 직면하게 되는 모호함과 유의한 결과를 찾고자 하는 욕구 때문이라고 해석하였다. 연구자들은 자료를 분석하는 과정에서 수많은 결정과 판단에 직면해야 한다. 예를 들어 반응시간을 측정하는 실험을 했다면, 어느 정도로 빠른 반응을 극단치(outlier)로 간주해서 제거해야 하는가? 상위 2.5%, 평균으로부터 2 표준편차 이상, 100ms 이내, 150ms 이내, 300ms 이내? 이처럼 '빠르다는 것'을 정의하는 기준은 매우 다양하기 때문에 모호함이 있는 분석적 결정을 내릴 때 연구자들은 자신의 가설을 정당화 해주는 방향으로 결정을 내리기 쉽다. 일부 강경한 학자들은 노골적으로 드러나는 사기보다 QRP가 학문의 발전에 악영향을 미칠 수도 있음을 지적하였다. QRP나 *p*-hacking을 이용해서 인위적으로 유의한 결과를 만들어내는 것은 견고하지 못한, 그러므로 결과적으로는 재현될 수 없는 결과가 학술지에 지속적으로 출판되는 결과를 초래하는데(Simonsohn et al., 2014), 이는 다음에 언급할 출판편향(publication bias) 문제와도 연결된다.

학계 내의 구조적 문제 : 출판편향

출판 과정에서 존재하는 출판편향이 연구 재현성 위기를 초래하는 데 기여했다는 분석이 있어 왔다. 학술지는 과학적인 아이디어와 연구방법, 그리고 경험적인 자료에 대해 연구자들이 의사소통하는 주요한 수단인데, 현재 학계에서는 가설에 부합하는 유의한 연구, 새롭고 참신한 연구만을 받아들이는 경향이 지배적이다(APS, 2013; Francis, 2012a; Ginger-Sorolla, 2012; Neuliep & Crandall, 1990, 1993; Nosek & Lakens, 2014; Open Science Collaboration, 2012, 2015; Yong, 2012). 이는 상당수의 학술지가 '연구의 독창성'이나 '기존 연구와의 차별성' 등의 평가항목을 투고논문의 주요 심사기준으로 포함한 것에서도 드러난다. 기존의 발견을 재확인하는 연구는 논문심사 시 정량적인 지표에서 상대적으로 낮은 점수를 받도록 심사기준이 설정되어 있다. 그렇기 때문에 학술지에 출판되는 논문은 새로운 발견이나 기존 연구에 비해 어떤 점에서든 차이가 있거나 차별화된 결과들이 주를 이룬다.

사실 이러한 문제는 오늘날 갑작스럽게 대두된 것은 아니다. Rosenthal(1979)은 이미 오래 전에 이러한 문제를 서랍함 현상(file-drawer problem)이란 용어를 통해 설명한 바 있다. 출

판편향의 해로운 형태 중의 하나인 서랍함 현상은 유의한 결과가 나온 논문만이 학술지에 게재되고, 유의하지 않은 것으로 나타난 연구결과는 게재되지 못하고 연구자의 서랍함에 들어가 공개되지 않는 현상을 의미한다. 즉, 영가설이 참으로 나타난 부적 결과는 학술지에 출판되지 않고, 영가설을 기각하는 정적인 결과만이 출판되는 것이다(Franco, Malhotra, & Simonovits, 2014).

문제는 출판편향으로 인해 출간된 연구결과물의 총화는 실제 현상을 올바르게 대표하지 못하기 때문에, 학문적인 차원을 넘어 사회적으로도 심각한 문제를 야기할 수 있다는 점이다(Pashler & Harris, 2012). 학계에서 출판편향은 광범위한 현상이다. 심리학뿐만 아니라 의학 연구들에서도 부적 결과들은 잘 보고되지 않으며, 일례로 미국 식품의약청에 등록된 항우울제 효과 연구 중에서 38개의 정적 결과 중 37개가 출판된 것에 비해, 24개의 부적 결과들 중에서는 3개만이 출판되었다는 점을 지적한 메타분석이 있었다(Turner, Matthews, Linardatos, Tell, & Rosenthal, 2008). 또한 Begley와 Ellis(2012)는 신약 발전 프로그램을 설계하기 위해 전임상(preclinical) 암 연구 분야에서 출판된 기념비적 논문 53개의 결과를 재확인하는 과정에서 전체 논문 중 오직 6개의 논문(11%)만이 재현된다는 사실을 발견하였다. 심각한 것은 정적 결과만이 보고되고 부적 결과는 묻히는 상황에서 재현되지 않는 연구들이 이미 확산되어 있고, '잘못된 사실'들이 이후 단계의 결정들에 영향을 주고 있다는 점이다. 이는 관련 기업 및 산업 전반에 걸쳐 부정적 결과를 초래할 수 있으며, 이미 임상 연구로 진전된 경우에는 다수의 환자들이 근거가 충분하지 않은 약물의 임상 실험 대상자가 되었음을 의미하므로 연구의 윤리성과 관련해서도 심각한 문제를 야기한다(Begley & Ellis, 2012; Pashler & Harris, 2012).

더불어 최근 들어 출판편향 문제를 악화시키는 외부적 보상체계의 문제와 관련하여 학계에서 학문적 성취에 대한 경쟁이 과열되고 있는 점을 들 수 있다(Chambers, 2014; Pashler & Wagenmakers, 2012). 저명한 학술지에 논문을 많이 게재하는 것이 연구자 자신에게 돌아올 보상(학계에서의 명성과 지위, 연구비 등 재정적, 제도적 지원)의 크기를 결정하는 규칙이 되어가고 있다(Open Science Collaboration, 2012). 이러한 보상 시스템은 시장의 논리에 의해 조성되고 작동된다. 누구나 선망하는 좋은 자리보다 그러한 자리를 갈망하는 연구자의 수가 더 많고, 연구자들이 게재를 희망하는 권위 있는 학술지의 한정된 지면보다 더 많은 수의 논문이 양산되고 있다(Stroebe et al., 2012). 이처럼 치열한 출판 병목(publication bottleneck)에서 생존하기 위해 연구자들은 과학적인 엄격함을 추구하기 보다는, 단기간에 유의한 결과를 다수 출간하거나 때로 외관상 결점이 없어 보이는 심미적인 결론을 도출하는데 노력을 기울이기도 한다(Ginger-Sorolla, 2012). 새로운 발견에 가치를 부여하는 학계의 기준, 학문적인 성취에 대한 과잉경쟁, 그리고 이러한 현실에 적응하고 생존하기 위한 연구자들의 출구전략(예: QRP, *p*-hacking, HARKing[6] 등)이 맞물린 결과는 의도치 않게 출판편향의 문제로까지 확산될 수 있다.

이상에서 살펴본 바와 같이 심리학 연구

6) HARKing이란 'hypothesizing after the results are known'의 두문자어 표현으로, 결과를 보고나서 사후적으로 가설을 수립하는 연구관행을 의미한다(Kerr, 1998).

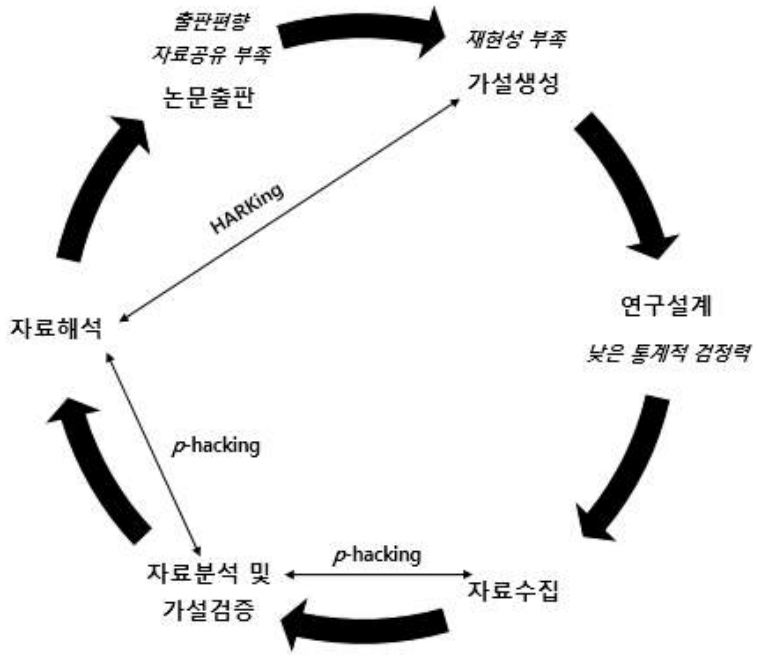

그림 2. QRP, 출판편향, 재현성 문제의 순환관계 (Chambers et al., 2014)

의 재현성 위기는 단일 요인에 의해서라기보다는 통계적 추론, 연구관행 상의 한계와 더불어 학계의 구조적인 문제를 반영하는 여러 요인들이 상호 복합적으로 연계되어서 발생하는 것으로 보인다. Chambers, Feredoes, Muthukumaraswamy와 Etchells(2014)은 의심스러운 연구관행, 출판편향, 낮은 재현성 문제의 상호연관성을 가설-연역적(hypothetico-deductive) 모형이라는 개념으로 설명하였는데, 그림 2에 도식화되어 있듯이 연구수행 과정에서 나타나는 연구관행이 유의한 결과의 출판에 영향을 미치고, 이러한 출판편향은 연쇄적으로 낮은 재현성 문제를 유발하게 된다. 또한 낮은 재현성은 새로운 연구의 통계적 검정력을 약화시키고, 이는 다시 유의한 결론에 도달하기 위한 의심스러운 연구관행의 사용을 '촉진'하는 악순환이 반복된다.

심리학적 연구의 재현성을 증진하기 위한 해결책들

지금까지 살펴보았듯이 2011년 이후로 주로 사회심리학 분야 내에서 심리학적 연구의 재현성 위기와 그 원인에 대한 대규모의 논쟁이 진행되어 왔으며, 대안과 해결책을 모색하는 논문들이 뒤이어 발표되고 있다. 그 대안과 해결책들은 사회심리학뿐만 아니라 심리학 내의 타 분과에도 유용하게 적용될 수 있는 것들로 생각된다. 이에 본고에서는 현재까지 여러 연구자들이 제안하였던 잠정적인 해결책들을 다음과 같이 정리하여 제시하고자 한다 - 반복검증에 대한 재평가, 신뢰구간과 효과 크기의 사용, 표본 크기의 증가, 메타분석의 활용, 베이지안 통계의 적용, 연구 방법론의 투명성 제고.

