Health warning: Might contain multiple personalities The problem of homonyms in Thomson Reuters Essential Science Indicators

Anne-Wil Harzing

Version August 2015

Accepted for Scientometrics

Prof. Anne-Wil Harzing
Middlesex University
The Burroughs, Hendon
London NW4 4BT
Email: anne@harzing.com
Web: www.harzing.com

Health warning:
Might contain multiple personalities

The problem of homonyms in Thomson Reuters Essential Science Indicators

Anne-Wil Harzing

Middlesex University
The Burroughs, Hendon, London NW4 4BT
Email: anne@harzing.com, Web: www.harzing.com

Abstract

Author name ambiguity is a crucial problem in any type of bibliometric analysis. It arises when several authors share the same name, but also when one author expresses their name in different ways. This article focuses on the former, also called the "namesake" problem. In particular, we assess the extent to which this compromises the Thomson Reuters Essential Science Indicators (ESI) ranking of the top 1% most cited authors worldwide. We show that three demographic characteristics that should be unrelated to research productivity – name origin, uniqueness of one's family name, and the number of initials used in publishing – in fact have a very strong influence on it.

In contrast to what could be expected from Web of Science publication data, researchers with Asian names – and in particular Chinese and Korean names – appear to be far more productive than researchers with Western names. Furthermore, for any country, academics with common names and fewer initials also appear to be more productive than their more uniquely named counterparts. However, this appearance of high productivity is caused purely by the fact that these "academic superstars" are in fact composites of many individual academics with the same name. We thus argue that it is high time that Thomson Reuters starts taking name disambiguation in general and non-Anglophone names in particular more seriously.

Keywords

Author disambiguation; homonyms, Essential Science Indicators, research productivity

Health warning:
Might contain multiple personalities

The problem of homonyms in Thomson Reuters Essential Science Indicators

Background

Author name ambiguity is a crucial problem in any type of bibliometric analysis. It arises when several authors share the same name, but also when one author expresses their name in different ways. This article focuses on the former, also called the "namesake" problem. In particular, we assess the extent to which this compromises the Thomson Reuters Essential Science Indicators (ESI) ranking of the top 1% most cited authors worldwide. ESI is a database that is part of the Thomson Reuters Web of Knowledge, which also includes better-known databases such as the Web of Science (including, amongst others, the (Social) Science Citation Index) and the Journal Citation Reports (JCR, source of the well-known Journal Impact Factor).

Thomson Reuter's Web of Knowledge is the oldest – and, in the eyes of many, still the authoritative – provider of bibliometric data. One would therefore expect Thomson Reuters to apply the most sophisticated methods of author disambiguation (for recent developments as well as a comprehensive review of methods see Shin, Kim, Choi & Kim, 2014; Wu, Li, Pei & He, 2014; Zhu, Yang, Xie, Wang & Hassan, 2014). However, after a quick perusal of the top-100 most cited academics, we consider it highly unlikely that Thomson Reuters has applied author disambiguation effectively.

The top-100 most cited academics are a surprisingly homogenous group: 68 of them have Chinese family names and 24 have Korean names, with the remaining 8 having Japanese or Indian names. Of the top-100 most cited academics, well over half are called Wang, Zhang, Li, Kim or Lee, which – not coincidentally – are also the most frequently occurring Chinese and Korean family names. Although Chinese and Korean academics have dramatically increased their share of the world production of papers in the last 10 years, it is hard to believe they now make up more than 90% of the world's most highly cited academics, leaving all their Western colleagues behind.

