The iSchool Community: A Case Study of iConference Reviews

Toine Bogers[1], Elke Greifeneder[2]
[1]Department of Communication & Psychology, Aalborg University Copenhagen, Copenhagen, Denmark
[2]Berlin School of Library and Information Science, Humboldt Universität zu Berlin, Berlin, Germany

Abstract
A fair review process is essential to the success of any scientific conference. In this paper we present an analysis of the reviewing process of the 2014-2015 iConferences as well as a demographic analysis of the iConference community as a whole. The results show a clear need for making the reviewer pool more representative of the iSchool community as a whole by including more women and more researchers from Asian institutions. Other recommendations are to improve the continuity of the reviewer pool and to provide clearer instructions to reviewers to ensure that written reviews explicitly cover all the aspects represented by the review scores. The results of our study provide the iSchool community with a descriptive analysis of its community and a better understanding of its review process.
Keywords: peer review, gender bias, cultural differences, iConference, iSchools, bias
doi: 10.9776/16247

Acknowledgements: We would like to thank Vera Hillebrand and Leyla Dewitz for their valuable help in collecting background information on the authors and reviewers. Similarly, we wish to express our thanks to Lihong Zhou and his team at Wuhan for their help with data collection about Asian authors and reviewers. This study was approved by the iCaucus executive committee.
Contact: toine@hum.aau.dk; greifeneder@ibi.hu-berlin.de

1 Introduction

A fair peer review process is an important part of any successful academic conference. Conference chairs aim to recruit informed and experienced reviewers and guide them in the review process through instructional letters. Yet even the best instructions in the world do not help if there is a fundamental difference between (sub)groups of reviewers in their cultural attitude towards reviewing and providing praise and criticism. Germans, for instance, are known to be reluctant with praise compared to Americans. If German reviewers always give lower ratings than their US counterparts, this would introduce a bias in the review process. Other imbalances in the reviewer corps could introduce additional biases.

Cultural biases gain in importance when conferences become more international. The two major information science conferences, ASIS&T and the iConference, have both seen a large influx of international submissions in recent years. The iSchool community in particular has changed tremendously: today, over half of all iSchools are located outside North America. However, the iConference's traditional double-blind peer reviewing process had only varied minimally up to 2013. In 2014, the paper chairs attempted to take internationalization into account by assigning each submission from non-North American (NNA) countries a reviewer from a NNA country as well, in order to allow for a better cultural match. The subjective impression afterwards, however, was that this harmed submissions from NNA countries, because reviewers from NNA countries appeared to give lower ratings on average than North American (NA) reviewers.

The aim of this work is to examine the iConference community and its submission reviewing process in more detail. We use the reviewing data from the 2014 (Berlin, Germany) and 2015 (Newport Beach, USA) iConferences in our analyses. Our contributions with this paper are three-fold:

- A demographic analysis of the iConference author and reviewer community.
- An analysis of the review scores assigned by the reviewers according to demographical features with the aim to identify any possible imbalances or biases in the iConference review process.
- A content analysis of the reviews, combined with a comparison to the official review scores assigned by reviewers and the final outcomes for each submission.

2 Background

The iSchools are a rapidly growing international organization of Library and Information Science schools. It was founded in 1988 as an informal collaboration and was expanded to five members in 2001 and to ten members in 2003. For the first twenty years of its existence, the iSchool community consisted only of

US-based institutions, which changed in 2008 when the University of Toronto joined and in 2009 when schools from three non-native English-speaking countries joined—Germany, China, and Denmark. By 2015, the iSchools group had increased to 65 members and has more non-US members for the first time in its history. This development has been reflected in the annual conference of the iSchools, the iConference, which has seen a large influx of non-US contributions and reviews.

In times of restricted research funding, having one's conference submission accepted becomes essential for receiving travel funding. The growth in iSchool members has also caused a growth in the number of iConference submissions, even though the full-paper acceptance rate has remained stable at around 35%. At the 2015 iConference in Newport Beach, approximately 80% of attendees came from North America, 11% from Europe, and 8% from Asia, with a total of 25 countries being represented (Heideger, 2015). If travel funding depends on acceptance and if acceptance appears to be biased towards North American submissions, some change may be needed.

Reviewer bias can be described as the result of impartiality during the reviewing process (Lee et al., 2013) and is an ongoing topic of interest in many disciplines (Rooyen, 2001; Hirschauer, 2004; Syavash, 2014; Guthrie et al., 2015; Manchikanti et al., 2015; PEERE 2014–2018). While Information Science researchers have published about peer reviewing and biases in other fields, there appears to be a research gap when it comes to studying biases in the Information Science field itself. In addition, most previous studies have analyzed journal peer review processes, but not conference review processes. This may be due to the fact that permission to study review data from conferences is more difficult to obtain, because of the various stakeholders that are involved in a conference that changes organizers every year. To best of the authors' knowledge, the only similar study was published by Kumar et al. (2008), who examined the review process of the SIGCSE conference, which was used to present guidelines for better reviewing processes. In this paper, we focus on only one form of bias: the *presumptive* bias. It includes biases evoked through institutional, gender, or cognitive differences (Lee et al., 2013).