반복검증에 대한 재평가

반복검증은 우리가 이미 알고 있는 지식에 새롭게 더해 줄 것이 적으며, 따라서 연구로서의 참신함이나 유용성이 떨어진다는 견해가 심리학 내에 존재해 왔던 것이 사실이다(Laws, 2016). 이와 일관되게 심리학 내에서 반복검증 연구들의 출판 비율이 약 1%에 지나지 않는다는 보고가 있으며, 자연과학에 비하여 사회과학 학술지의 편집자와 심사자들은 반복검증보다는 새로운 결과를 선호하는 경향이 강하다는 설문 결과도 있다(Madden, Easley, & Dunn, 1995; Makel et al., 2012; Neuliep & Crandall, 1993). 즉, 반복검증 연구는 "획기적이지 않으며, 과학이라는 벽을 쌓을 때 그저 하나의 벽돌일 뿐"(Kail, 2012; *Psychological Science*의 편집자)이라는 회의주의는 반복검증 연구의 가치가 저평가 받고, 다른 연구자들과 공유할 수 있는 학술지 논문으로 출판되는 과정에서 배제되는 출판편향이 발생하는 데 기여하였던 것으로 생각된다.

그러나 더욱 견고하고 누적적인 심리과학을 위해서 반복검증은 필수적이며, 중요한 부분이라는 인식이 강화되어야 할 필요가 있다. 단 한 번의 연구에 의해 검증된 결과는 아직 우리가 찾고자 하는 사실이 아니다. 과학은 공동의 지식 창출 과정이며, 어떤 연구자에 의해 제기된 주장이 유의미한 것으로 여겨진다면 거듭된 실험 및 관찰을 통하여 잠정적인 사실의 범주에 들게 되는 승인(canonization)을 받는 검증 과정을 거치는 것이 바람직하다(Nissen et al., 2016). 일반적으로 성공적인 반복검증이 늘어날수록, 연구결과에 대한 확신은 증가한다(Stangor & Lemay, 2016). 연구의 결론을 내리는 데 있어 1종 오류의 가능성을 감소시키기 위해 하나의 논문에도 여러 개의 반복검증 연구를 포함하는 것이 필요하다는 주장이 있어 왔다(Murayama, Pekrun, & Fiedler, 2014). 실제로 반복검증의 횟수가 연구결과의 긍정 예측도(positive predictive value: 이하 PPV; 유의한 연구결과가 참일 확률)에 어떤 영향을 주는지를 시뮬레이션한 연구에서는 사전 확률(pre-study odds; 효과가 실재할 것으로 연구에 앞서 가정한 확률)이 .5이고 검정력이 80%일 때 1회의 유의한 결과의 PPV는 .55에 불과하였으나, 2번의 유의한 반복검증이 추가될 시 PPV는 .98로 증가하였다(Moonesinghe, Khoury, & Janssens, 2007). 이와 더불어 학술지 차원에서도 특정 이론이나 가설에 관한 수렴 증거를 찾는 과정으로서 어느 정도의 중복성(redundancy)이 필요하다는 생각을 수용할 필요가 있으며(Crotty, 2014), 유의하지 않은 부적 결과나 가설과 반대 방향으로 나타난 결과에 대한 적극적인 설명을 장려해야 한다.

이와 관련된 한 가지 이슈가 과연 어떤 방식으로 반복검증을 해 나가는 것이 유용한가 하는 것이다. 반복검증에 대한 보다 세분화된 분류도 존재할 수 있겠으나(예: Hüffmeier, Mazei, & Schultze, 2016), 지금까지 이와 관련된 대부분의 논의는 직접(direct or exact) 대 개념적(conceptual) 반복검증의 축을 따라 이루어져 왔다(Crandall & Sherman, 2016; Crocker & Cooper, 2011; Nosek, Spies, & Motyl, 2012; Pashler & Harris, 2012; Schmidt, 2009; Simons, 2014; Stroebe & Strack, 2014). 직접 반복검증은 원 연구에서 사용된 자극과 절차 등을 가급적 모두 동일하게 재현하여 동일한 결과가 산출되는지 확인하는 것으로, Moonesinghe 등(2007)은 모든 진정한 반복검증은 직접 반복검증이어야 한다고 주장하였다. 현재까지 진행된 4

개의 대규모 반복검증 프로젝트들도 직접 반복검증에 속한다고 볼 수 있다. 이와 달리 개념적 반복검증은 원 연구와 동일한 개념이나 가설을 다른 방법론에 의거하여 반복검증하는 것이며, 초점은 원 연구에서 사용된 절차의 신뢰도보다는 기저의 이론적 예측의 타당성에 있다(Crandall & Sherman, 2016).

결국 어떤 방식의 반복검증이 유용한지는 목적이 무엇인가에 따라 달라질 것이다. 원 연구결과가 위양성인지를 직접적으로 확인하고자 할 때에는 직접 검증이 필요하겠지만, 이론의 확증과 반증을 통한 누적적인 과학적 발전의 측면에서는 개념적 반복검증이 유용한 정보들을 제공할 수 있을 것으로 생각된다. 더불어 심리학에서 연구 대상으로 삼는 주제들은 특정 맥락에 의존하는 현상일 가능성이 높으며, 동일한 절차를 사용한다고 해도 "우리는 같은 강물을 두 번 건널 수 없다"(Crandall & Sherman, 2016, p. 94)[7]는 점을 고려할 필요가 있어 보인다. 어떤 면에서 모든 반복검증은 개념적이며, 직접 대 개념적 반복검증의 구분은 범주적이라기보다 차원적일 수도 있다(Crandall & Sherman, 2016; McGrath, 1981; Stroebe & Strack, 2014). 즉, 이들은 개념적 반복검증의 가치에 방점을 두면서, 직접 대 개념적 반복검증이라는 방법이 상호배타적인 양자 선택의 문제가 아닐 수도 있다는 점을 지적하였다.

동시에 반복검증이 우리가 수행하는 심리학적 연구의 결과가 갖는 타당도나 진실성에 대한 논란에 최종적인 해결책을 제공해 주는 것은 아니라는 점을 분명히 인식할 필요가 있다(Coyne, 2016; Stangor & Lemay, 2016; Stroebe, 2016). 앞서 기술하였듯이 성공적인 반복검증 횟수가 늘어날수록, 연구결과에 대한 확신은 증가할 수 있다(Moonesinghe et al., 2007). 그러나 그 효과가 실재하는 것인지를 확인하기 위해서 반복검증은 유용하지만 불완전한 수단이다(Stangor & Lemay, 2016). 만약 원 연구의 반복검증에 성공하였다고 하자. 그렇다면 과연 몇 번의 반복검증이 그 연구결과를 타당한 것으로 받아들이기에 충분한가? 그리고 반복검증 연구조차 p 값에 근거한 선택적인 보고와 출판편향, 작은 표본 크기와 낮은 검정력 문제로부터 자유롭다고 간주할 수 있는가? 반대로 원 연구의 반복검증에 실패할지라도 이것이 원 효과가 거짓이었다는 결론을 내리게 해주는 것은 아니다(Open Science Collaboration, 2012). 이에 대한 절대적인 답은 존재하지 않는 것으로 보인다. 대신 전향적으로 우리가 과학을 하는 방식, 즉 방법론의 견고함과 데이터의 투명성을 높이는 것을 포함한 근본적인 변화를 시도해야 할 필요성이 있다.

신뢰구간과 효과 크기의 사용

이미 앞서 기술한 바와 같이 전통적으로 심리학에서 사용되어 온 영가설검증과 p 값에 근거한 추론이 갖는 많은 한계점들이 노출되어 왔다. 따라서 영가설 유의성 검증을 보완하는 대안으로서 신뢰구간과 효과 크기를 추

7) 이와 관련하여 원 연구에서 사용된 자료와 절차를 사용하여 직접 검증을 시도하더라도, 그것이 동일한 이론적 변인을 반영하지 않을 가능성이 있다는 주장이 있다(Klein et al., 2014; Stroebe, 2016). 일례로 Many Labs 프로젝트에서 반복검증에 실패한 성조기 효과의 경우에 원 연구가 시행되었던 2009년과 반복검증 연구가 실시된 5년 후의 사회문화적 상황에서의 차이로 인하여 동일한 실험 절차가 동일한 효과를 나타내지 않을 수 있다는 것이다.

가적으로 보고하는 것을 장려하고 제도화하려는 노력이 필요할 것이다. 신뢰구간은 점 추정에 비하여 추정 오차에 대한 정보도 함께 제공하는 구간 추정이라는 장점이 있으며, 추정 오차가 크면 구간의 범위가 넓어지고 추정 오차가 작으면 좁아진다. 신뢰구간은 추정 오차와 모집단의 모수(parameter)에 대한 정보를 제공함으로써, p 값에 근거한 영가설검증의 한계점을 보완해줄 수 있다(김청택, 2011; Cumming & Finch, 2005). 만약 p 값이 유의하다고 해도 신뢰구간이 너무 넓거나 실험조건 간 상당 부분 중첩된다면, 이것이 통계적으로는 유의하더라도 실제적인 함의를 가지는 유의미한 차이인지 의문을 가져볼 필요가 있다.

또한 영가설검증의 문제점을 보완하기 위한 대안으로 효과 크기를 사용해야 한다는 것도 여러 연구자에 의해 제기되어온 주장이다(Cohen, 1988, 1994; Cumming, 2012; Ellis, 2010; Francis, 2012b; Funder et al., 2014; Lakens, 2013; Wasserstein & Lazar, 2016). 연구자가 p 값을 통해 알 수 있는 것은 영가설검증의 기각 여부뿐이며, p 값이 작을수록 어떤 처치나 조작의 효과가 더 강하다고 결론내릴 수는 없다. 그에 비해 효과 크기는 독립변인이 종속변인에 미치는 영향력의 크기를 해석할 수 있도록 표준화한 지표로, 효과가 얼마나 큰지를 알려준다(Ellis, 2010; Lakens, 2013). 사용하는 분석 방법에 따라 Cohen's d, Cohen's f, partial η^2 등 여러 가지 종류의 효과 크기 지표들이 존재하며(Ellis, 2010), 일례로 집단 평균비교 시 자주 사용되는 Cohen's d의 경우에는 일반적으로 .20을 작은, .50을 중간의, .80을 큰 효과 크기로 해석한다(Cohen, 1988, p. 40). 만약 효과 크기가 미미하다면, p 값 상 통계적으로 유의한 결과라고 해도 그 현실적 함의는 크지 않을 것이다. 효과 크기는 이와 같은 부가적인 정보를 포함하는 장점이 있으므로 국내에서도 더 많은 학술지들이 주요 결과에 대한 효과 크기의 보고를 장려하거나 의무화할 필요가 있으며(Hyde, 2001), 연구 재현성 위기의 맥락에서도 역시 중요한 보완책이 될 수 있다.

표본 크기의 증가

사회심리학뿐만 아니라 심리학 전반에서 연구결과의 타당도에 심각한 위협이 되는 고질적인 문제로 지적되어 온 것이 바로 작은 표본 크기와 낮은 통계적 검정력(statistical power)이다(Cohen, 1962; Finkel, Eastwick, & Reis, 2015; Fraley & Vazire, 2014; Schaller, 2016; Sedlmeier & Gigerenzer, 1989). 통계적 검정력은 영가설이 실제 거짓일 때 통계적 검정에서 영가설을 기각하는 능력(1-β, 즉, 위음성 결과를 산출하는 2종 오류(β)를 범하지 않는 것; Button et al., 2013)과 관련되며, 일반적으로 효과 크기와 표본 크기에 달려있다. Fraley와 Vazire(2014)는 검정력이 낮은 연구는 실재하는 효과를 적절히 탐지해낼 수 없고, 학계에 위양성 결과들의 비율이 증가하게 하며 따라서 재현 가능하지 않은 결과들을 양산할 수 있음을 지적하였다. 이러한 문제의식에서 시작하여 이들은 사회 및 성격심리학 분야의 대표적인 학술지 6개에서 표본 크기와 검정력을 조사하였는데, 그 결과 평균 표본 크기는 104이고 중간 크기(r=.2, d=.4)의 효과를 찾아낼 검정력은 약 50%였다. 이는 관습적으로 권고되는 검정력의 기준(80%; Cohen, 1988)보다 낮은 것으로, 연구가설이 실제 참인 경우 영가설을 기각하고 연구가설을 채택할 확률이 50%에 불과하다는 의미로 해석될 수 있다(Maxwell et al., 2015).

사회, 성격 심리학 외에 상담 및 임상심리학(Rossi, 1990)이나 인접 학문인 신경과학 분야(Button et al., 2013)도 작은 표본 크기와 낮은 검정력의 이슈를 공유하고 있는 것으로 보인다.

만약 예상되는 효과 크기에 비해 표본 크기가 너무 작다면, 연구의 시작부터 결과에 대한 잘못된 결론을 이끌어낼 위험성을 안고 있는 것이다(Stroebe, 2016). 대부분의 심리학적 연구에서 다루는 효과들은 작거나 중간 정도의 효과 크기를 갖기 때문에, 표본 크기가 작은 경우 실제로 존재하는 효과를 놓치는 2종 오류를 범할 수 있고 이는 기저 이론에 대한 혼란을 가져올 수 있다(Stangor & Lemay, 2016). 따라서 연구자의 현실적 한계 내에서 충분히 큰 표본을 사용하여 연구를 진행해야 하며, 발견하고자 하는 효과 크기에 견주어 적절한 표본 크기를 설정하기 위한 검정력 분석 등의 방법들(예: G*Power; Faul, Erdfelder, Lang, & Buchner, 2007)을 부가적으로 활용할 필요가 있다(Ellis, 2010).

메타분석의 활용

연구 재현성 위기에 대한 대안으로 메타분석을 활용할 것을 주장한 연구자들이 있다(Bargh, 2012; Braver, Thoemmes, & Rosenthal, 2014; Cumming, 2014; Stroebe, 2016). 이들은 낮은 검정력을 가진 반복검증 연구가 갖는 문제로 새로운 연구 역시 납득할만한 유의도 수준에 도달하지 못할 가능성이 있는데, 이러한 경우 그 연구결과가 폐기되고 출판편향이 나타나게 되므로, 대안적으로 연구의 신뢰성과 타당성을 확립하는 데 있어 메타분석이 도움을 줄 것이라는 낙관적인 전망을 제시하였다. 일례로 Braver 등(2014)은 '지속적인 메타분석의 누적(continuously cumulating meta-analysis)'이라는 용어로, 연구자들이 반복검증을 하나씩 할 때마다 연속적인 방식으로 메타분석을 실시할 것을 권고했다(Stroebe, 2016에서 재인용). 여기서 합산된 수치는 개별 연구 자료에 근거한 것보다 더 신뢰로울 것이라는 가정에서이다.

그러나 동시에 재현성 위기의 해결책으로 메타분석에 대한 회의적인 의견을 개진한 연구자들도 있다(Feinstein, 1995; Laws, 2016). 메타분석은 이미 출판된 연구들에서 패턴을 찾아내는 통계적 기법이므로, 메타분석을 가설을 확증하거나 연구의 재현성을 확인하기 위한 용도로 사용하는 데에는 제한점이 있다는 것이다(Hyman, 2010: Laws, 2016에서 재인용). 유의한 효과를 만들어내기 위하여 합쳐진 메타분석 결과는 신뢰롭지 않으며, 선택적으로 메타분석을 실시하면 이 또한 다른 형태의 부적합한 연구관행인 '메타 해킹(meta-hacking)'이 될 우려가 있다(Egger, Smith, Schneider, & Minder, 1997; Sakaluk, 2016). 현재까지의 중론은 메타분석이 재현성 이슈에 있어 상당한 역할을 할 수 있으나, 일정한 제한점이 있기 때문에 메타분석을 통한 사후적 확인이 적절한 검정력을 갖춘 실험을 사용한 전향적 재검증의 대체물은 될 수 없다는 것으로 보인다(Taylor & Munafò, 2016). 또한 이러한 논란과 관련하여 메타분석 시 자료 공유와 투명성 제고를 위한 노력(Lakens, Hilgard, & Staaks, 2016), 메타분석 시 *p*-curve 분석[8]의 보완적인 병용(Simonsohn et

8) *P*-curve 분석은 기존 문헌들에서 얻어진 p 값의 분포를 통하여 증거적 가치(evidential value)를 찾는 방법으로, 정적으로 편포된(right-skewed) p 값의 분포는 해당 효과가 견고하다는 의미로 해석

al., 2014; Taylor & Munafò, 2016) 등의 제안점도 고려해볼만 한 것으로 생각된다.

한편 반복검증된 견고한 연구결과에 대한 강조가 큰 표본의 확증적 연구들을 선호하는 경향으로 이어져, 소규모의 창의적이고 탐색적인 연구들을 억제하게 되지 않을까 하는 우려도 표명된 바 있다(Baumeister, 2016). 과학의 중요한 두 가지 축이 탐색과 확증이며, 양자 간의 균형이 우리가 지향해야 할 바라는 점에 대해서는 이견의 여지가 별로 없을 것이다(Dovidio, 2016; Sakaluk, 2016). 이와 관련하여 Sakaluk(2016)은 '소규모 탐색, 대규모 확증(Exploring Small, Confirming Big)'이라는 2단계 연구 전략을 제안하였다. 이는 1단계에서 탐색적 모델을 개발하기 위하여 작은 표본에서 전통적인 영가설 유의성검증을 실시하고, 2단계에서는 대규모의 표본에서 추가 변인들을 포함하여 1단계의 탐색적 모형을 더 엄격하고 확증적으로 검증하는 것이다. 모든 연구들이 확증적 방식으로 이루어져야 할 필요는 없으며, 재현성 위기가 미래의 심리학을 누구나 이미 알만한 안전한 가설들만을 검증하는 진부한 심리학('Bubba' psychology; Kelley, 1992)으로 이끌어 가서는 안 될 것이다(Stroebe, 2016). 이론을 검증 또는 반증하기 위한 반직관적이고 확률이 낮은 가설(risky prediction; Popper, 1959, 1963)의 가치는 불변하는 것이며, 대규모의 확증적 연구에서의 견고함과 더불어 소규모의 창의적이고 탐색적인 연구들도 지속되어야 할 것이다.

베이지안 통계의 적용[9]

추론 통계 내에서는 두 갈래의 전통이 존재한다 - 빈도주의 통계와 베이지안(Bayesian) 통계(Maxwell et al., 2015). 재현성 위기의 원인 중 하나인 영가설 유의성 검증의 문제점들을 보완하기 위하여 신뢰구간 등을 활용하는 것이 빈도주의적인 해결책이라면, 베이지안 통계를 도입하는 것은 통계적 추론의 패러다임을 전환하는 것이라고 할 수 있다. 앞서 기술하였듯이 빈도주의에서는 무수히 많은 시행의 반복표집을 가정하며 통계적 추론은 고정된 상수인 모수를 추정하는 것으로, 여기서 다루는 확률은 가설 확률이 아닌 일종의 빈도 확률을 의미한다(김청택, 2013; Greenland et al., 2016). 반면 베이지안 통계에서는 모수도 분포를 가지는 변수이며, 현재 관찰된 자료와 이전의 자료, 연구자의 믿음 등을 고려하여 모수의 분포를 추정하고 가설에 대한 믿음을 갱신한다(김용대, 김혜중, 오만숙, 오현숙, 정윤식, 2001; 김청택, 2013).

박준석(2015)은 심리학에서 통계적 추론 상의 문제점을 해결하고 연구결과의 재현성을 높이기 위하여 베이지안 통계가 대안이 될 수 있음을 세 가지 측면에서 일목요연하게 설명한 바 있다. 첫째, 베이지안 통계는 주어진 자료가 영가설 또는 연구가설을 지지하는 정도

된다(Simonsohn et al., 2014). 이 통계적 방법에 대한 더 상세한 정보는 Simonsohn 등(2014)을, 이 방법을 활용하여 세로토닌 유전형질의 조절효과에 대한 연구들을 메타분석한 예시로는 Taylor와 Munafò(2016)를 참고하라.

9) 미리 밝혀두건대 저자들은 베이지안 통계의 전문가가 아니며, 베이지안 통계에 대한 심층적인 논의는 본고의 주 목적에서도 벗어날 것이다. 다만 본고에서는 재현성 위기와의 관련성이라는 좁은 주제 내에서만 베이지안 통계와 관련된 내용을 간략히 소개하고자 한다. 이에 대한 보다 자세한 설명과 논의를 위해서는 해당 참고문헌을 보기 바란다(예: Berger, 1985; Wagenmakers, Lee, Lodewyckx, & Iverson, 2008).

를 수량화함으로써 연구자가 알고자 하는 정보를 직접적으로 제공한다. 둘째, 베이지안 통계는 표본 크기를 증가시키거나 임의로 데이터 수집을 중단함으로써 자신의 연구가설에 호의적인 결과를 의도적으로 만들어내는 *p*-hacking 방법들에 크게 영향 받지 않는다. 셋째, 베이지안 통계를 사용하면 오해석되는 경향이 감소할 것이다. 또한 Gelman(2015)도 사회과학이 다루는 현상의 변동성(variability)과 맥락의존성을 강조하며, 이러한 연구 주제의 특성이 연구결과가 재현되지 않는데 기여할 수 있음을 지적하였다. 그는 영가설검증의 더욱 핵심적인 문제점이 데이터 요약 수치로서 p 값을 사용하는 데 있는 것이 아니라 고정효과(constant effect)를 가정하는 기저 모델에 있다고 주장하면서, 변동성 있는 효과를 다루기 위해서는 숨겨진 조절변인을 고려한 상호작용 분석, 위계적인 베이지안 분석이 해결책이 될 수 있다고 주장하였다.

더불어 반복검증 연구의 성공 여부를 수량화하여 평가하기 위해서도 베이지안 분석을 적용하는 움직임이 있어 왔다. 빈도주의 통계 내에서 반복검증 시도의 성공은 원 연구와 반복검증 연구 사이의 p 값이나 효과 크기를 비교함으로써 결정되거나, 개별 연구들의 효과 크기를 메타분석함으로써 결정되었다(Verhagen & Wagenmakers, 2014). 그러나 원 연구와 반복검증 연구 간의 검정력 차이 등이 결론을 편향시킬 수 있다는 단점이 지적되어 왔으며, 베이지안 방식으로 반복검증 시도의 성공을 판별하는 여러 방법들이 제안되어 왔다(Bayarri & Mayoral, 2002; Rouder & Morey, 2011; Verhagen & Wagenmakers, 2014). 일례로 Verhagen과 Wagenmakers(2014)가 개발한 방법은 두 경쟁가설들의 적절성을 비교하는 베이즈 요인(Bayes factor)을 계산함에 있어, 회의론자의 가설(H_0; 효과가 신빙성이 없다, 즉 효과 크기가 0이라는 영가설)과 옹호론자의 가설(H_r; 원 연구에서 발견한 효과와 반복검증 연구의 효과가 일치한다는 것이며, 여기서 효과는 사후 분포로 수량화됨) 사이의 우도(weighted likelihood ratio)를 반복검증이 성공인지 실패인지 증명하는 증거로 삼는다. 이와 유사한 흐름에서 최근 Etz와 Vandekerckhove(2016)는 앞서 언급된 대규모 반복검증 프로젝트 중 하나인 Reproducibility 프로젝트(Open Science Collaboration, 2015)에 포함된 72개 연구를 대상으로 베이즈 요인을 계산하여 재현성의 정도를 평가하였는데, 비록 증거의 크기가 약하기는 했지만 상당수(75%)가 원 연구와 반복검증 연구 간의 증거 크기가 유사하다는 점을 보고하였다. 부가적으로 이들은 Reproducibility 프로젝트에서 원 연구와 반복검증 연구 간 나타난 차이점들은 대부분 표본 크기에 의해 설명될 수 있음을 입증하여, 작은 표본 사용 시 발생할 수 있는 낮은 검정력 이슈도 함께 지적하였다.

한편 베이지안 통계가 재현성 위기를 해결할 대안이 될 수 있을지에 대하여 다소 신중하거나 유보적인 입장을 취하는 학자들도 있다. Greenland 등(2016)은 베이지안 통계가 가설이 참일 확률(가설 확률)에 대한 직접적인 정보를 포함하는 장점이 있지만, 아직 p 값이나 신뢰구간만큼의 대중성을 얻지 못한 이유로 베이지안 모델의 철학적 가정에 대한 이견, 베이지안 분석 사용에 관한 관습이 아직 정립되지 못한 점을 들었다. 실제로 베이지안 통계는 그동안 사용되어 온 빈도주의적 추론 통계와는 다른 인식론적, 철학적 가정에 근거하고 있으며, 사후 확률 분포에 영향을 줄 수 있는 사전 분포(prior distribution) 선택 문제 등

이 논란이 되어 왔다(김용대 등, 2001; 박준석, 2015; Greenland et al., 2016; Trafimow & Marks, 2015). 또한 확률 등 수학적 이론과 계산에 대한 지식이 요구되고, 사용자 친화적인 그래픽 기반 통계 패키지가 부재한 등 연구 실제에 적용하기 위한 진입 장벽이 높다는 점도 베이지안 통계가 아직 연구자들에게 널리 보급되지 못하는 이유 중 하나로 꼽히고 있다(박준석, 2015; Greenland et al., 2016).

연구 방법론의 투명성 제고

마지막으로 여러 학자들이 공통적으로 지적하는 바는 어떠한 통계적 분석법도 오해석되거나 오용될 가능성이 있으며, 양질의 분석이 양질의 데이터 수집을 대체할 수 없다는 것이다(Wasserstein & Lazar, 2016). 또한 재현성을 높이기 위해서 데이터와 분석 관련 정보들을 충분히 공개하고, 연구 자료에 대한 접근성을 높여야 한다는 주장이 반복적으로 제기되어 왔다(Gelman, 2015; Greenland et al., 2016). 과학이 진실을 찾기 위한 작업이라고 할 때, 종국적으로 연구 재현성 위기를 극복하기 위해서는 연구 방법론의 투명성 제고가 핵심적인 과제일 것이다(Asendorpf et al., 2013; Hales, 2016; Stangor & Lemay, 2016). 임상심리학자인 Coyne (2016)은 반복검증을 계속해 나가는 것이 현재 심리학이 마주하고 있는 재현성 위기에 해답이 될 수는 없다고 비판하면서, 대신 연구의 투명성 강화가 근본적인 해결책임을 주장하였다. 비록 영가설이 실제로 참인지의 여부는 알 수 없지만, 수집한 증거의 질이 양호한지 불량한지에 관해서는 좀 더 경험적으로 알 수 있으며(Hales, 2016), 연구 데이터의 질과 투명성을 향상시키기 위한 노력은 연구의 재현성을 증진시키기 위하여 매우 핵심적인 부분이다. 이를 위해 앞서 언급된 QRP를 지양하는 것은 물론 논문에서 방법론 섹션을 가급적 상세하게 기술하고, 요청 시 동료 연구자들에게 연구에 사용된 자료 및 절차를 공유하고 연구자간 필연적인 이견을 조율하려는 개방적인 태도가 필요하다(Eich, 2014; John et al., 2012; Kahneman, 2014; Simmons et al., 2011).

이와 관련하여 연구 사전등록제(pre-registration)라는 새로운 제도적 해결책이 최근 들어 시도되고 있다. 연구 사전등록제는 연구를 시작하기 이전에 가설과 분석 등 연구계획을 공개적으로 등록하여 사전에 검증을 받고, 그 결과 또한 공개적으로 공유하는 것이다(박준석, 2016; Lindsay, 2015; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012). 만약 연구계획이 방법론적으로 건전하고 이론적으로 중요하다는 것이 동료 평가를 거쳐 확인되면, 저자들은 그를 바탕으로 논문을 작성하여 결과가 유의한지에 상관없이 원고를 출판하게 된다. 이와 같은 방식으로 연구의 투명성이 확보될 수 있으며, 학술지에 유의하지 않은 결과도 게재하게 됨으로써 출판편향의 폐해를 줄일 수 있을 것으로 기대된다(Hales, 2016; Stroebe, 2016). 사전등록제가 모든 유형의 연구에 적용 가능한 것은 아니며 아직 높은 비율로 실행되고 있지는 않으나, 최근의 몇몇 반복검증 연구는 사전등록제의 방식으로 진행된 바 있다. 예를 들어 *Perspectives on Psychological Science*에서는 Registered Replication Report라는 명칭 하에 2017년 초 현재까지 일련의 사전 등록된 반복검증 연구결과를 출판하였다(예: Alogna et al., 2014; Cheung et al., 2016; Eerland et al., 2016; Wagenmakers et al., 2016).

해결책 종합 요약

재현성 위기를 일거에 해결할 수 있는 최종적인, 단일 해결책은 존재하지 않는다(박준석, 2015; Gelman, 2015). 재현성 위기는 다면적 원인을 가지고 있기 때문에, 그에 따라 필요한 처방이 달라질 것이다. 신뢰구간과 효과 크기의 사용, 메타분석과 베이지안 통계의 적용은 재현되지 않는 위양성 결과를 양산할 우려가 높은 현행 통계적 추론 상의 문제점들을 보완하기 위한 해결책이며, 표본 크기를 증가시키고 검정력이 높은 연구를 설계하는 것, 연구 방법론의 투명성을 제고하는 것은 어떤 통계적 분석 전략을 채택하더라도 공통적으로 적용될 수 있는 해결책이다. 한편 출판편향과 같은 구조적인 문제는 사전등록제, 학술지 편집 방침의 변화 등의 해결책을 요할 것이다. 또 다른 관점에서 지금까지 제안된 해결책들은 재현성 평가의 중요성을 강조하는 것(반복검증에 대한 재평가)과 재현성이 높은 연구들이 출판되도록 하는 방안(신뢰구간과 효과 크기의 사용, 표본 크기의 증가, 메타분석의 활용, 베이지안 통계의 적용, 연구 방법론의 투명성 제고)으로도 분류될 수도 있을 것이다.

그렇다면 이러한 제안들에 따라 원 연구자, 반복검증 연구자, 학술지 편집자들에게 각기 요구되는 요소는 무엇인가? 우선 원 연구자들은 사전에 충분한 통계적 검정력을 갖출 수 있는 연구를 계획해야 하며, 여기에는 예상되는 효과 크기를 고려한 적절한 크기의 표본을 확보하는 것도 포함된다. 또한 자료 분석 시 *p*-hacking 등 QRP를 지양하며, 논문 작성 시 연구의 목적, 설계, 참여자를 포함 또는 배제한 기준, 원 분석 계획과 실제 실행된 분석 절차를 가급적 상세히 기술해야 한다(Greenland et al., 2016). 결과 보고 시에는 유의한 값만을 선택적 보고하는 것을 지양해야 하며(ASA, 2016; Gelman & Geurts, 2017; Greenland et al., 2016), p 값 외에 신뢰구간, 효과 크기 등의 정보를 함께 제공하는 것이 바람직할 것이다. 다음으로 반복검증 연구자들에게는 원 연구자에게 요구되는 것에 더하여, 2종 오류를 방지하기 위하여 충분히 큰 검정력을 갖춘 반복검증 연구를 계획하는 것의 중요성이 강조되어야 할 것이다. 검정력의 선택은 어느 정도 주관적일 수밖에 없으나, 특히 반복검증 연구의 경우 검정력이 .80 이상으로 큰 표본에서 진행되어야 한다는 점은 여러 연구자들이 거듭 지적해 왔다(Etz & Vandekerckhove, 2016; Maxwell et al., 2015). 또한 직접 반복검증 시 원 연구 절차를 주의 깊게 준수하여야 하며, 개념적 반복검증을 통하여 이론의 적용 범위와 경계 조건들을 확인해 나가는 노력도 필요할 것이다. 더불어 이러한 과정에서 원 연구자를 비롯하여 의견이 다른 연구자들 간의 소통 노력도 과학 공동체의 상호신뢰와 발전을 위해 필요할 것이다.

마지막으로 학술지 편집자들에게 요구되는 것은 잘못된 통계적 관행의 만연과 선택적 보고 및 출판편향을 감소시키기 위한 제도적인 변화 대책이다. 이를 위해 학술지 심사 및 편집 방침에서 기존 영가설 유의성 검증에 대한 의존을 줄이고 대안적인 해결책들(신뢰구간, 효과 크기, 메타분석, 베이지안 통계 등)의 적용을 장려하도록 명시하는 것이 필요할 것이다. 실제로 점차 더 많은 학술지들이 신뢰구간이나 효과 크기를 보고하도록 요구하고 있다(Greenland et al., 2016). 또한 편집자의 입장에서 반복검증 연구의 가치에 대한 재평가가 필요하며, 유의한 결과에만 출판에 인센티브

를 주는 현 관행에 대한 재고도 이루어져야 할 것이다. 이와 관련하여 사전등록제를 통하여 이론적, 방법론적으로 건전한 연구라면 연구결과가 유의한지의 여부와 관계없이 출판될 수 있도록 하는 방안도 출판편향을 줄이는 데 도움이 될 것이며, 특히 확증적 목적의 반복검증 연구는 사전 등록되는 것이 필요하다(Verhagen & Wagenmakers, 2014).

재현성 위기에 대처하는 국내외 학계 동향

지금부터는 재현성 위기에 대처하기 위하여 지금까지 제안되어 온 해결책들이 실제로 국내외 학계에서 연구 실제와 논문 출판에 어떠한 영향을 미치고 있는지를 살펴보고자 한다. 해외의 경우 재현성 향상과 연구의 투명성 제고를 위한 대표적인 움직임은 다음과 같다. 먼저 연구 사전등록제의 실시이다. 연구 사전등록제는 앞서 살펴본 Many Labs 프로젝트(Klein et al., 2014), Registered 프로젝트(Nosek & Lakens, 2014) 등의 재현성 검증 프로젝트에 적용된 바 있다. 또한 *Perspectives on Psychological Science*에서는 사전 등록된 반복검증 연구결과들을 출판하였다(Alogna et al., 2014; Cheung et al., 2016; Eerland et al., 2016; Wagenmakers et al., 2016).

다음으로 통계적 분석방법과 관련된 학술지 편집방침의 변화가 있다. 대표적으로 *Basic and Applied Social Psychology*(이하 *BASP*)는 2015년 첫 호의 사설을 통해 향후 *BASP*에 게재되는 논문에서 영가설검증 사용의 금지를 선언했다(Trafimow & Marks, 2015). *BASP*는 p 값을 비롯하여 이와 관련된 t, F 값과 신뢰구간의 사용까지도 금지하였는데, 이러한 조치는 통계학자들 사이에서 그 이득에 대한 논란을 일으키기도 하였다(Ashworth, 2015). *BASP*는 베이지안 분석에 대해서는 사례별로 판단을 내리기로 하여 요구 또는 금지를 하지 않고 가능성을 열어 두었으며, 대신 효과 크기, 강력한 기술통계, 데이터의 빈도나 분포의 제시, 큰 표본 크기 등의 대안을 강조함으로써(Trafimow & Marks, 2015), 심리학계의 재현성 논란에서 노정된 한계점들을 보완하기 위한 실질적인 조치들을 이행하는데 선구적 역할을 한 것으로 평가된다.

또 다른 학술지인 *Psychological Science*에서도 2014년 사전등록과 더불어 데이터 공개, 세부적인 연구 방법론의 요소들의 공개를 장려하는 투고 및 편집 방침을 표명한 바 있다(Eich, 2014). 여기에는 재현 가능한 연구들을 싣기 위한 보완책들이 포함되어 있는데, 이는 영가설검증의 대안으로 제안된 효과 크기, 신뢰구간 등을 보고하도록 장려하는 것과 함께 QRP와 관련된 해결책들(표본 크기를 결정한 근거와 자료 수집을 중지한 규칙을 밝힘, 연구에 포함된 모든 독립 및 종속변수들을 공개, 원고의 방법과 결과 부분을 상세히 기술할 수 있도록 해당 부분의 글자 수 제한 폐지)을 망라하고 있다. 또한 투명성을 증진하기 위한 대책으로 연구 자료, 도구(절차와 분석)를 공개하거나 사전등록제를 실시한 연구를 인정하는 차원에서 배지(badge)를 달아 표시하도록 하였다(예: Open Data badge, Open Materials badge, Preregistered badge). 흥미롭게도 이러한 편집방침 변화 이후 *Psychological Science* 논문 투고와 출간 현황이 어떻게 달라졌는지를 추적한 연구들이 있으며(Lindsay, 2017), 다른 학술지들에 비하여 *Psychological Science*에서 연구 자료 등을

표 2. 미국심리학회(APA) 산하 주요 학술지의 재현성 관련 투고 및 편집규정

[*Psychological Bulletin*]
• Call for papers: Replication and reproducibility: http://www.apa.org/pubs/journals/bul/call-for-papers-replication.aspx
▷ 재현 연구 특별호(혹은 특별 섹션)를 위한 논문 투고 공지 중
• Editorial(Albarracin, 2015)
▷ 투고 논문은 보고된 방법이 재현될 수 있음을 입증하는 충분한 논의가 이루어져야 함
▷ 재현성 문제는 특정한 방법론에 대한 이상화로는 해결되지 않으며, 우수한 연구의 종합(syntheses)이 중요함, 기존 연구를 종합적 관점에서 다루는 *Psychological Bulletin*이 재현성 문제의 개선에 기여할 수 있음
[*Behavioral Neuroscience*]
• Manuscript Submission: http://www.apa.org/pubs/journals/bne/index.aspx?tab=4
▷ 투고 대상 논문은 새로운 발견을 제시하는 전형적인 연구 논문 외에 기존 연구와 반대되는 발견, 반복검증 논문, 사전 등록 연구(registered reports) 등도 포함함
• 반복검증 논문의 경우 사전등록을 권장함
[*Journal of Personality and Social Psychology*]
• Manuscript Submission: http://www.apa.org/pubs/journals/psp/index.aspx?tab=4
▷ 반복검증이 본 학술지의 핵심 미션은 아니지만, 기존의 중요한 발견을 재현하는 논문의 투고를 장려함
▷ 반복검증 논문의 주요 출판 기준
- 재현된 발견의 이론적 중요성
- 반복검증 연구의 검정력
- 원 연구와 방법론, 절차, 자료 등의 유사성
- 동일한 발견이 보고된 선행 반복검증 연구의 횟수 및 검정력
- 우선권 대상 : 원 연구자가 아닌 연구자의 연구, 개념적 재현 보다는 직접적 재현, 원 연구에 포함된 여러 연구(multi-study) 가운데 하나 이상의 재현 시도
- 반복검증 연구는 온라인으로만 제공되며, 인쇄본에는 목차만 제시됨

공개하는 비율이 증가하였고(Giofrè, Cumming, Fresc, Boedker, & Tressoldi, 2017; Kidwell et al., 2016), 영가설검증 외의 통계적 대안들을 적용하는 사례가 늘어난 것으로 나타났다(Giofrè et al., 2017). 이는 비록 완벽한 개선에는 시간이 걸릴지라도, 학술지의 투고 및 편집 방침의 변화가 연구자 및 심사자, 편집자의 관행에 일정한 긍정적인 변화를 미칠 수 있다는 점을 예증하는 것이라고 할 수 있겠다.

한편 상기와 같은 선언적 사례 이외에 재현성 관련 제도적인 추세를 파악하기 위해 미국심리학회(American Psychological Association: 이하 APA) 산하 학술지 가운데 주요 분야별로 한 종씩 총 여섯 개 분야의 학술지[10]를 선정하여

10) *Psychological Bulletin*('Core of Psychology' 분야), *Behavioral Neuroscience*('Neuroscience & Cognition' 분야), *Journal of Experimental Psychology: Applied* ('Basic/Experimental Psychology' 분야), *Journal of Abnormal Psychology*('Clinical Psychology' 분야), *Journal*

표 3. 한국심리학회 산하 학회의 재현성 관련 규정 현황

주요 규정	비고 (효과, 의미)
• 연구의 위조, 변조, 조작의 금지 및 판정 관련 규정 ▷ 학회 : 임상, 상담, 산업 및 조직, 발달, 인지 및 생물, 건강, 문화 및 사회문제	• 연구윤리의 제고 • 재현성 저하를 유발하는 위·변조 견제
• 타 연구자의 결과 재검증(반복검증)을 위한 연구자료 공유 ▷ 학회 : 상담, 발달, 중독, 문화 및 사회문제	• 연구진실성 제고 및 후속 반복검증 연구 지원
• <Brief Report>의 주요 투고 대상 가운데 “의미 있는 반복검증” 논문을 적시함 ▷ 학회 : 임상	• 반복검증 연구의 활성화

재현성과 관련된 내용이 명시되거나 강화되었는지 규정을 검토하였다. 비록 포괄적인 전수조사는 아니지만, 그 결과 *Journal of Experimental Psychology: Applied*, *Journal of Abnormal Psychology*, *Journal of Applied Psychology* 등은 재현성과 관련된 특별한 강조 사항을 발견할 수 없었다. 나머지 세 개의 학술지는 재현성과 관련된 의미 있는 규정이 명시되어 있었으며, 주요 내용은 표 2와 같다.

국내의 경우 재현성 재고를 위한 학계 및 학술지 차원의 대응은 해외 동향 대비 미미한 상황으로 판단된다. 한국심리학회 산하 학회들의 홈페이지[11]에 탑재된 각종 규정을 검토한 결과 재현성과 관련된 규정은 표 3에 제시된 것과 같이 파악되었다. 가장 많은 학회가 제시하고 있는 것은 연구의 위조, 변조, 조작의 금지와 관련된 규정이다. 심리학 연구의 재현성 문제가 촉발된 계기 중 하나가 Stapel의 데이터 조작이었다는 점에서 이러한 규정은 재현성 향상의 필요조건이지만, 이는 재현성 이슈에 국한되기보다는 기본적인 연구윤리에 포함되는 사항들로 볼 수 있다. 두 번째는 반복검증을 위한 연구자료 공유 의무로서, 상담, 발달, 중독, 문화 및 사회문제 분과의 학회지들에서 채택하고 있었다. 마지막으로 한국임상심리학회에서 발행하는 *Korean Journal of Clinical Psychology*는 투고 가능한 논문 형태 중 Brief Report의 투고 대상 가운데 하나로 “의미 있는 반복검증”을 지정하고 있다. 전술한 바와 같이 반복검증 논문에 대한 학술지들의 전반적인 관심 수준이 낮다는 점을 고려할 때, 주요 투고 대상으로 반복검증 논문을 명시적으로 적시한 것은 적지 않은 의미를 지닌다고 하겠다.

이상 살펴본 바와 같이 국내 심리학 학술지들의 재현성 관련 규정은 많지 않고 그 내용도 대체로 원론적인 수준이다. 이에 한국심리학회 산하 각 학회의 편집위원장들에게 이메일을 통해 홈페이지에 탑재된 공식적인 규정과는 별도로 재현성과 관련하여 논의되었거나 논의되고 있는 편집방침 관련 내용이 있는지 부가적으로 문의하였다. 문의 결과 8개 학회

of Personality and Social Psychology(Social Psychology and Social Processes 분야), *Journal of Applied Psychology*(‘Industrial/Organizational Psychology’ 분야)

11) 15개 산하 학회 가운데 2017년 6월 기준 접속이 불가한 한국법심리학회를 제외한 14개 학회의 홈페이지 탑재 규정을 검토하였다.

지의 편집위원장들로부터 회신이 왔으며 그 내용을 요약하면 다음과 같다. 첫째, 재현성은 향후 고려해야 할 중요 문제이지만 현재까지는 학회 차원의 방침이나 논의가 미미하다. 둘째, 현재 재현성 제고를 위해 진행 중인 논의는 없으며, 이번 논문이 논의의 계기가 될 수도 있겠다. 셋째, 비록 현재 진행 중인 논의나 계획은 없으나, 투고와 심사 과정에서 연구의 재현성을 판단할 수 있는 정보가 성실히 보고되었는지는 중요한 심사 준거이다. 넷째, 향후 재현성 강화를 위한 조치를 추진할 경우 편집위원회 상정, 논의, 정기 이사회 심의 등에 상당한 시간이 걸릴 것으로 예상된다. 이와 같이 현재로서는 재현성 강화와 관련하여 국내 학회 차원의 주목할 만한 논의는 진행되지 않고 있는 상황으로 파악된다.

결 어

지금까지 최근 심리학 내에서 재현성 위기와 관련된 담론들이 촉발되고 진행된 과정 및 그 원인과 해결책들에 관하여 개관하였다. 요약하면, 심리학 내의 재현성 위기는 심리학 연구의 반복검증 성공률이 낮다는 문제 제기에 의하여 촉발되었으며 이후 그 원인을 면밀하게 파악하고 대안을 제시하기 위한 자정 노력들이 이루어져 왔다. 낮은 재현성의 원인으로서 영가설검증에 의존하는 통계적 추론절차, 연구의 타당성을 저해하는 의심스러운 연구관행, 출판편향 등이 집중적인 조명을 받아왔고, 이에 대한 대안으로서 신뢰구간과 효과 크기의 사용, 메타분석의 활용, 베이지안 통계로의 전환 그리고 연구 방법론의 투명성 제고 등에 대한 강조를 중심으로 해결책들이 제시되고 있다. 다만 제안된 대안과 해결책들이 실제 심리학계와 주요 학술지의 실제에 충분히 파급되기에는 아직 시간과 노력이 더 필요한 상황으로 여겨진다. 모든 변화는 양면성을 내포하고 있으며, 위기는 기회이기도 하다. 특히 이러한 담론의 과정에서 우리가 일상적으로 또는 관성적으로 과학을 해온 방식에 대한 내성과 통찰을 시도하게 된 것은 재현성 위기를 통하여 얻게 된 가장 큰 유익이라고 생각된다.

본 논문은 국내에서 재현성 위기와 관련된 이슈를 직접적으로 다루는 첫 번째 시도로서 의의를 갖는다. 즉, 국내 심리학계에 재현성 이슈의 중요성을 환기하고 앞으로 더욱 견고하고 투명한 방식으로 연구가 진행될 수 있는 개선 방안들이 공론의 장에서 논의될 수 있는 기초를 마련하고자 하는 것이 주 목적이다. 다만 본 논문은 그러한 논의의 첫 걸음으로서 재현성 위기와 관련된 제반 사항들을 개괄적으로 두루 소개하고 있기에 그에 따른 한계를 동시에 가지고 있다. 이 복잡한 이슈에 내재된 여러 문제점들과 잠정적인 해결책들을 모두 심도 있게 다루는 것은 지면상의 제한뿐만 아니라 저자들의 개인적 역량을 넘어서는 일이 될 것이다. 더불어 이는 재현성 이슈가 한 개의 논문에서 충분한 깊이로 다루기에는 실로 거대한 주제라는 특성에도 기인하는 것으로 생각된다. 국외에서 재현성과 관련된 매우 방대한 양의 학술 논문들이 다수 발표되어 있는 점도 이를 방증한다. 본 논문을 시작으로 향후 다양한 전문 분야의 심리학자들에 의하여 국내 학계의 논의가 세부적으로 발전되어 가기를 바라며, 특히 사회심리학과 계량심리학을 전공하는 심리학자들이 후속 논의에 동참하여 본 논문의 한계를 보완해 주기를 기대한다.

참고문헌

교보문고 (2016). 종합 연간 베스트. http://www.kyobobook.co.kr/bestSellerNew/bestseller.laf?range=1&kind=3&orderClick=DAC&mallGb=KOR&linkClass=A에서 2016, 12, 13 자료 얻음.

김용대, 김혜중, 오만숙, 오현숙, 정윤식 (2001). 베이지안 통계학의 과거 · 현재 · 미래. 한국통계학회논문집, 8, 47-64.

김청택 (2011). 통계적 가설검증의 절차와 문제점 그리고 대안. 서울: 민속원.

김청택 (2013). Bayesian 통계의 소개. 강의자료 (2013. 6. 2.)

박준석 (2015). 영가설 유의성검증 절차의 문제점들에 대한 해결책으로서의 베이지언 통계학: 심리학의 경우. 과학철학, 18(2), 135-147.

박준석 (2016). '연구 사전등록제', 재현성 위기의 제도적 해법. 사이언스온. http://scienceon.hani.co.kr/media/414001에서 2017, 1, 7 자료 얻음.

한국금융신문 (2015). 또 한 번의 개가(凱歌). http://www.fntimes.com/paper/view.aspx?num=142353에서 2016, 12, 15 자료 얻음.

한국연구재단 (2016). 한국학술지인용색인 (Korea Citation Index). https://www.kci.go.kr/kciportal/po/search/poCitaSearList.kci?years=5&search=directory&year=2015&field=*&page=1에서 2016, 12, 13 자료 얻음.

Albarracín, D. (2015) Editorial. *Psychological Bulletin, 141*(1), 1-5.

Alogna, V. K., Attaya, M. K., Aucoin, P., Bahník, Š., Birch, S., Birt, A. R., ... & Buswell, K. (2014). Registered replication report: Schooler and engstler-schooler (1990). *Perspectives on Psychological Science, 9*(5), 556-578.

American Statistical Association (2016). American Statistical Association releases statement on statistical significance and *p*-values. https://www.amstat.org/search?search=American%20Statistical%20Association%20releases%20statement%20on%20statistical%20significance%20and%20p-values에서 2017, 7, 3 자료 얻음

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J., Fiedler, K., ... & Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality, 27*(2), 108-119.

Ashworth, A. (2015). Veto on the use of null hypothesis testing and *p* intervals: Right or wrong? Taylor & Francis Editor. http://editorresources.taylorandfrancisgroup.com/veto-on-the-use-of-null-hypothesis-testing-and-p-intervals-right-or-wrong에서 2017, 7, 1 자료 얻음.

Association for Psychological Science (2013). *Registered replication reports*. http://www.psychologicalscience.org/publications/replication에서 2016, 12, 1 자료 얻음.

Balluerka, N., Gómez, J., & Hidalgo, D. (2005). The controversy over null hypothesis significance testing revisited. *Methodology, 1*(2), 55-70.

Banerjee, P., Chatterjee, P., & Sinha, J. (2012). Is it light or dark? Recalling moral behavior changes perception of brightness. *Psychological Science, 23*(4), 407-409.

Bargh, J. A. (2012). Priming effects replicate just fine, thanks. *Psychology Today, 11*. https://www.psychologytoday.com/blog/the-natural-unconsciou

s/201205/priming-effects-replicate-just-fine-thanks에서 2017, 1, 7 자료 얻음.

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology, 71*(2), 230-244.

Baumeister, R. F. (2016). Charting the future of social psychology on stormy seas: Winners, losers, and recommendations. *Journal of Experimental Social Psychology, 66*, 153-158.

Bayarri, M. J., & Mayoral, A. M. (2002). Bayesian design of "successful" replications. *The American Statistician, 56*(3), 207-214.

Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature, 483*(7391), 531-533.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology, 100*(3), 407-425.

Berezow, A. B. (2012). *Why psychology isn't science*. http://articles.latimes.com/2012/jul/13/news/la-ol-blowback-pscyhology-science-20120713에서 2016, 7, 14 자료 얻음.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis* (2nd ed). New York: Spinger.

Bhattacharjee, Y (2013). *The mind of a con man*. http://www.nytimes.com/2013/04/28/magazine/diederik-stapels-audacious-academic-fraud.html?pagewanted=all에서 2016, 12, 1 자료 얻음.

Bissell, M. (2013). The risks of the replication drive. *Nature, 503*(7476), 333-334.

Bollen, K., Cacioppo, J. T., Kaplan, R. M., Krosnick, J. A., & Olds, J. L. (2015). *Social, behavioral, and economic sciences perspectives on robust and reliable science*. www.nsf.gov/sbe/AC_Materials/SBE_Robust_and_Reliable_Research_Report.pdf에서 2016, 12, 1 자료 얻음.

Brandt, M. J., IJzerman, H., & Blanken, I. (2014). Does recalling moral behavior change the perception of brightness? *Social Psychology, 45*(3), 246-252.

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., ... & Van't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology, 50,* 217-224.

Braver, S. L., Thoemmes, F. J., & Rosenthal, R. (2014). Continuously cumulating meta-analysis and replicability. *Perspectives on Psychological Science, 9*(3), 333-342.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*(5), 365-376.

Carter, T. J., Ferguson, M. J., & Hassin, R. R. (2011). A single exposure to the American flag shifts support toward Republicanism up to 8 months later. *Psychological Science, 22*(8), 1011-1018.

Caruso, E. M., Vohs, K. D., Baxter, B., & Waytz, A. (2013). Mere exposure to money increases endorsement of freemarket systems and social inequality. *Journal of Experimental Psychology: General, 142*(2), 301-306.

Center for Open Science (2013). *Our mission is to increase openness, integrity, and reproducibility of research.* https://cos.io/about/mission/에서 2016, 1, 10 자료 얻음.

Cesario, J. (2014). Priming, replication, and the hardest science. *Perspectives on Psychological Science, 9*(1), 40-48.

Chambers, C. (2014). *Physics envy: Do 'hard' sciences hold the solution to the replication crisis in psychology?* http://www.theguardian.com/science/head-quarters/2014/jun/10/physics-envy-do-hard-sciences-hold-the-solution-to-the-replication-crisis-in-psychology에서 2016, 7, 14 자료 얻음.

Chambers, C. D., Feredoes, E., Muthukumaraswamy, S. D., & Etchells, P. (2014). Instead of "playing the game" it is time to change the rules: Registered Reports at *AIMS Neuroscience* and beyond. *AIMS Neuroscience, 1*(1), 4-17.

Cheung, I., Campbell, L., & LeBel, E. P., ... & Yong, J. C. (2016). Registered Replication Report: Study 1 from Finkel, Rusbult, Kumashiro, & Hannon (2002). *Perspectives on Psychological Science, 11*(5), 750-764.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology, 65*(3), 145-153.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). NJ: Lawrence Earlbaum Associates.

Cohen, J. (1994). The earth is round ($p<.05$). *American Psychologist, 49*(12), 997-1003.

Coyne, J. C. (2016). Replication initiatives will not salvage the trustworthiness of psychology. *BMC Psychology,* 4:28.

Crandall, C. S., & Sherman, J. W. (2016). On the scientific superiority of conceptual replications for scientific progress. *Journal of Experimental Social Psychology, 66*, 93-99.

Crocker, J., & Cooper, M. L. (2011). Addressing scientific fraud. *Science, 334*(6060), 1182-1182.

Crotty, D. (2014). *When crises collide: The tension between null results and reproducibility*. https://scholarlykitchen.sspnet.org/2014/09/10/when-crises-collide-the-tension-between-null-results-and-reproducibility/에서 2016, 7, 18 자료 얻음.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. NY: Routledge.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*(1), 7-29.

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist, 60*(2), 170-180.

Dovidio, J. F. (2016). Commentary: A big problem requires a foundational change. *Journal of Experimental Social Psychology, 66*, 159-165.

Doyen, S., Klein, O., Pichon, C. L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLoS One, 7*(1), e29081.

Driscoll, R., Davis, K. E., & Lipetz, M. E. (1972). Parental interference and romantic love: The Romeo and Juliet effect. *Journal of Personality and Social Psychology, 24*(1), 1-10.

Dunn, D. S. (2015). *The Oxford handbook of undergraduate psychology education*. Oxford University Press.

Eerland, A., Sherrill, A. M., Magliano, J. P., Zwaan, R. A., Arnal, J. D., Aucoin, P., ... & Crocker, C. (2016). Registered replication report: Hart & Albarracín (2011). Perspectives on *Psychological Science, 11*(1), 158-171.

Eich, E. (2014). Business not as usual. *Psychological Science, 25*(1), 3-6.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ, 315*(7109), 629-634.

Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge University Press.

Etz, A., & Vandekerckhove, J. (2016). A Bayesian perspective on the reproducibility project: Psychology. *PLoS One, 11*(2), e0149794.

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *PLoS One, 5*(4), e10068.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175-191.

Feinstein, A. R. (1995). Meta-analysis: Statistical alchemy for the 21st century. *Journal of Clinical Epidemiology, 48*(1), 71-79.

Finkel, E. J., Eastwick, P. W., & Reis, H. T. (2015). Best research practices in psychology: Illustrating epistemological and pragmatic considerations with the case of relationship science. *Journal of Personality and Social Psychology, 108*(2), 275-297.

Fraley, R. C., & Vazire, S. (2014). The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PLoS One, 9*(10), e109019.

Francis, G. (2012a). Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin & Review, 19*(6), 975-991.

Francis, G. (2012b). The psychology of replication and replication in psychology. *Perspectives on Psychological Science, 7*(6), 585-594.

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science, 345*(6203), 1502-1505.

Funder, D. C., Levine, J. M., Mackie, D. M., Morf, C. C., Vazire, S., & West, S. G. (2014). Improving the dependability of research in personality and social psychology recommendations for research and educational practice. *Personality and Social Psychology Review, 18*(1), 3-12.

Galak, J., LeBoeuf, R. A., Nelson, L. D., & Simmons, J. P. (2012). Correcting the past: Failures to replicate psi. *Journal of Personality and Social Psychology, 103*(6), 933-948.

Gelman, A. (2015). The connection between varying treatment effects and the crisis of unreplicable research: A Bayesian perspective. *Journal of Management, 41*(2), 632-643.

Gelman, A., & Geurts, H. M. (2017). The statistical crisis in science: How is it relevant to clinical neuropsychology? *The Clinical Neuropsychologist*, 1-15.

Gergen, K. J. (1973). Social psychology as history.

Journal of Personality and Social Psychology, 26(2), 309-320.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics, 33*(5), 587-606.

Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The Null Ritual: What you always wanted to know about significance testing but were afraid to ask. In D. Kaplan (Ed.), *The Sage book of quantitative methodology for the social sciences* (pp. 391-408). SAGE publications.

Gilbert, D. T. (2014). *Some thoughts on shameless bullies*. http://www.psychol.cam.ac.uk/cece/blog 에서 2014, 9, 11 자료 얻음.

Ginger-Sorolla, R. (2012). Science or art? How aesthetic standards grease the way through the publication bottleneck but undermine science. *Perspectives on Psychological Science, 7*(6), 562-571.

Giofrè, D., Cumming, G., Fresc, L., Boedker, I., & Tressoldi, P. (2017). The influence of journal submission guidelines on authors' reporting of statistics and use of open research practices. *PloS One, 12*(4), e0175583.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, *P* values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology, 31*, 337-350.

Hales, A. H. (2016). Does the conclusion follow from the evidence? Recommendations for improving research. *Journal of Experimental Social Psychology, 66*, 39-46.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of *p*-hacking in science. *PLoS Biology, 13*(3), e1002106.

Hedges, L. V. (1987). How hard is hard science, how soft is soft science? The empirical cumulativeness of research. *American Psychologist, 42*(5), 443-455.

Hüffmeier, J., Mazei, J., & Schultze, T. (2016). Reconceptualizing replication as a sequence of different studies: A replication typology. *Journal of Experimental Social Psychology, 66*, 81-92.

Husnu, S., & Crisp, R. J. (2010). Elaboration enhances the imagined contact effect. *Journal of Experimental Social Psychology, 46*(6), 943-950.

Hyde, J. S. (2001). Reporting effect sizes: The roles of editors, textbook authors, and publication manuals. *Educational and Psychological Measurement, 61*(2), 225-228.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Med, 2*(8), e124.

Jasny, B. R., Chin, G., Chong, L., & Vignieri, S. (2011). Again, and again, and again…. *Science, 334*(6060), 1225-1225.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*(5), 524-532.

Johnson, D. J., Cheung, F., & Donnellan, M. B. (2014). Does cleanliness influence moral judgments? *Social Psychology, 45*(3), 209-215.

Judd, C. M., & Gawronski, B. (2011). Editorial comment. *Journal of Personality and Social Psychology, 100*(3), 406.

Kahneman, D. (2014). A new etiquette for

replication. *Social Psychology, 45*(4), 310-311.

Kail, R. V. (2012). Reflections on five years as editor. *Observer, 25*(9).

Kelley, H. (1992). Common-sense psychology and scientific psychology. *Annual Review of Psychology, 43*(1), 1-23.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*(3), 196-217.

Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L. S., ... & Errington, T. M. (2016). Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. *PLoS Biology, 14*(5), e1002456.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahník, Š., Bernstein, M. J., ... & Nosek, B. A. (2014). Investigating variation in replicability. *Social Psychology, 45*(3), 142-152.

Krueger, J. (2001). Null hypothesis significance testing: On the survival of a flawed method. *American Psychologist, 56*(1), 16-26.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for *t*-tests and ANOVAs. *Frontiers in Psychology, 4*(863).

Lakens, D., Hilgard, J., & Staaks, J. (2016). On the reproducibility of meta-analyses: Six practical recommendations. *BMC Psychology*, 4:24.

Laws, K. R. (2016). Psychology, replication & beyond. *BMC Psychology,* 4:30.

Levine, T. R., Weber, R., Hullett, C., Park, H. S., & Lindsey, L. L. M. (2008). A critical assessment of null hypothesis significance testing in quantitative communication research. *Human Communication Research, 34*(2), 171-187.

Lilienfeld, S. O. (2012). Public skepticism of psychology: Why many people perceive the study of human behavior as unscientific. *American Psychologist*, *67*(2), 111-129.

Lindsay, D. S. (2015). Replication in psychological science. *Psychological Science, 26*(12), 1827-1832.

Lindsay, D. S. (2017). *Editor's perspective: What happened when Psychological Science began encouraging the use of the new statistics?* Presentation in the 29th Association for Psychological Science (APS) Convention. Boston, MA. https://thenewstatistics.com/itns/wp-content/uploads/2017/05/APS-Editor%E2%80%99s-Perspective.pdf에서 2017, 7, 11 자료 얻음.

Lynott, D., Corker, K. S., Wortman, J., Connell, L., Donnellan, M. B., Lucas, R. E., & O'Brien, K. (2014). Replication of "Experiencing physical warmth promotes interpersonal warmth" by Williams and Bargh (2008). *Social Psychology, 45*(3), 216-222.

Madden, C. S., Easley, R. W., & Dunn, M. G. (1995). How journal editors view replication research. *Journal of Advertising, 24*(4), 77-87.

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher, 43*(6), 304-316.

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science, 7*(6), 537-542.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does “failure to replicate” really mean? *American Psychologist, 70*(6), 487-498.

McGrath, J. E. (1981). Dilemmatics: The study of research choices and dilemmas. *American Behavioral Scientist, 25*(2), 179-201.

Moonesinghe, R., Khoury, M. J., & Janssens, A. C. J. (2007). Most published research findings are false: But a little replication goes a long way. *PLoS Medicine, 4*(2), e28.

Murayama, K., Pekrun, R., & Fiedler, K. (2014). Research practices that can prevent an inflation of false-positive rates. *Personality and Social Psychology Review, 18*(2), 107-118.

Neuliep, J. W., & Crandall, R. (1990). Editorial bias against replication research. *Journal of Social Behavior and Personality, 5*(4), 85-90.

Neuliep, J. W., & Crandall, R. (1993). Reviewer bias against replication research. *Journal of Social Behavior and Personality, 8*(6), 21-29.

Nissen, S. B., Magidson, T., Gross, K., & Bergstrom, C. T. (2016). Publication bias and the canonization of false facts. *Elife, 5*, e21451.

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology, 45*(3), 137-141.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7*(6), 615-631.

Nuzzo, R. (2014). Statistical errors: *P* values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume. *Nature, 506*(7487), 150-153.

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science, 7*(6), 657-660.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), 943-951.

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science, 7*(6), 531-536.

Pashler, H., Harris, C. R., & Coburn, N. (2011). *Elderly-related words prime slow walking*. http://www.PsychFileDrawer.org/replication.php?attempt=MTU%3D에서 2016, 12, 9 자료 얻음.

Pashler, H., & Wagenmakers, E. J. (2012). Editors introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science, 7*(6), 528-530.

Popper, K. R. (1959). *The logic of scientific discovery*. London: Hutchinson.

Popper, K. R. (1963). *Conjectures and refutations: The growth of scientific knowledge*. London: Routledge & Kegan Paul.

PsychFileDrawer (2012). *Archives for the attempts at replication in experimental psychology*. http://psychfiledrawer.org/에서 2016, 6, 27 자료 얻음.

Ritchie, S. J., Wiseman, R., & French, C. C. (2012). Failing the future: Three unsuccessful attempts to replicate Bem's retroactive facilitation of recall effect. *PLoS One, 7*(3), e33423.

Robinson, E. (2011). Not feeling the future: A failed replication of retroactive facilitation of memory recall. *Journal of the Society for Psychical Research, 75*(904), 142-147.

Roediger, H. L. (2012). Psychology's woes and a partial cure: The value of replication. *APS Observer, 25*(2), 9. http://www.psychologicalscience.org/issue/february-12에서 2016, 12, 26 자료 얻음.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*(3), 638-641.

Rouder, J. N., & Morey, R. D. (2011). A Bayes factor meta-analysis of Bem's ESP claim. *Psychonomic Bulletin & Review, 18*(4), 682-689.

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology, 58*(5), 646-656.

Sakaluk, J. K. (2016). Exploring small, confirming big: An alternative system to the new statistics for advancing cumulative and replicable psychological research. *Journal of Experimental Social Psychology, 66*, 47-54.

Schaller, M. (2016). The empirical benefits of conceptual rigor: Systematic articulation of conceptual hypotheses can reduce the risk of non-replicable results (and facilitate novel discoveries too). *Journal of Experimental Social Psychology, 66*, 107-115.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1*(2), 115-129.

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology, 13*(2), 90-100.

Schnall, S. (2014). *Simone Schnall on her experience with a Registered Replication Project*. http://www.spsp.org/blog/simone-schnall-on-her-experience-with-a-registered-replication-project에서 2014, 9, 11 자료 얻음.

Schnall, S., Benton, J., & Harvey, S. (2008). With a clean conscience cleanliness reduces the severity of moral judgments. *Psychological Science, 19*(12), 1219-1222.

Schweinsberg, M., Madan, N., Vianello, M., Sommer, S. A., Jordan, J., Tierney, W., ... & Srinivasan, M. (2016). The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. *Journal of Experimental Social Psychology, 66*, 55-67.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105*(2), 309-316.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359-1366.

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science, 9*(1), 76-80.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). *P*-curve: A key to the file-drawer. *Journal of Experimental Psychology: General, 143*(2), 534-547.

Sinclair, H. C., Hood, K. B., & Wright, B. L. (2014). Revisiting the Romeo and Juliet effect (Driscoll, Davis, & Lipetz, 1972). *Social Psychology, 45*(3), 170-178.

Spellman, B. A. (2012). Introduction to the special section on research practice. *Perspectives on Psychological Science, 7*(6), 655-656.

Stangor, C., & Lemay, E. P. (2016). Introduction to the special issue on methodological rigor and replicability. *Journal of Experimental Social Psychology, 66*, 1-3.

Stroebe, W. (2016). Are most published social psychological findings false? *Journal of Experimental Social Psychology, 66*, 134-144.

Stroebe, W., Postmes, T., & Spears, R. (2012). Scientific misconduct and the myth of self-correction in science. *Perspectives on Psychological Science, 7*(6), 670-688.

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science, 9*(1), 59-71.

Taylor, A. E., & Munafò, M. R. (2016). Triangulating meta-analyses: The example of the serotonin transporter gene, stressful life events and major depression. *BMC Psychology*, 4:23.

Thaler, R. H. (2016). 똑똑한 사람들의 멍청한 선택 (박세연 역). 서울: 리더스북. (원전은 2015년에 출판).

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology, 37*, 1-2.

Turner, E. H., Matthews, A. M., Linardatos, E., Tell, R. A., & Rosenthal, R. (2008). Selective publication of antidepressant trials and its influence on apparent efficacy. *New England Journal of Medicine, 358*(3), 252-260.

Verhagen, J., & Wagenmakers, E. J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General, 143*(4), 1457-1475.

Wagenmakers, E. J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams Jr, R. B., ... & Bulnes, L. C. (2016). Registered Replication Report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science, 11*(6), 917-928.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on *p*-values: Context, process, and purpose. *The American Statistician, 70*(2), 129-133.

Wagenmakers, E. J., Lee, M., Lodewyckx, T., & Iverson, G. J. (2008). Bayesian versus frequentist inference. In H. Hoijtink, I. Klugkist, & P. A. Boelen (Eds.), *Bayesian evaluation of informative hypotheses* (pp. 181-207). New York: Spinger Verlag.

Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7*(6), 632-638.

Williams, L. E., & Bargh, J. A. (2008). Experiencing physical warmth promotes interpersonal warmth. *Science, 322*(5901), 606-607.

Wilson, T. (2012). *Stop bullying the 'soft' sciences.*

http://articles.latimes.com/2012/jul/12/opinion/la-oe-wilson-social-sciences-20120712에서 2016, 7, 14 자료 얻음.

Yong, E. (2012). In the wake of high profile controversies, psychologists are facing up to problems with replication. *Nature, 485*(17), 298-300.

1차원고접수 : 2017. 02. 20.
수정원고접수 : 2017. 07. 14.
최종게재결정 : 2017. 09. 01.

한국심리학회지 : 일반
Korean Journal of Psychology: General
2017, Vol. 36, No. 3, 359-396
http://dx.doi.org/10.22257/kjp.2017.09.36.3.359

Replication crisis in psychology: A review of its causes and solutions

Bin-Na Kim
Department of Psychology
Seoul National University

Joonwon Choi
Department of Business Administration
Sangmyung University

Hyunseok Ko
Department of Liberal Studies
Korea Air Force Academy

Recently, there has been an ongoing controversy surrounding replication crisis in psychology. Some of popular experiments such as priming effect failed to be replicated. And subsequent large-scale replication projects of psychological studies revealed that the success rate of replication was less than ideal, which threatened the status of psychology as science. Through the course of mass debate, problems such as null hypothesis statistical testing, questionable research practices, and publication bias were discussed as causes of current replication crisis. Accordingly, solutions (use of confidence interval, effect size, meta-analysis, Bayesian statistics, and efforts to increase transparency in methodology) were proposed in order to enhance reproducibility of psychological research. Depending on how it is resolved, the present crisis in replication also holds opportunity for future development of psychology. Therefore, we attempted to introduce how replication crisis unfolded in psychology and to comprehensively review causes and potential solutions in the hope of establishing basic ground for further discussion in Korea.

Key words : replication, reproducibility, publication bias, methodology