On average these top-100 most cited academics have gathered over 135,000 citations each and published 11,465 papers each in just 10 years. Hence, on average each of these super-authors managed to publish more than three papers *a day*, every single day, for a decade. This is clearly not a feasible proposition unless these "academic superstars" are in fact composed of multiple individuals. The Chinese expression 张三李四 ("three Zhang, four Li", meaning "anyone" or "just everybody") seems to be particularly appropriate here. To be fair, Thomson Reuters does acknowledge the namesake problem in the ESI helpfile under the heading "Name conflation":

> "*Scientists having the same last name and initials may represent multiple individuals. This is especially likely in the case of common surnames. The ability to break out the name by field may to some degree disambiguate person X in field Y from person X in field Z; however, keep in mind that a listed name can still represent more than one scientist within the same field.*"

In order to find this crucial caveat, however, an ESI user would first of all need to take the initiative to consult the helpfile, something most users of computer programs and websites rarely do. The likelihood of any user consulting the helpfile is further reduced by the location of the link to the helpfile. Rather than displaying a prominent link in the results area, the link is shown in a very small font size at the top right hand of the page, next to the language choice, not an area that draws immediate intention. Second, even if the ESI user *did* consult the helpfile, they would need browse and read it systematically. The caveat is not listed until the last section of the helpfile named "Citation thresholds", not a section a naïve reader would be likely to consult to find out more about name disambiguation. Finally, even if the ESI user would somehow manage to discover the caveat, its wording is so "mild" that readers would be forgiven for interpreting it simply as a minor footnote.

We argue, however, that more than a minor caveat is needed. Even when moving further down the ESI ranking of the most highly cited academics, we only spotted an occasional non-Asian name. These rare non-Asian names were usually academics of the calibre of Nobel Prize winners such as Andre Geim. Ranking the list of highly cited academics by the number of papers published, rather than the number of citations, made the problem even worse. In contrast, sorting the ESI ranking by the number of citations per paper makes Asian names vanish almost completely from the top of the ranking, which is now dominated by Anglo and European names. This seems to strongly suggest that any highly cited Asian academics were in fact merged with many of their lowly cited namesakes. Strotmann & Zhao (2012) noted the same problem in their study of the top-200 most cited authors in stem cell research and Heeffer, Thijs & Glänzel (2013) had to exclude Chinese academics from their study because of the namesake problem.

As Strotmann & Zhao explain, the preponderance of the namesake problem for Chinese and Korean academics is not entirely surprising as these countries have much less variety in terms of family names than most other countries do. Two dozen Chinese names account for nearly half of the Chinese population and nearly a quarter of the Chinese population is called Wang, Zhang or Li (https://en.wikipedia.org/wiki/ Chinese_surname). In Korea, the potential for namesake problems looms even larger as just three family names (Kim, Lee and Park) account for nearly half of the population (https://en.wikipedia.org/wiki/Korean_name). Interestingly, the opposite is true for given names, which – unlike most given names in Western countries – tend to be unique, as each name is chosen individually.

So, in fact, the *full* names of East Asian academics *are* unique when written in their respective character based languages. However, when East Asian names are romanized, many different ideographic names are mapped onto the same romanized name. Combined with Thomson Reuter's unfortunate choice to use only initials, rather than full given names, this leads to an enormous potential for erroneous aggregration of individual academics with the same family name. As the number of papers published by Chinese and Korean academics is still increasing rapidly[1], we can expect this problem to only get worse in the near future. In this paper, we therefore decided to investigate the current scale of the namesake problem in the Essential Science Indicators in more detail.

[1] The increase in the number of papers published between 2005-2009 and 2010-2014 is 98% for China, 57% for Korea and only 17% for the USA (Essential Science Indicators, May 2015).

Methods

Data

The data for this paper consist of the top 1% most highly cited academics as listed in the Thomson Reuters Essential Science Indicators (ESI). This database is updated every two months. We used the 7 May 2015 version of the data, which covers 10 years + 2 months (1 January 2005 - 28 February 2015). In our analyses we used both the overall ranking, which includes nearly 83,000 academics, and four subject specific rankings: Chemistry (9,172 academics), Clinical Medicine (19,738 academics), Economics & Business (1,568 academics) and Physics (6,064 academics). Economics & Business was subsequently chosen for in-depth analysis for two reasons. First, the relatively small absolute number of highly cited academics in this discipline allowed us to review each and every individual. Second, Economics & Business is the author's home discipline; focusing on a familiar discipline allows for a more reliable assessment of the namesake problem. Chemistry, Clinical Medicine and Physics were included mainly as contrasting disciplines. As the four key scientific areas in which Nobel Prizes are awarded these four disciplines have also been subject to prior bibliometric analysis (e.g. Harzing, 2013a).

Coding

The Essential Science Indicators ranking includes the top 1% most highly cited academics. However, as it is easier to assess the extent of the namesake problem by looking at the most productive academics (i.e. those who produced the largest number of papers), we sorted the respective rankings by the number of papers instead of the number of citations. For the overall ranking we subsequently coded the name origin (e.g. Chinese, Korean, Anglo) of the top-1000 academics with the largest number of published papers, which equated to 1.2% of the total number of authors. We did the same for any academic who had published more than 1000 papers in the 10-year period. Finally, we coded the names of all academics in the list that shared their family name with at least 4 others in the list. So the origin of *all* names that appeared at least 5 times in the list were coded. As a result, one third of the nearly 83,000 names were coded, indicating that namesakes were a very frequent occurrence.[2] Analogous to the overall rankings we coded the top 1.2% most productive academics for the discipline specific rankings in Chemistry (110 academics), Clinical Medicine (238 academics) and Physics (73 academics), whereas for Economics & Business we coded all academics.

We coded names as Anglo, Chinese (incl. variants used in Singapore, Hong Kong, Taiwan, Malaysia etc.), Japanese, Korean, Indian, and European (Germanic, Dutch, Italian, French, Hispanic, Swedish, Danish)[3]. Name origin coding was based purely on the academic's name, using the author's in depth knowledge of naming practices in different countries (see e.g. Harzing, 2001) and a variety of web-based tools. Names that were ambiguous were left unclassified. Obviously, the origin of an academic's name does not necessarily coincide with their nationality; someone with a Chinese name for instance might well have British, Australian, Malaysian or any other nationality. However, our interest in this paper is not in nationality as such. We simply try to assess the extent to which different name origins suffer from the namesake problem.

[2] Obviously this understates the extent of the occurrence of namesakes as there were many cases where the same name occurred only 2-4 times.
[3] Names originating in other European countries did not occur in multiples of 5 or more.

Results

After a brief illustration of the namesake problem through a review of the world's most productive researchers, we conduct three separate tests to establish the extent of the problem in the database as a whole. Subsequently, we drill down to the disciplinary level to establish whether this eliminates the namesake problem.

Y Wang: the world's most productive academic

Looking at the three most productive academics in the ESI ranking provides a first indication of the extent to which the namesake problem distorts the underlying reality. Y Wang, Y Zhang and Y Li make every other academic look like a slacker. Each has published around 30,000 papers in just 10 years, for an average of nearly 9 papers a day. Our most productive academic, Y Wang, is a true homo universalis; he/she published more than 100 papers in *each* of 73 distinct research areas, ranging from business economics to computer science, biochemistry to astronomy, oncology to materials science, and toxicology to mathematics. Y Wang is affiliated with more than 500 different universities in nearly 100 different countries.

However, when we look at the actual list of publications, we discover the secret behind this incredible productivity. Y Wang suffers from a serious case of multiple personalities. Even looking at the last 10 papers he/she published, we discover that the initial Y covers six different given names: Yang, Yi, Yuan, Ying, Yan, and Yu. The problem doesn't stop there, however. Every further refinement produces a new case of multiple personalities. Searching for Yang Wang results in some 2,500 papers, still a very respectable 5 papers a week. Yang Wang also manages to hold multiple appointments; in fact he/she is affiliated with more than 100 universities. Even more interestingly, Yang Wang holds many appointments within the **same** university. At Fudan University, he/she is affiliated with, to name just a few, the Department of Mechanical & Engineering Science, the Institute of Genetics, the Department of Neurosurgery, the Department of Anatomy, the Department of Urology, the Department of Pharmacology, the Department of Radiology & Oncology, the Department of Macromolecular Science and the School of Public Health.

It is almost impossible to establish how many academics called Y Wang were amalgamated to create this one academic superstar. However, based on our investigation of Yang Wang alone, we estimate Y Wang is likely to include thousands of academics. When looking at the lesser mortals amongst the no less than 321 highly cited authors called Wang, the namesake problem still raises its ugly head. Even the seemingly uniquely named CCL Wang, one of only eight Wangs who has published *less* than 100 papers, is the amalgamation of two different individuals: Charlie CL Wang and Cecilia CL Wang. Hence, we consider it very likely that most, if not all, of the academics called Wang (or Zhang or Li or Kim or Lee) included in the ESI highly-cited list are in fact amalgamations of multiple academics.

Some common sense tests for author name ambiguity

Obviously, we cannot review all 83,000 academics in the ESI highly cited database to assess the extent of the namesake problem. However, we suggest three separate tests to provide an approximation. All three tests look at characteristics of individual academics' names that should, in theory, not have any systematic influence on one's research productivity: national origin of the name, the uniqueness of one's family name, and the number of initials used when publishing.

1. Does your nationality determine your research productivity? Yes, but not in the way you expect!

The second column of Table 1 presents the relative share of the main countries/country groupings in terms of publications in the Web of Science. In the Anglo category we combined papers from academics affiliated with universities in the USA, Canada, UK, Australia, New Zealand and Ireland.[4] In the European category, we aggregated papers published from the EU-25 plus Russia, Turkey and Switzerland. The Chinese category included papers published from Mainland China, Taiwan, and Singapore. As expected, academics affiliated with Western universities make up a large share, nearly 70% in fact, of the papers published in the WoS. We would therefore expect them to make up a substantial proportion of the top most productive academics as well.

Table 1: Representation of academics with different name origins

Name origin	% of WoS papers at country level	% of top 1,000 ESI academics	% of ESI academics with 1,000+ papers
Anglo	35.3%	1.1%	3.5%
Chinese	11.6%	62.8%	58.1%
European	33.5%	0.8%	3.7%
Indian	2.7%	5.2%	5.8%
Japanese	5.0%	15.4%	17.9%
Korean	2.6%	14.7%	10.5%
Rest of the world	9.5%	0.0%	0.0%

However, as Table 1 shows this is not the case at all. Academics with Anglo or European names make up less than 2% of the top 1,000 most productive academics and just over 7% of the academics with more than 1,000 published papers. In contrast academics in Asian countries, who contribute only 22% of the number of papers published in the WoS, make up more than 98% of the top 1,000 most productive academics and nearly 93% of the academics with more than 1,000 papers.

Hence, there is a very strong discrepancy between the representation of (mainly East) Asian countries in the overall world publication output and the representation of academics with Asians names amongst the most productive academics worldwide. This provides us with a strong indication that the namesake problem might be distorting the Essential Science Indicators ranking. In the next two sections we'll investigate this further by looking at two other name characteristics: the uniqueness of one's family name and the number of initials used in publishing.

2. Does your last name determine your research productivity? Yes it does!

The nationality linked to a particular name origin could still be argued to influence research productivity, though more likely in the opposite direction to the results we found above. However, one would be hard pushed to make an argument that the (lack of) uniqueness of one's family name should influence one's research productivity. Our second test thus compares academics that share their name with at least four other academics

[4] Classifying papers by country of affiliation obviously does not provide an identical result compared with classifying authors' names by the origin of their name as there are many Asian academics working at Western universities. However, the effects discussed in this paper are so large that it is unlikely that this limitation negates our results.

with those that do not. As Table 2 shows, the former on average produce four times as many papers, and have twice as many citations as academics with more unique names. On average academics with a common family name publish nearly one paper a week.

Table 2: Productivity and citation levels of academics with common vs. more unique names

Nationality	# of Papers	# of Cites	Cites per Paper
Uncoded (Names occurring < 5 times)	115	3509	64.3
Coded (Names occurring > 5 times)	450	7074	31.6
Coded name origins	**# of Papers**	**# of Cites**	**Cites per Paper**
Anglo	169	4685	43.0
European	229	5198	37.5
Indian	598	8243	25.6
Chinese	762	9330	18.6
Japanese	930	12802	17.0
Korean	1118	13332	16.5

When disaggregating the coded names by name origin, academics with common names of Anglo origin seem to be least productive, closely followed by their European counterparts. However, these academics are still 1.5 to 2 times as productive as academics with more unique names. Academics with common names of Indian, Chinese, Japanese or Korean origin appear to be between five and ten times as productive as those with more unique names. It is thus very likely that many, if not most, of the academics with more common family names are in fact composed of multiple academics. As a result, the citations per paper for these "individuals" are substantially lower than for academics with more unique names for the simple reason that the former's citation performance is deflated by amalgamation of namesakes with low citation rates.

As indicated above, China and Korea have particularly concentrated naming practices for family names: a quarter of the Chinese are called Zhang, Wang and Li and half of the Koreans are called Kim, Lee or Park. Hence for our sample of coded Chinese and Korean names we compared those Chinese and Koreans that had one of their country's three most common names with counterparts that were the bearer of slightly less common name. As table 3 shows any academic called Li, Zhang or Wang has on average published some 200 articles a year, i.e. nearly four articles a week. Academics called Lee, Kim or Park are only slightly less productive with three articles a week. Around a quarter of both groups are in the top 1,000 most productive academics worldwide and around half have published more than 1,000 papers in the 10-year period. The namesake problem is thus clearly even more prominent for common Chinese and Korean names than it is for Asian names in general.

Table 3: Research productivity for the three most common Chinese and Korean names

Name	# of Papers	# of Cites	% of top 1,000 ESI academics	% of ESI academics with 1,000+ papers
Zhang, Wang, Li	2007	21780	28.1%	53.8%
Other Chinese names	612	7838	5.1%	15.0%
Kim, Lee, Park	1642	19210	24.5%	45.5%
Other Korean names	484	6207	2.9%	11.7%

Looking at the names of European origin in more detail, we found that, with an average of 146 papers and 3714 citations, academics with a commonly occurring Dutch name differ least from the group of uncoded academics (i.e. those with more unique names). No doubt this is due to the much higher tendency of Dutch parents – especially Catholic parents – of blessing their children with three or more given names. Coded Dutch academics in our sample were six times more likely than the average to have 3 initials and twelve times more likely to have 4 initials; none of the other nationalities even came close. Hence, even if Dutch academics have a common family name, they are likely to distinguish themselves by their initials. This brings us to our third final test: does the number of initials you publish with influence your research productivity?

3. Does the number of initials you publish with influence your research productivity? Yes it does!

Another naming characteristic that should be unrelated to an academic's productivity is the number of initials an academic customarily uses in their publications. The frequency of using multiple given names (and thus initials) differs substantially by country: most Japanese academics only have one initial, whereas most academics from neighbouring Korea have more than one. Hence we need to examine the effect of the number of initials on research productivity on a country-by-country level.

Table 4: Research productivity of academics with 1 initial vs. those with more initials

Name origin	# of Papers	# of Cites per paper	% of top 1,000 ESI academics	% of ESI academics with 1,000+ papers
Anglo 1 initial	289	31.9	0.4%	3.0%
Anglo >1 initial	120	47.7	---	---
Chinese 1 initial	1585	15.6	19.4%	38.8%
Chinese > 1 initial	556	19.4	4.6%	14.1%
European 1 initial	348	50.9	0.3%	3.6%
European > 1 initial	114	23.8	---	---
Indian 1 initial	820	16.7	8.7%	24.8%
Indian > 1 initial	274	38.4	0.8%	4.7%
Japanese 1 initial	975	14.9	9.9%	31.2%
Japanese > 1 initial	102	56.4	---	---
Korean 1 initial	1703	15.5	21.6%	38.3%
Korean > 1 initial	1001	16.7	13.3%	26.7%
Unclassified 1 initial	133	61.2	---	---
Unclassified > 1 initial	88	69.1	---	---

Table 4 shows that academics who have common family names (i.e. those that occur at least five times in the ESI database) and are publishing with only one initial, have published substantially more papers, are more likely to be amongst the top 1,000 most productive academics, and are more likely to have published more than 1,000 papers. They also have fewer citations per paper than their counterparts who also have a common family name, but publish with more initials. This is true for any name origin, whether it is Anglo, European, Chinese, Indian, Japanese or Korean, although the difference is largest for the Japanese. The rare Japanese with more than one initial on average published only a tenth as much as those with the traditional single initial. Nearly a third of the Jap-

anese with a single initial have published more than 1,000 papers, whereas none of the Japanese with more than one initial reached this high level of productivity. Standing out seems to have a seriously adverse effect on one's research productivity in Japan! However, before we start constructing an elaborate psychological theory about why standing out in a collectivist society would make an academic less productive, we should consider the common sense alternative: academics with a common family name and only one initial are more likely to be confused with their namesakes. Table 4 shows that this is even the case for academics with less common family names. Even for those academics whose names we did not classify as their family name occurred only 1-4 times in the ESI database, having only one initial "buys" them an additional 45 papers over a 10-year period.

Drilling down to a disciplinary level

Drilling down to a disciplinary level should diminish the namesake problem, as academics that share the same name, but work in different disciplines, would automatically be separated. We therefore looked at the top 1.2% (analogous to the top-1,000 academics in the overall ranking) most productive academics in Chemistry, Physics, Medicine and Economics & Business. As Table 5 shows, this did not change the dominance of Asian names amongst the most productive academics. In fact, in Chemistry a full 86% of the top 1.2% most productive academics were of Chinese origin. Only in Economics & Business did any Anglo names make it to the top, and European names were absent in all of the four disciplines.

Table 5: Different name origins in the top 1.2% for different disciplines

Discipline (n)	Chinese	Korean	Japanese	Indian	Anglo	European
Chemistry (110)	86%	11%	1%	2%	0%	0%
Physics (73)	73%	18%	5%	4%	0%	0%
Medicine (238)	41%	24%	31%	4%	0%	0%
BusEco (20)	47%	42%	0%	5%	5%	0%
Total (top-1000)	63%	15%	15%	5%	1%	1%

For Economics & Business we subsequently coded the name origin of every name in the ESI ranking. As Table 6 shows European and Anglo names clearly dominate in Economics & Business, making up more 80% of the names that could be coded. However, they are underrepresented in the top-100 most productive academics, making up only 43% of this group. The other extreme is represented by academics with Korean and Chinese names that make up less than 12% of the total sample, but represent 47% of the most productive academics in Economics & Business.

Table 6: Representation of different name origins in Economics & Business

Coded name origins	# of Papers	Cites per Paper	% of total ESI academics	% of top-100 academics
Anglo (543)	20	37.9	40%	25%
European (560)	20	37.7	41%	18%
Indian (113)	23	35.2	8%	10%
Chinese (136)	38	20.9	10%	34%
Korean (22)	77	13.9	2%	13%

Name conflation in the top-100 and a closer look at Korean and Chinese academics

Reviewing the top-100 most productive academics in Economics & Business in detail, we found that more than half of them (54) were composed of multiple academics. This was true for *all* of the Korean names in the top-100, 88% of the Chinese names and 60% of the Indian names. In contrast, only 4 out of the 25 (16%) of the Anglo names and one of the 18 European names were composites. Given that Korean and Chinese names presented the biggest problem, we subsequently looked in detail at *all* 22 Korean names in the ESI ranking and all Chinese that were called Zhang, Wang or Li (in addition to those in the top-100).

Apart from one Bai, all Koreans in the Economics & Business ESI list are called Lee or Kim. Out of the 22 Koreans in the list, only 3 are **not** a composite of different academics. All three have two initials and work at a US university. Of the remaining 19, 12 have only one initial and have published between 38 and 243 papers. In most cases, even a casual inspection revealed at least 10 different given names and often many more. Of the seven academics with multiple initials, three contain at least 10 academics, and two more at least 5. The remaining two seem to represent "only" two individuals. Although obviously names with fewer papers are less likely to be multiple academics, names with as few as 13 publications contained 5 different academics.

Out of the 59 Chinese names (34 from the top-100 and a further 25 Zhang, Wang or Li's) only nine are not composites of multiple academics. The remaining 50 all contained multiple authors. In most cases, even a casual inspection revealed at least 10 different given names and often many more. Although the conflated individuals were more likely to have only one initial, 14 out of the 50 had two initials and several of these still contained more than 10 academics. Seven of the nine academics that were not conflated had multiple initials, only two were working in China. In sum, nearly all of the Korean names in the Business & Economics ESI ranking represent multiple academics, whereas this is likely to be the case for *at least* half of the Chinese academics[5].

Uniqueness of family name and the number of initials

Replicating the analyses we conducted for the aggregate ESI ranking, we investigated whether uniqueness of the family name and the number of initials impacted on research productivity. Given the much smaller sample size, there are far fewer names that are not unique. In fact, whereas in the total sample a third of the names occurred 5 times or more, in Business & Economics nearly 80% of the names are fully unique and most of the non-unique names occurred only 2 or 3 times. That said, there is still a very significant difference ($t=12.409$, $p = 0.000$) in research productivity between academics with unique names (average of 19 papers) and those with non-unique names (average of 33 papers). The group of academics with non-unique names also has a significantly higher ($t=10.054$, $p=0.000$) proportion appearing amongst the top-100 most productive academics: 18% vs. 3%.

We also coded the *number* of times that the name appeared in the ESI ranking and found a strong (0.453) and highly significant ($p=0.000$) correlation between the number of times a name appeared in the list and the number of papers an individual with that name had published. Although the strength of this correlation was largely driven by

[5] We investigated 43% of the Chinese names, of which 85% were conflated. Although conflation is less likely for the remaining, less common, Chinese names, these academics on average still produced significantly more papers than Anglo and European academics. Hence, we consider it very likely that many of them still contain multiple academics.

names that occurred more than 10 times (mostly Chinese and Korean names), academics that shared their name with only 1-3 other academics also published significantly (t=3.866, p=0.000) more papers than academics that did not have namesakes in the list.

Finally, Table 7 shows that academics that publish with only one initial appear to have a significantly higher productivity than those that publish with more than one initial. This is true for every country except India, where although the effect is in the expected direction, it is not significant.

Table 7: Research productivity of academics with 1 initial vs. those with more initials

Coded name origins	# of papers, academics with 1 initial	# of papers, academics with > 1 initial	t-value	Significance
Anglo (543)	22.7	18.2	-3.749	.000
European (560)	20.8	18.0	-2.596	.010
Indian (113)	23.7	21.9	-0.430	.668
Chinese (136)	48.5	26.7	-4.597	.000
Korean (22)	115.6	29.4	-4.378	.000

Overall, we therefore find that although the namesake problem is not as big in Economics & Business as it is in the overall ESI list, it is still prominently present. Moreover, we should realise that Economics & Business is one of the "smallest" disciplines and the discipline in which Asian names are least common. Hence, the namesake problem is likely to create a bigger distortion in all other Web of Science disciplines.

Discussion

In this paper we investigated the extent of the namesake problem in Thomson Reuter's Essential Science Indicators. We showed that three demographic characteristics that should be unrelated to research productivity – name origin, uniqueness of one's family name and the number of initials used in publishing – in fact have a very strong influence on it. In contrast to what would be expected from Web of Science publication data, researchers with Asian names – and in particular Chinese and Korean names – appear to be far more productive than researchers with Western names. For any country, academics with common names and fewer initials appear to be more productive than their more unique counterparts. However, this appearance of productivity is caused purely by the fact that many researchers in the ESI ranking with Asian names, as well as those with less unique name/initial combinations are in fact composites of many individual academics with the same name. Drilling down to the disciplinary level reduced the namesake problem, but by no means eradicated it. We looked in detail at the most conservative case, the discipline of Economics & Business, which includes a much smaller number of academics than other disciplines, as well as a substantially smaller proportion of academics with Asian names. Even here name origin, uniqueness of family name and the number of initials still significantly impacted on research productivity.

We showed that the namesake problem impacted on the accurateness of the ESI ranking for all nationalities. However, the use of a simplistic last-name-plus-initial(s) falls down completely in countries such as China and Korea, in which a very small number of family names account for a large percentage of the population and where the first name is essential to distinguish individuals (Strotmann & Zhao, 2012). As Qiu (2008) indicates this might have serious career limiting consequences for Asian academics. As it is difficult to

uniquely identify Asian authors, they are less like to be asked to be reviewers or editorial board members or to participate in collaborative research projects. This makes it hard for Asian academics to compete on an equal footing with academics with more unique names. Finally, in the context of the Essential Science Indicators ranking, the namesake problem also disadvantages academics with unique names, who cannot "compete" with the super-authors created by the amalgamation of many namesakes, and are thus ranked much lower in the list of highly cited authors than they should have been.

It is thus high time that Thomson Reuters starts taking name disambiguation and non-Anglophone names more seriously. We therefore fully endorse Strotmann & Zhao's (2012: 1830) comment that: "American and European information systems are lagging behind in their efforts to keep up with the changing demands on accurate author searching". Their suggestion of providing authors the opportunity of having full names listed in their original languages would seem an excellent solution. As they argue this is entirely feasible with modern information technology. Journals published by the American Physical Society already offer the option to include the author's name in their own language in brackets (see e.g. http://journals.aps.org/prl/PhysRevLett.99.230001).

We do acknowledge that Thomson Reuters has partnered with the Chinese Academy of Sciences to offer the Chinese Science Citation Database. However, this is a separate database that only covers Chinese journals and requires a separate subscription. What is urgently needed is not a "patch-up" by adding additional databases to cover "non-standard" publications. Science is a global enterprise and thus requires globally integrated coverage. We already argued before that Thomson Reuters seems to be misunderstanding the Social Sciences (Harzing, 2013b). In this paper we showed that Thomson Reuters also seems to have serious difficulty with non-Western names. Thomson Reuters' Anglophone, Science-based view of the world might well have been tolerable in the past, but it has long ceased to be acceptable in the 21st century.

References

Harzing, A. W. (2001). Who's in charge? An empirical study of executive staffing practices in foreign subsidiaries. *Human Resource Management*, 40(2), 139-158.

Harzing, A. W. (2013a). A preliminary test of Google Scholar as a source for citation data: a longitudinal study of Nobel prize winners. *Scientometrics*, 94(3), 1057-1075.

Harzing, A. W. (2013b). Document categories in the ISI Web of Knowledge: Misunderstanding the social sciences?. *Scientometrics*, 94(1), 23-34.

Heeffer, S., Thijs, B., Glänzel, W., (2008) Are Registered Authors More Productive? *ISSI Newsletter*, 2013, 9 (2), 29-32.

Qiu, J. (2008). Scientific publishing: Identity crisis. *Nature News*, 451(7180), 766-767.

Shin, D., Kim, T., Choi, J., & Kim, J. (2014). Author name disambiguation using a graph model with node splitting and merging based on bibliographic information. *Scientometrics*, *100*(1), 15-50.

Strotmann, A., & Zhao, D. (2012). Author name disambiguation: What difference does it make in author-based citation analysis? *Journal of the American Society for Information Science and Technology*, 63(9), 1820-1833.

Wu, H., Li, B., Pei, Y., & He, J. (2014). Unsupervised author disambiguation using Dempster–Shafer theory. *Scientometrics*, *101*(3), 1955-1972.

Zhu, J., Yang, Y., Xie, Q., Wang, L., & Hassan, S. U. (2014). Robust hybrid name disambiguation framework for large databases. *Scientometrics*, *98*(3), 2255-2274.

Zhou, P., & Leydesdorff, L. (2006). The emergence of China as a leading nation in science. *Research Policy*, *35*(1), 83-104.