Gender is usually the first category that is raised as potentially problematic. Pittinsky et al. (2000) and Budden (2010) found that female authors were at a disadvantage, and Borsuk et al. (2009) showed that female reviewers were stricter than male reviewers. Primack et al. (2009) found no evidence of a difference among genders. Kliewer et al. (2005), Ceci & Williams (2011) and Walker (2015) confirmed the findings of Primack et al. using different samples. More recent work indicates that gender does not appear to be a bias-inducing factor. The verdict about age is also unclear. While Kliewer et al. (2005) discovered that (especially younger) age has an effect on the review outcome, Primack et al. (2009) found the opposite to be true. Age will not be studied in the present study, since we were unable to collect reliable data about age. The same also holds true for language bias: some studies argue in favor of such a bias (Tregenza, 2002; Ross et al., 2006) while others argue against it (Loonen et al., 2005; Walker 2015).

Nationality appears to be a more critical presumptive bias. Only Primack et al. (2009) discovered no evidence of a bias between nationalities. In contrast, Marsh et al. (2008) found that American reviewers have a reputation of being harsh towards their countrymen and lenient towards non-American authors. Another study showed that reviewers from China judge their compatriots more harshly than reviewers from other countries (Campos-Arceiz, 2015). American reviewers are said to be more lenient than their colleagues from the UK or Germany (Budden, 2010). While these studies show a bias because of different reviewer nationalities, they have one fundamental flaw: they examined nationality as a proxy of the current country of submitting author (Cronin, 2009). We took care to avoid this issue in our study.

In this study, we attempt to locate presumptive biases in the form of rating scores, overall recommendations for acceptance, and review texts. The aforementioned studies use quantitative scores and available author and reviewer data to detect biases. Review texts have rarely been studied, however, and the few existing studies show that acceptance or rejection depends on relevance and research design, not on presumptive biases (Bornmann et al., 2010).

3 Methodology

The iConference review process is managed using the conference management system ConfTool[1]. For the study described in this paper, we received access to the reviewing results from the 2014 and 2015 iConferences. Institutional review board approval was obtained from Humboldt University and from the iCaucus executive committee. Since the review data did not provide enough information on the cultural background of authors and reviewers nor on the style and tone of reviews, we collected additional data. Figure 1 provides an overview of the data sources used in our analysis.

[1] http://www.conftool.net

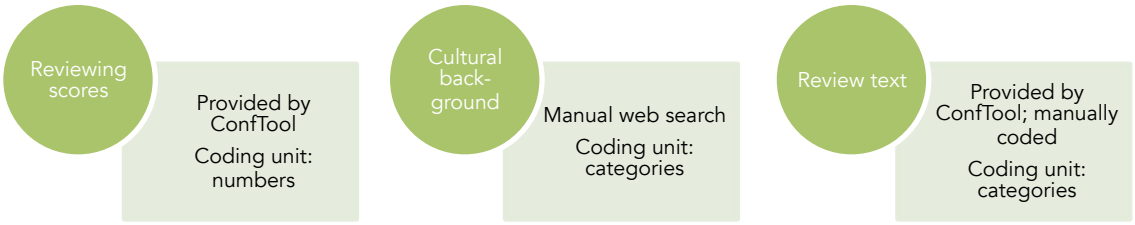

Figure 1. Data sources used in our analysis

3.1 Review scores

We extracted the review scores using ConfTool's export functionality. This data contains the following elements for each separate review:

- Authors (ID and full name), reviewer (ID and full name), chairs (ID and full name), paper ID and title
- Review text (review, summary, and internal comments)
- Review scores on five different aspects on a scale from 1-5 (with official iConference review instructions for each aspect)
  a) Soundness (*"Is the research represented in this submission convincing? Are the research questions well defined? Is the approach sound and well chosen for the authors' aims? Are the findings or results supported by evidence from the research?"*)
  b) Significance (*"How significant is the work? If the ideas are novel, will they also be useful or inspirational? If the results are sound, are they also important?"*)
  c) Originality (*"How original is the work reported in this submission? Are the ideas novel? Is the approach innovative or creative? Do the authors identify new problems, ask new questions, or move in new directions?"*)
  d) Clarity (*"Is this submission well-written and well-structured? Would the explanation benefit from more examples or illustrations? Are the results appropriately interpreted, explained, and put into context? Is there sufficient detail for an expert to validate the work?"*)
  e) Relevance (*"How closely related is the topic of the submission to the iConference theme? What level of interest do you expect it to elicit among attendees at the conference?"*)
- Overall recommendation of acceptance (on a scale from 1-5)
- Confidence of the review in their assessment (on a scale from 1-4)
- Final outcome for the submission (acceptance or rejection)

In 2014, each submission had two automatically assigned reviewers from a general pool of reviewers. In 2015 each paper submission had two automatically assigned reviewers as well as a meta-reviewer, whereas notes and posters were only assigned two and one reviewers respectively. All reviewers came from a single reviewer pool. In addition, 2015 reviewers could bid for submissions to allow for a closer match between their expertise and assigned submissions. Before analysis, duplicates were removed from the data set. For example, an author might have signed up to ConfTool in 2014 with their first name and surname and signed up again as a reviewer in 2015 with their first name, middle name, and surname. All data was then normalized, cleaned, and imported into a MySQL database. All personal information was anonymized by assigning random, numerical IDs. All subsequent analyses described in this paper were only performed on this anonymized data set.

3.2 Cultural background

The ConfTool data alone does not provide enough information about the authors and reviewers to make any valid statements about the influence of cultural background. ConfTool only provides information about the current country of residence, yet researcher mobility is an increasingly big part of academic careers and the current country of residence may not be the same as the country of origin or where the PhD degree was obtained (Cronin, 2009). PhD traditions can differ from country to country and thereby have an effect on writing and reviewing behavior at the iConference. To examine these types of issues, we augmented our data set by collecting additional demographic information on our authors and reviewers: (1) gender; (2) PhD completed (yes/no); (3) university of PhD graduation; (4) country and continent of PhD completion; (5) country and continent of Master's degree; and (6) country and continent of origin.

Two student assistants were tasked with collecting this additional data from the Web. At no point did these students have access to the review results; they were only provided with a list of names. We collected this information only for all authors, reviewers, and chairs—745 people in total. We limited ourselves to only first authors for several reasons. Author ordering conventions differ from field to field, but in information science the first author is typically the person who contributed most to a submission. Last authors commonly—but not reliably—act as mentors and/or perform copy-editing of the original text. Finally, extending the data collection process for all of a submission's co-authors as well would have more than doubled the amount of effort necessary, so we only considered first authors.

If a specific piece of information could not be identified, it was marked as "Unclear". Taiwan was counted as a part of Mainland China and Puerto Rico was included in the US counts. Finally, if the country of origin could not be found, we used the country where the Bachelor degree was completed, assuming that most people complete a Bachelor in their home country.

3.3 Review text

To determine whether the assigned review scores, which are intended to summarize the review, were representative of the written reviews, we wanted to compare official review scores with a manual coding of the review text. As manually coding 1,265 reviews was not feasible, we restricted ourselves to a randomly selected subset of 200 reviews. All reviews were coded at the sentence level. We excluded meta-reviews from the coding process, since they typically summarize the other reviews. We used the following codes in our content analysis:

- **Category 1**: Review texts were coded according to how they represented the review components
  a) *Soundness*, *Significance*, *Originality*, and *Clarity* were the main codes (see the definitions in Section 3.1). Each code was coded with the following child codes: (1) *High* (corresponding to a rating of 4 or 5 on a scale of 1-5); (2) *Neutral* (if the main code was not mentioned in the review); and (3) *Low* (corresponding to a rating of 1-3 on a scale of 1-5). Text coded as *High* was usually praise, whereas text coded as *Low* was typically criticism. In additional, the *Quality of English* used was coded with the same child codes.
- **Category 2:** Tone of the review
  a) Six main codes *Very positive*, *Positive*, *Slightly positive*, *Slightly negative*, *Negative*, and *Very negative*. Child codes were for each the presence of the code with *Yes* and *No*. In theory, one review could contain elements on all tonal levels, and indeed, several reviewers use both very positive and very negative tone examples in the same review. In these cases, we selected the dominant code.
  b) Additional codes *Language slip*, *Personal perspective*, *Desire to help*, *Authors tried to*, and *Hammer* or *Flower*. Language slips are phrases with a negative connotation; one example from the sample was: "*this submission is like an embryo – not quite ready*". Reviewers use personal perspectives when they phrase their review from their point of view; for example: *"I still do not understand"*. Some reviewers expressed the desire to help authors, e.g., *"I hope this helps you improve the paper"*. Some reviewers expressed that authors tried to do something, but did not quite achieve their goal(s), e.g. *"This poster promises to present…"*. Finally, we coded the initial sentences of reviews as negative (by bringing down the *Hammer*, e.g., *"Overall, more proofreading is needed."*) or positive (*Flower*, e.g., *"Overall this is a solid piece of work"*).

4 Results & Analysis

This section presents the results of our analysis of the iConference reviewing data and coding. Sections 4.1 and 4.2 aim to analyze the demographics of the iConference reviewer and author community respectively. Section 4.3 examines the review scoring in more detail, whereas Section 4.4 takes a closer look at the content analysis we performed of the reviews and how this relates to the official review scores.

4.1 Author demographics

Our augmented data set from the 2014 and 2015 iConferences provides us with an excellent opportunity to analyze the author and reviewer communities of the iConference community in those two years. We only present results aggregated over both years, unless interesting differences exist between the two years. We start by analyzing the authors and their submissions. In 2014, the Berlin iConference received 276 submissions (113 full papers, 74 notes, and 89 posters), whereas the 2015 iConference in Newport Beach received 325 submissions (136 full papers, 66 notes, and 123 posters), which represents a 17.8% increase in submissions. Of these 601 submissions, 53.1% were accepted for publication and

presentation. The prestige of the submission type is commensurate with the difficulty of getting accepted: only 35.3% of all full papers were accepted, 53.6% of all notes, and 73.6% of all posters.

4.1.1 Gender

The 601 submissions originated from 278 female and 221 male first authors for a total of 499 unique first authors. Female first authors were responsible for 56.7% (N = 601) of all submissions and they submitted 1.23 papers on average compared to 1.18 for male first authors. This productivity difference is not significant, however, according to a two-tailed independent-samples t-test with a mean difference between the genders of 0.050 ($t(497) = 0.938$, $p = .349$, ES = 0.0018, 95% CI [-0.055, 0.155]). The average number of submissions per first author was 1.20, although this is likely to be higher when co-authorship is taken into account. Figure 2 shows the distribution of authors, submissions, reviewers, reviews, and chairs over the different genders.

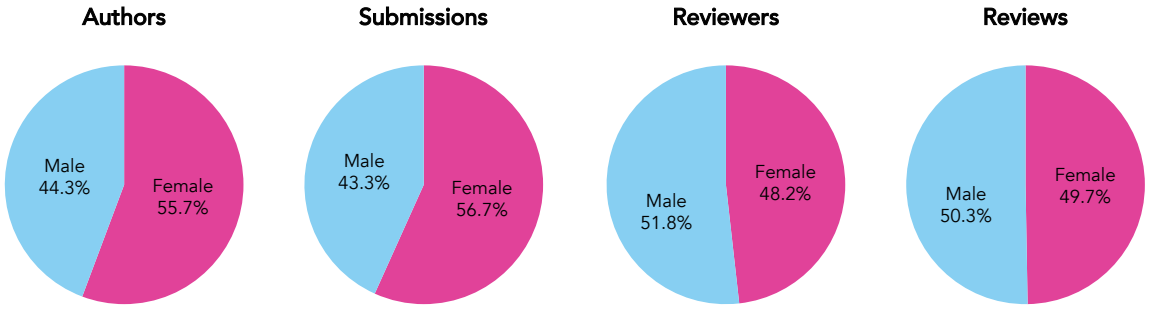

Figure 2. Distribution of authors, submissions, reviewers, and reviews by gender.

Female first-author submissions were accepted at a higher rate of 53.4% (N = 341) compared to male first-author submissions at 52.7% (N = 260). This difference is not significant however, according to a Chi-square test ($\chi^2(1, N = 601) = 0.028$, $p = .869$).

4.1.2 Location

Where do these 601 submissions come from? When we look at first-author affiliation, we find that the University of Illinois at Urbana-Champaign and the University of Washington are tied for first place as the two most prolific universities at the 2014 and 2015 iConferences with 34 (or 5.7%) submissions each. Rounding out the top five are the University of Pittsburgh with 29, Florida State University with 27, and the University of Wisconsin Milwaukee with 19 submissions. Wuhan University is the first non-US institution on the list and the seventh most productive institution with 17 submissions.

Perhaps unsurprisingly, the geographical location of a conference has a strong influence on where the submissions come from. This is most likely due to higher travel costs when submitting authors do not live in the same country or on the same continent as where the conference is being held. The 2014 European iConference saw 109.8% more submissions from Europe relative to 2015, whereas the 2015 North American iConference saw 19.5% more submissions from North America than in 2014. An interesting observation is that while submissions from researchers based in Europe dropped by 52.3% from 2014 to 2015, the acceptance rate of these European submissions jumped from 32.4% to 51.6%. One possible explanation for this could be that gaining more experience with the themes and conventions of the iConference has paid off over time for the European researchers. Another explanation could be that European researchers are targeting submission types with higher acceptance rates. However, this does not appear to be the case: in 2014 48.5% of all European submissions were full papers—the category with the lowest acceptance rate—which increased to 54.8% in 2015.

What influences where and how much is submitted? In addition to travel costs, the continent or country that researcher obtained their PhD degree in could also play a role. The PhD period is often a formative time in which researchers learn the traditions, conventions, and work ethic of their field. If we look purely at continent of current residence, researchers living in North America and Asia are the most prolific in terms of submissions with averages of 1.24 and 1.25 respectively. Researchers living in Europe only produced 1.11 submissions on average. This lower number is probably due to lower pressure to publish in the European non-tenure-track systems. This may change in the future, as more European universities are moving towards tenure-track systems. Figure 3 shows a general overview of the distribution of authors and submissions over the different continents.

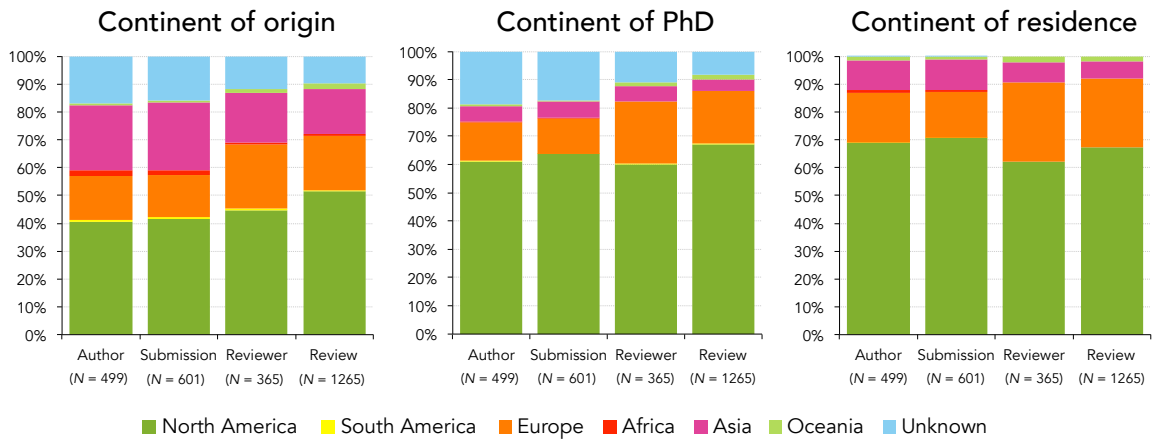

Figure 3. Distribution of the authors, submissions, reviewers, and reviews by continents of origin, PhD degree and current residence. The 'Unknown' category represents those cases where it was impossible to find the desired information on the authors' homepages.

In terms of countries, the US is number one in terms of submissions with Canada and the UK usually rounding out the top three. This holds for both country of PhD degree and country of current residence. Usually, submissions with first authors native to North American or European countries have an acceptance rate close to the global average of 53.1% for the iConference. For example, the US citizens have a combined acceptance rate of 57.3% (N = 225), Canadian citizens of 52.2% (N = 23), and the researchers originally from the United Kingdom have an above-average acceptance rate of 58.1% (N = 31). In contrast, Asian and African countries suffer more rejections. South Korea is a positive exception to this rule with a 66% acceptance rate of its 33 submissions, whereas Denmark appears to be a negative exception with only a 16% acceptance rate (N = 6). If we look at country of origin, China and South Korea are the second and third-most productive nations.

While it is ultimately quality rather than quantity that makes research impactful, it is interesting to look at the productivity of different nationalities: which nationalities submit the most (on average)? The global average is 1.20 submissions and four nationalities (that have at least 5 submissions) are more productive than this on average: Iran (M = 1.44), Germany (M = 1.28), Australia (M = 1.25) and China (M = 1.24). UK citizens are less productive than the global average (M = 1.11) as are the US (M = 1.09), Canada (M = 1.00), and South Korea (M = 1.00), although the latter has a 66% acceptance rate.

4.1.3 PhD completion

The iConference appears to be very welcoming to PhD students: the proportion of submissions from authors with and without a PhD degree is very similar at 50.6% and 49.4% (N = 567). It is reasonable to expect that being awarded a PhD degree would give one a better idea of how to do research and therefore come with a higher acceptance rate. However, this does not turn out to be the case. The conditional probability of having your submissions accepted given that you have a PhD is 0.5087, whereas the overall acceptance rate is 53.1%. A likely explanation for this may be that we can only take first authors into account. Successful papers are typically collaborations between multiple authors and PhD students tend to collaborate with their supervisors, which should eliminate any differences between these types of submissions. In general, it is also a sign that the iConference is accepting of student submissions and that PhD students deliver high-quality work.

Researchers with North American PhD degrees have the highest acceptance rate at 60.5% (N = 382). Researchers with European PhD degrees achieve a 44.6% (N = 74) acceptance rate and Asian PhD degrees achieve 22.9% acceptance (N = 35). However, of those 231 accepted North American submissions, 27.2% come from researchers originally from Asia and only 52.4% come from North American citizens. Asian researchers in general have an acceptance rate of 53.4% (N = 146), whereas researchers originally from North America saw 56.8% (N = 250) of their submissions get accepted.

4.2 Reviewer demographics

In total, 1265 reviews were completed by 365 unique reviewers for an average number of 3.47 reviews per reviewer. While the number of reviewers decreased from 232 in 2014 to 207 in 2015, the number of completed reviews actually increased from 594 to 671. This represented a statistically significant increase

in review burden from 2.56 to 3.24 average reviews per reviewer according to a two-tailed independent-samples *t*-test ($t(289.240) = 2.670$, $p < .01$, ES = 0.016, 95% CI [0.197, 1.165]).This difference is most likely due to a change in the reviewing setup made for the 2015 iConference, when a meta-reviewing tier was introduced for full paper reviewing. As a result, the average number of reviews per full paper increased from 2.13 to 3.01—a significant mean increase of 0.882 according to a two-tailed independent-samples t-test ($t(20.979) = 126.513$, $p < .001$, ES = 0.64, 95% CI [0.799, 0.965]). The average number of reviews per note dropped significantly from 2.22 to 2.00 ($t(3.463) = 92.267$, $p < .005$, ES = 0.08, 95% CI [0.087, 0.345]) and the average number of reviews per poster dropped even more from 2.12 to 1.05—also a significant difference ($t(141.069) = 26.776$, $p < .001$, ES = 0.77, 95% CI [0.995, 1.154]).

Interestingly, reviewing for the iConference appears to provide the added benefit of a higher acceptance rate of one's own submissions. The conditional probability of having at least one accepted submission, given that one has reviewed for the iConference is 0.558, which is higher than the overall acceptance rate of 53.1%. The most likely explanation for this is that accepted authors are more visible at the iConference and therefore more likely to be asked as reviewers. Moreover, research expertise in general is likely to be a mediating variable for success in both activities.

4.2.1 Gender

The 1265 iConference reviews were completed by 176 female and 189 male reviewers; Figure 2 shows the distribution of gender over the reviews and reviewers. Despite a larger share of male reviewers, more reviews were completed by female reviewers on average ($M = 3.57$) than by male reviewers ($M = 3.37$). This difference is not significant according to an independent-samples *t*-test ($t(363) = 0.685$, $p = .494$).

4.2.2 Location

Figure 3 also shows the distribution of reviews and reviewers over the different continents. Looking at the number of reviews by continent, we find that North America and Oceania provide the most reviews on average, both in terms of nationality and where the PhD degree was obtained. North American citizens provided 4.01 reviews on average ($N = 653$) and Oceania citizens 5.00 ($N = 25$). Below the global average of 3.47 reviews per reviewer are the European citizens, who provided 2.95 ($N = 245$) reviews on average, and the Asian citizens ($M = 3.17$, $N = 206$). Looking at the continent of residence, it is interesting to note that European residents make up a much larger share of the reviewer population compared to Asian researchers (28.5% vs. 7.1%), than they do for the pool of authors (17.8% vs. 10.8%). This suggests a possible lack of representativeness in the reviewer corps.

At first glance, the conference location should not have any influence on where most of the reviews come from, as it does not require physical travel to perform reviews. However, the same pattern observed for submissions is present for reviews as well: the 2014 Berlin iConference had 140% more reviews from Europe, while the 2015 Newport Beach iConference had 41.8% more reviews from North America. It is not completely surprising that there are differences. The general and program chairs were overwhelmingly from the same conference continents, so their suggestions for reviewers drawn from their personal networks are likely biased towards their continent of residence and PhD degree. However, this neither explains nor justifies the drop in the share of European reviewers in 2015. In fact, only 31.9% ($N = 232$) from 2014 continued reviewing in 2015. There can be multiple reasons for why a reviewer does not continue from year to year: lack of time, poor reviews, or because they were forgotten. However, a low continuity rate of 31.9% coupled with a large change in the geographical distribution of the reviewer corps suggests that more effort could be made towards stability and representativeness of the reviewer corps.

4.2.3 PhD completion

Having completed a PhD degree appears to be fairly important for reviewing for the iConference: 990 of the 1220 reviews (or 81.1%) where the PhD completion status was clear had a PhD degree. This suggests that research expertise is required to become an iConference reviewer.

4.3 Review scores

For individual iConference reviewers, the final output of the reviewing process was a written review detailing the strengths and weaknesses of the submission under review and a set of scores rating different properties of the paper. In this section we examine the numerical review scores in more detail.

Review scores play an important role in the chairs' decisions on which submissions to accept or reject. Often, chairs rank submissions by the total review score—a weighted sum of the individual component scores—and review text is only inspected in borderline cases or award nominations. As a result, it is important to examine the review scores and determine whether specific factors have a meaningful influence on these scores and thereby the final acceptance outcome.

Table 1 shows the average review component scores divided by gender. In contrast to the related work by, for instance, Primack et al. (2005), we do find a difference in behavior between female and male reviewers. Female reviewers assign higher review scores than male reviewers on all categories (except for confidence). These differences are significant for three components: soundness, originality, and overall recommendation. The mean difference in soundness score between female and male reviewers of 0.178 is statistically significant according to a two-tailed independent-samples *t*-test ($t(1263) = 2.547$, $p < .011$, ES = 0.005, 95% CI [0.041, 0.316]). The mean difference in originality scores of 0.161 is also significant ($t(1263) = 2.534$, $p < .011$, ES = 0.005, 95% CI [0.036, 0.285]). Finally, female reviewers assign higher overall recommendation scores than men with a mean difference of 0.158 ($t(1263) = 1.982$, $p < .048$, ES = 0.003, 95% CI [0.002, 0.314]). The significance of these differences in soundness and overall recommendation is especially meaningful as they typically play a big part in the total score.

Review component score	Gender		Continent of PhD degree				Total
	Female	Male	N. Am.	Europe	Asia	Oceania	
Soundness	**3.22****	3.04	3.14	**3.28**	2.92	2.67	3.15
Significance	**3.17**	3.06	3.13	**3.21**	3.19	3.05	3.15
Originality	**3.16****	3.00	3.09	**3.17**	3.02	2.81	3.10
Relevance	**3.02**	2.93	3.00	3.01	**3.08**	2.48	3.00
Clarity	**3.58**	3.45	3.52	**3.72**	3.23	3.19	3.54
Overall recommendation	**3.20***	3.05	3.13	**3.27***	2.83	2.76	3.14
Confidence	2.94	**2.95**	2.99	3.00	3.06	**3.62**	3.01
Total score	**32.31**	30.99	31.74	**33.07**	29.29	27.24	31.28

Table 1. Average review component scores distributed by gender and continent of PhD. Continents with ≤ 5 instances were removed. Highest scores per attribute are marked in bold. Scores marked by * and ** signal a significant difference compared to the other groups at the .05 and .025 level respectively.

In contrast to our hypothesis formulated in Section 1, the cultural background—as represented by the continents of origin, PhD degree, and residence—do not appear to have a meaningful influence on the assigned review scores. Apart from the score for relevance, no significant differences could be found between the current continents of residence and the review component scores according to one-way ANOVA tests. In terms of the continent of PhD (shown in Table 1), there were only significant differences between the continents for the clarity score according to a one-way ANOVA; no other components were significantly different. There were no significant differences for the current continent of residence. Factorial ANOVAs also could not uncover any interaction effects between gender and continent of PhD or between gender and continent of residence. An interesting observation is that, when comparing the nationalities of the different continents, female reviewers nearly always assigned higher overall recommendation scores than male reviewers, except for Asia, where this situation was reversed. However, according to a factorial ANOVA this interaction effect was not significant either.

Research expertise could have an influence on the confidence a reviewer has in their judgment and thus the score assigned to this component. Our only way of measuring research expertise—the completion of a PhD degree—did not show a significant difference in confidence scores according to a two-tailed independent-samples *t*-test ($t(1218) = 1.907$, $p = .057$, ES = 0.003, 95% CF [-0.010, 0.723]).

One of the factors that influence the clarity of a submission could be language proficiency. Whether or not a reviewer is a native speaker of English could influence the assigned clarity score. This effect could go either way: either non-native speakers are more lenient because they can relate, or native speakers are more lenient because understanding broken English is easier for them. We define native speakers as those reviewers originally from Anglo-Saxon countries with more than 5 iConference reviewers: Canada, Ireland, the UK, and the US. Non-native speakers hail from any other country with more than 5 reviewers. Non-native speakers did assign higher clarity scores on average ($M_{native} = 3.50$ vs. $M_{non\text{-}native} = 3.59$), but this difference was not significant according to a two-tailed independent-samples t-test ($t(649.713) = 1.050$, $p = .294$, 95% CI [-0.083, 0.274]).

4.4 Review text

The longest review submitted to the 2014 and 2015 iConferences had a length of 1346 words, while the shortest was only 9 words long. The average review length was 266.6, with a median of 213.0 and a standard deviation of 207.8, suggesting a skewed and highly varied distribution of review length. In the first phases of our content analysis we coded all review text according to how well they represented the aspects of *Soundness*, *Significance*, *Originality*, and *Clarity*. Our coded ratings for *Soundness* showed

only a moderate, positive correlation with the reviewers' scores (r = 0.492, p < .001, N = 200). *Soundness* was found to be present in 73.3% of the reviews, the most commonly mentioned aspect. In 21.1% of the reviews, *Soundness* was coded as low, and reviews that were coded with a low score for *Soundness* were rejected in 70.2% of the cases, suggesting that it is indeed an important aspect for reviewers.

Together with *Soundness*, a perceived lack of *Significance* was another 'kiss of death' from reviewers: 73.7% of these reviews ended up being rejected. *Significance* was only present in 40.7% of the reviews and showed a moderately positive correlation with the reviewers' score (r = 0.391, p < .001, N = 200). In contrast, papers coded for high *Significance* were accepted 60.5% of the time.

Originality showed a lower positive correlation with the reviewers' scores (r = 0.323, p < .001, N = 200). In fact, 42.2% of the time, reviewers did not cover *Originality* in their review. In 46.2% of the reviews the submission was described as original and in 11.6% cases as unoriginal. Part of the reason for the 46.2% share of original reviews was that a statement such as *"this is an interesting research project"* was coded as representing *High* originality, and many reviewers use this phrase as a standard introductory sentence. Word frequency counts further revealed that the word *"interesting"* was used 1030 times (7% of all words), while the word *"important"* was only used 505 times. *"New"* in the sense of novel research was used 362 times and *"attention"* in the sense of this research will draw attention was used 99 times. Reviews coded with a low score for *Originality* were rejected in 60.9% of the cases.

In 50.8% of the cases, *Clarity* was not mentioned at all in the paper and was moderately positively correlated with the reviewers' clarity score (r = 0.400, p < .001, N = 200). Reviews that were coded with a low score for *Clarity* were rejected in 63.8% of the cases, but this was not significant according to a Chi-square test. While the *Quality of English* was only explicitly mentioned in 6.0% of all reviews, a low quality coincided with a rejection rate of 54.5%. However, there was no meaningful correlation between *Quality of English* and the reviewers' scores for clarity (r = -0.063, p < .001, N = 200), suggesting that the quality of English is only a small part of the overall clarity of a paper. Overall, we observe that reviews are often not representative of the review scores that need to be assigned. Several reviewing aspects, such as significance, originality, and clarity are explicitly covered in only half of the reviews or less.

We also coded all 200 reviews for the tone expressed in them. Unsurprisingly, reviews written in either a *Very positive* or a *Positive* tone received significantly higher overall recommendation scores. Acceptance rates were not significantly affected, however, by reviews with *Very positive* elements in them according to a Chi-square test ($\chi^2(1, N = 198) = 2.925$, p = .087) or by reviews with *Positive* elements ($\chi^2(1, N = 199) = 2.753$, p = .097). The reverse was true for reviews coded as *Very negative* or *Negative*: they received significantly lower overall recommendation scores. *Negative* elements did show a significant increase in rejection rates ($\chi^2(1, N = 199) = 4.587$, p < .05) as the presence of *Very negative* elements ($\chi^2(1, N = 198) = 5.999$, p < .025).

We also coded the initial sentences of each review for whether they brought the *Hammer* down from the start or whether they contained a positive statement (*Flower*). iConference reviewers were overwhelmingly positive with 83.4% of all reviews being coded as positive. There was also a significant difference in the reviewers' overall recommendation scores between these two groups ($t(197) = 2.674$, p < .01, ES = 0.035, 95% CI [0.154, 1.019]). This suggests that the opening sentence of a review could have great predictive value for the final reviewer recommendation. There were no interesting findings with regard to the other four codes *Language slip*, *Personal perspective*, *Desire to help*, and *Authors tried to*.

5 Discussion & Conclusions

In this paper we have presented the results of a comprehensive analysis of the reviewing process at the 2014 and 2015 iConferences. Our analysis showed that female researchers tend to be more productive at the iConference, both in terms of submissions and reviews, even though female researchers are underrepresented in the review corps. They also assign significantly higher review scores than men. This is in contrast to more recent findings in the related work (e.g., Primack et al., 2009). We found no presumptive biases in terms of having a PhD degree: PhD students were accepted in equal proportion, making the iConference a welcoming venue for students.

Researchers working at North American institutions submitted the majority of the papers, notes and posters, with European and Asian institutions following at a respectable distance. The conference location does have a considerable influence on the submission rates from institutions on the same continent as the conference, which is understandable due to travel funding issues. A more surprising finding perhaps was that the review corps is also heavily skewed towards researchers working on the conference continent. Europe appears to be overrepresented in the review corps in general, but the overall continuity of the review corps is low. In terms of review scores assigned to submissions, it is actually European reviewers that assign the highest scores, despite the reputation of North American reviewers being more

generous with their praise. Overall, we observe that reviews are often not representative of the review scores that need to be assigned. Clear reviewer instructions on the necessity of explicitly covering all review components in the written review would likely address this problem.

Submission and review distributions also do not appear to be completely representative of the iSchool community as a whole. Our overall recommendation would therefore be to increase the representativeness of the reviewer corps by including more women and more researchers from Asian institutions. Given that conference location is a strong predictor of the reviewer corps distribution, it will be interesting to see whether the 2017 Wuhan iConference will see an increase in female and in Asian reviewers. For the sake of continuity, however, future iConference should also do a better job of creating a stable pool of reviewers that is representative of the iSchool community as a whole.

6 References

Bornmann, L., Mutz, R., & Daniel, H.-D. (2010). A reliability generalization study of journal peer reviews: A multilevel meta-analysis of inter-rater reliability and its determinants. *PLoS One*, 5(12), e14331.

Borsuk, R.M., Aarssen, L.W., Budden, A.E., Koricheva, J., Leimu, R., Tregenza, T. & Lortie, C.J. (2009). To name or not to name: The effect of changing author gender on peer review. *Bioscience*, 59(11), 985–989.

Budden, A. E. (2010). Diversity begets diversity: an analysis of relationships between author, reviewer, and editor populations. *European Science Editing*, 36(2), 31–34.

Campos-Arceiz, A. (2015). Reviewer recommendations and editors' decisions for a conservation journal: Is it just a crapshoot? And do Chinese authors get a fair shot? *Biol. Conservation*, 186(6), 22–27.

Ceci, S. J. & Williams, W. M. (2011). Understanding current causes of women's underrepresentation in science. *PNAS*, 108(8), 3157–3162.

Cronin, B. (2009). Editorial. Vernacular and vehicular language. *JASIST*, 60(3), 433.

Guthrie, J., Parker, L. D. & Dumay, J. (2015). Academic performance, publishing and peer review: peering into the twilight zone. *Accounting, Auditing & Accountability Journal*, 28(1), 2–13.

Heideger, C. (2015). *Internal report to the Directors of iSchools*.

Hirschauer, S. (2004). Peer Review Verfahren auf dem Prüfstand. Zum Soziologiedefizit der Wissenschaftsevaluation. *Zeitschrift für Soziologie*, 33(1), 62–83.

Kliewer, M. A., