연구평가지표를 위한 라이덴 선언

연구평가를 위한 열 가지 원칙

by Diana Hicks, Paul Wouters, Ludo Waltman, Sarah de Rijcke & Ismael Rafols

<http://www.leidenmanifesto.org/>

연구활동을 지원하고 운영하는데 있어 실증적인 데이터 사용이 증가하고 있다. 과거에는 필요한 경우에만 동료에 의해서 연구평가가 수행되었지만 최근에는 일상적으로 이루어질 뿐만 아니라 계량적인 평가지표에 크게 의존하고 있다[1]. 이 때 연구평가가 데이터 기반으로 이루어지고 심사자의 판단이 부재하다는 점은 문제이다. 이러한 상황에서 연구성과 관련 계량적 평가지표들이 빠르게 확산되고 있다. 올바른 연구성과평가를 위한 목적으로 평가지표가 개발되고 있으나, 개발된 평가지표들이 충분히 이해되지 못하거나 때론 잘못 사용하기도 한다. 계량적 지표에 근거한 연구성과평가가 점점 더 확산되고 있지만, 많은 기관에서는 실제로 어떻게 수행하고 해석해야 하는지에 대한 충분한 지식이나 도움이 부재한 상황이다. 연구성과평가 체계를 향상시키기 위해 만든 도구인 평가지표로 인해 도리어 평가 체계 자체가 위협받고 있다.

2000년도 이전에는 Institute for Scientific Information (ISI)이 Science Citation Index를 CD-ROM 형태로 발간하였으며 전문가의 분석 업무에 사용되었다. 2002년에는 톰슨 로이터스가 통합 웹 플랫폼을 발족하여 Web of Science 데이터베이스의 광범위한 접근을 가능하게 하였다. 엘스비어의 Scopus(2004년 발표)와 Google Scholar(2004년 베타버전 발표)의 등장으로 인용색인 데이터베이스가 다양화되었다. Web of Science 기반의 InCites와 Scopus 기반의 SciVal과 같이 기관 단위의 연구 생산성과 영향력을 쉽게 비교할 수 있는 웹 기반 도구들이 소개되고, Google Scholar를 활용하여 개인 저자 단위 인용 프로필을 분석하는 소프트웨어(Publish or Perish, 2007년 발표)도 등장하였다.

2005년에는 미국 캘리포니아 주립대학교 샌디에고 캠퍼스에 재직중인 물리학자 Jorge Hirsch가 h-지수를 제안하여 개별 연구자의 인용 평가를 대중화했다. 1995년 이래로 저널영향력지수에 대한 관심은 꾸준히 증가하고 있다('저널영향력지수에 대한 집착(impact-factor obsession*)' 참조).

최근 연구성과의 사회적인 이용과 온라인 상의 의견을 포괄하는 평가지표가 점차 중요해지고 있다. 그 예로 F1000Prime(2002년 설립), Mendeley(2008년 설립), Altmetric.com(Nature Publishing Group를 가지고 있는 Macmillan Science and Education의 후원 아래 2011년 설립)을 들 수 있다.

과학계량학자, 사회과학자, 연구관리자로서 우리는 만연해있는 평가지표의 오용 정도가 심

각해지고 있음을 목도해왔다. 세계의 대학들이 국제 대학 순위(상하이 랭킹, Times Higher Education 리스트 등)에서의 순위에 집착하는 것은 평가지표 오용의 수많은 예 중 하나이다. 심지어 우리 관점에서 볼 때 이 순위들은 부정확한 데이터와 편협하고 비논리적인 지표에 기반한 것으로 보이지만 세계 대학들은 이에 집착하고 있다.

일부 인사담당자는 임용지원자에게 h-지수를 요청하기도 한다. 몇몇 대학들은 일정 값 이상의 h-지수와 영향력지수가 높은 저널에 실린 논문의 수에 따라 승진을 결정한다. 연구자의 CV는 이러한 지표 값을 내세우는데 사용되어 왔으며 특히 의생물학 분야에서 이러한 경향이 두드러진다. 지도교수는 아직 준비가 채 되지 않은 박사과정 학생에게 영향력지수가 높은 저널에 논문을 게재하고 외부 연구비를 수주하도록 요구하는 일이 만연하고 있다.

스칸디나비아와 중국의 몇몇 대학들은 획일화된 지표 값에 따라 개별 연구자에게 연구비나 보너스를 지급한다. 예를 들어, 개별 연구자의 영향력 점수를 계산하여 연구비를 배정하거나 저널영향력지수 값이 15보다 높은 저널에 논문을 게재한 경우 연구자에게 보너스를 지급하는 식이다 [2].

대부분의 경우 연구자와 평가자는 여전히 균형 잡힌 연구성과평가를 위해 노력하고 있다. 그러나 연구성과 평가지표의 오용은 간과하기 어려울 정도로 광범위하게 이루어지고 있다.

이러한 상황 때문에 **라이덴 선언(Leiden Manifesto)**을 발표하게 되었다. 이 이름은 라이덴 선언이 구체화된 학회 개최지(http://sti2014.cwts.nl 참조)를 따랐다. 라이덴 선언의 열 가지 원칙은 과학계량학자에게는 새롭지 않지만 지금까지 성문화되지 않았기 때문에 명확하게 제시하기 어려웠다. ISI의 창립자인 Eugene Garfield와 같은 이 분야의 권위자들은 이 원칙들 중 몇 가지를 공식적으로 언급한 적이 있다[3,4]. 그러나 이러한 권위자들의 권고사항을 대학 행정 담당자가 잘 인지하고 있어서 연구성과평가보고서를 살펴볼 때 고려할 것이라고 기대하기는 어렵다. 왜냐하면 이들은 성과평가 방법론의 전문가가 아니기 때문이다. 또한 평가 대상이 된 연구자가 평가에 대한 이의를 제기하고자 할 때 이러한 원칙들이 전문적인 저널에 흩어져 있어 접근이 어렵다.

우리는 계량 지표 기반 연구성과평가에 벤치마킹이 될 만한 방법론으로서 열 가지 원칙을 제안한다. 이를 통해 연구자는 평가자를 이해할 수 있고 평가자는 계량적 평가지표를 이해할 수 있다.

열 가지 원칙

1. **정량적 평가는 정성적 평가와 전문가 평가를 지원하는데 사용되어야 한다.** 정량적 평가지표는 동료평가에서 발생할 수 있는 편향성에 대한 이의 제기와 검토를 용이하게 한다. 일련의 관련 정보 없이 동료에 대한 평가를 내리는 것은 어렵기 때문에 정량적 평가는 동료평가에서 활용되어야 한다. 평가자는 연구성과에 대한 전반적 평가 관련 의사결정을 계

량지표 값 자체로 대체해서는 안 된다. 평가지표가 평가자의 숙련된 판단을 대체할 수 없으며, 평가자는 평가지표에 자신의 평가에 대한 책임을 전가해서는 안 된다.

2. **기관, 연구 집단, 연구자의 목표에 따라 성과를 측정하라.** 목표가 처음부터 명시되어야 하며, 성과평가에 사용되는 지표는 그 목표와 분명하게 연관되어야 한다. 평가지표의 선택과 적용방법은 사회/경제적, 문화적 맥락을 충분히 고려하여야 한다. 연구자의 목표는 다양하다. 선구적 학술 탐구 연구는 사회 문제 해결에 중점을 두는 연구와는 다르다. 그리고 리뷰는 학문적 아이디어의 우수성보다 정책, 산업, 공공의 요구에 기반을 둘 수 있다. 따라서 하나의 평가 모델이 모든 맥락에 동일하게 적용될 수 없다.

3. **지역적으로 가치 있는 연구의 우수성을 인정하라.** 세계 도처에서 영어로 쓰여진 출판물이 곧 우수한 연구로 간주되고 있다. 예를 들어, 스페인 법은 스페인 연구자가 영향력이 높은 저널에 논문을 게재하는 것을 권고한다**. 저널영향력지수는 미국에서 출반된 영문저널 중 Web of Science에 색인된 저널들을 기반으로 산출된다. 이러한 편향은 지역적, 국가적인 요구에 기반한 연구가 주를 이루는 사회과학과 인문학에서 특히 문제가 된다. 또한 사하라사막 이남 아프리카의 HIV 전염병학처럼 인문사회학이 아닌 많은 다른 분야도 국가적, 지역적 가치를 지닌다.

   연구의 다양성과 사회적 연관성은 저널영향력지수 값이 높은 영문 저널에 논문을 출판하려는 연구 경향에 의해 그 성장이 억제되고 있다. Web of Science에서 인용되는 스페인 사회학자들은 추상적 모델 연구나 미국의 데이터를 대상으로 연구한다. 지역 노동법, 노령자를 위한 가정의료, 이민노동자와 같은 지역 특수성을 반영한 연구를 하는 사회학자들의 이름은 Web of Science에 색인된 주요 스페인어 저널에서도 자취를 감추었다[5]. 수준 높은 비영어 학술문헌에 기반한 계량적 평가지표가 제공된다면 지역적으로 가치 있는 연구의 우수성을 규명하고 적절히 평가하는데 기여할 수 있을 것이다.

4. **데이터 수집과 분석 기법은 공개되어야 하며 투명하고 단순해야 한다.** 평가를 위한 데이터베이스 구축은 확실하게 명시된 규칙을 따라야 하며 그 규칙은 평가를 완료하기 전에 제시되어야 한다. 이 과정은 수 십 년 동안 계량서지학적 평가방법을 개발한 학계와 영리단체가 따른 통상적 관례이다. 그리고 이 관례는 동료심사제를 거친 문헌에 게재된 프로토콜이다. 이러한 투명성은 엄정한 평가를 가능하게 했다. 예를 들어, 2010년 우리 그룹 중 하나인 네덜란드 라이덴 대학교 과학기술연구센터(the Centre for Science and Technology Studies at Leiden University in the Netherlands)에서 사용하는 주요 평가지표 계산 방법을 공개토론을 통해 수정하였다[6]. 이 분야에 새로이 진입하는 영리단체는 이러한 규정을 준수해야 한다. 불투명한 블랙박스 같은 연구성과평가 과정은 수용될 수 없다.

   이해하기 쉬운 명확하고 간단한 연구성과평가지표는 평가의 투명성을 높여준다. 그러나 단순화된 평가지표가 연구활동 전반을 반영할 수 없으므로, 왜곡의 소지가 있다(원칙 7 참조). 따라서 평가자는 연구 과정의 복잡성과 평가지표의 단순성 간의 균형을 반드시 유

지해야 한다.

5. **평가 대상자가 평가 데이터와 분석과정을 확인할 수 있도록 하라.** 연구성과 평가에 사용되는 데이터의 품질을 확실히 하기 위하여 평가에 관련된 모든 연구자는 연구성과가 정확하게 반영되었는지를 확인할 수 있어야 한다. 평가과정을 감독하고 관리하는 평가 담당자는 자체검증 또는 외부감사를 통하여 데이터의 정확성을 담보하여야 한다. 대학의 연구정보시스템(RIS)에는 데이터 검증 모듈이 포함되어야 하며, 이는 연구정보시스템 공급업체 선정의 기준이 되어야 한다. 정확한 고품질 데이터를 수집하고 처리하기 위해서는 시간과 재원이 필요하므로 이를 위한 예산 할당이 필요하다.

6. **학문 분야에 따른 출판과 인용관행의 다양성을 인정하라.** 최선의 방법은 여러 개의 연구성과 평가지표 세트를 만든 후 학문 분야별로 적절한 평가지표를 선택하도록 하는 것이다. 몇 년 전, 유럽의 역사학자들이 국내 동료심사 평가에서 상대적으로 낮은 점수를 받았다. 이는 그들이 Web of Science에 색인된 저널에 논문을 게재하기 보다 단행본을 출판했기 때문이다. 이들은 운 나쁘게도 소속이 심리학과였다. 역사학자들과 사회과학자들은 단행본과 자국어로 쓴 문헌을 그들의 출판물 수에 포함시켜 줄 것을 요구하고, 컴퓨터 과학자들은 학술대회 논문을 출판물 수에 포함시켜 주길 요구한다.

   인용률은 학문 분야에 따라 다르다. 수학 분야에서 상위저널의 영향력지수 값은 약 3 정도이고, 세포 생물학 분야에서 상위저널의 영향력지수 값은 약 30 정도이다. 정규화된 평가지표가 필요하며, 가장 안정적인 정규화 방법은 백분위(percentile)이다. 이 때 개별 논문은 해당 분야의 인용 분포에 따른 백분위 중 어디에 해당하느냐에 기반하여 가중치를 받게 된다(예를 들어, 상위 1%, 10%, 20%). 인용 평균에 기반하게 되면 한 두 개의 출판물에 의해 특정 대학의 순위가 크게 변동될 수 있으나, 백분위에 기반하게 되면 한두 논문에 의한 큰 순위 변동이 발생하지 않는다[7].

7. **개별 연구자 평가는 연구자의 전체 연구실적에 대한 정성적 판단에 기초하여야 한다.** 새로운 논문을 발표하지 않더라도 연구자의 경력이 오래될수록 h-지수가 높다. 또한 h-지수는 학문 분야에 따라 그 값이 다르게 나타나는데, 탁월한 생명과학자의 h-지수는 200, 물리학자는 100, 사회과학자의 경우 20-30으로 나타난다(원문의 ref 8 참조). h-지수는 데이터베이스에 따라서도 다르게 나타나는데, 예를 들면 컴퓨터공학 연구자들의 h-지수가 Web of Science에서는 약 10 정도로 산출되지만 Google Scholar에서는 20-30 정도로 산출된다[9]. 따라서 연구자의 논문을 읽고 판단하는 것이 특정 계량지표 값 하나에 의존하는 것보다 훨씬 바람직하다. 여러 연구자를 비교할 때도 개인의 전문지식, 경험, 활동, 영향력에 관한 다양한 정보를 고려하는 것이 최선의 방법이다.

8. **구체성 오류(misplaced concreteness)와 정확성 과신(false precision)에 주의하라.** 과학기술 지표는 개념적으로 모호하고 불분명한 경향이 있고 일반적이지 않은 엄격한 가설을 필요로 한다. 예를 들어 인용빈도가 가지는 의미는 오랫동안 논쟁의 대상이 되어 왔다.

그러므로 더욱 안정적이고 다원적인 평가를 위해서는 여러 개의 평가지표를 사용하는 것이 최선의 방법이다. 만약 오차 구간(error bar) 등을 사용하여 불확실성과 오류를 정량화할 수 있다면, 이를 평가지표 값과 함께 제공해야 한다. 이것이 불가능하다면, 평가지표 제공자는 적어도 정확성을 과신하지 말아야 한다. 예를 들어, 저널영향력지수는 동점을 피하기 위해 소수점 이하 셋째 자리까지 공개된다. 그러나 인용빈도의 개념적 모호함과 임의적 변동성을 고려해볼 때, 매우 작은 영향력지수 값의 차이를 가지고 저널들을 구분하는 것은 의미가 없다. 정확성을 과신하지 마라. 소수점 첫째 자리까지가 적정하다.

9. **평가지표와 평가가 연구환경에 미칠 수 있는 영향을 염두에 두라.** 평가지표는 보상체계를 통해 연구환경에 영향을 미칠 수 있으므로 예상 가능해야 한다. 따라서 어떤 경우라도 여러 개의 평가지표를 사용하는 것이 바람직하다. 단일 지표를 성과평가에 사용하게 되면 연구환경이 혼탁해지거나 평가지표 값을 높게 받는 것 자체가 연구의 목적이 되는 상황을 초래할 수 있다(즉, 평가지표 자체가 연구 목표가 되어 버리는 것이다). 한 예로 1990년대 호주 정부는 연구 기관에서 발표한 논문의 수에 크게 의존한 평가지표를 사용하여 대학의 연구비 지원을 결정하였다. 호주 대학들은 논문의 '가치'를 이 논문이 실린 학술지에 근거하여 연구비로 환산하여 계산할 수 있었다. 예를 들어, 2000년을 기준으로 한 논문의 연구비 가치가 호주달러로 800불(미국화폐로 약 480불)인 것으로 나타났다. 예상대로 호주 연구자가 발표한 논문의 수는 증가하였다. 하지만 그 논문들은 인용이 많이 되지 않는 저널에 실렸다. 즉, 논문의 수준은 하락하였다고 볼 수 있다[10].

10. **정기적으로 평가지표를 철저히 검토하고 개정하라.** 연구목표와 평가의 목적은 변화하고, 연구 시스템은 그와 함께 진화한다. 한때 유용했던 평가지표들이 부적절해질 수 있으며, 새로운 지표들도 등장한다. 연구성과 평가지표 시스템을 검토해야 하고 필요한 경우 개정해야 한다. 앞서 언급한 지나치게 단순화된 연구성과 평가 방식이 연구체계에 미친 영향을 알게 된 호주는 2010년에 연구 품질을 강조하는 다원화된 Excellence in Research for Australia 이니셔티브를 도입하였다.

향후 전망

위의 열 가지 원칙 준수를 통해 연구성과평가는 과학의 발전과 과학과 사회간의 상호작용에 있어서 중요한 역할을 수행할 수 있을 것이다. 연구성과평가지표는 개인의 전문지식으로는 수집하거나 이해하기 어려운 중요한 정보를 제공할 수 있다. 그러나 평가지표가 제공하는 정량적 정보가 연구성과평가 도구에서 연구의 목적으로 바뀌어서는 안 된다.

견고한 통계 기법과 함께 평가 대상이 되는 연구의 목적과 본질을 신중하게 고려하였을 때 최선의 연구성과평가가 가능하다. 평가를 위해서는 정량적인 근거와 정성적인 근거 모두 필요하며, 이 둘은 각각의 방식으로 객관성을 지닌다. 학문에 관한 의사결정은 양질의 데이터를 바탕으로 한 양질의 평가 과정에 기초하여야 한다.

참고문헌

1. Wouters, P. in Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact (eds Cronin, B. & Sugimoto, C.) 47–66 (MIT Press, 2014).

2. Shao, J. & Shen, H. Learned Publ. 24, 95–97 (2011).

3. Seglen, P. O. Br. Med. J. 314, 498–502 (1997).

4. Garfield, E. J. Am. Med. Assoc. 295, 90–93 (2006).

5. López Piñeiro, C. & Hicks, D. Res. Eval. 24, 78–89 (2015).

6. van Raan, A. F. J., van Leeuwen, T. N., Visser,

M. S., van Eck, N. J. & Waltman, L. J. Informetrics 4, 431–435 (2010).

7. Waltman, L. et al. J. Am. Soc. Inf. Sci. Technol. 63, 2419–2432 (2012).

8. Hirsch, J. E. Proc. Natl Acad. Sci. USA 102, 16569–16572 (2005).

9. Bar-Ilan, J. Scientometrics 74, 257–271 (2008).

10. Butler, L. Res. Policy 32, 143–155 (2003).

한국어 버전 번역자 (Translators listed alphabetically by last name)

EunKyung Chung (정은경, Professor, Department of Library and Information Science, Ewha Womans University), Jae Yun Lee (이재윤, Professor, Department of Library and Information Science, Myongji University), Boram Lee (이보람, PhD Candidate, Department of Library and Information Science, Ewha Womans University), and So Young Yu (유소영, Assistant Professor, Department of Library and Information Science, Hannam University)

---

* 역주1. 원문 431 페이지에 있는 그림 참조

** 역주2. Jiménez-Contreras, E., et al. (2002) Impact-factor rewards affect Spanish research, Nature, 417(27 June 2002), 898. < doi:10.1038/417898b>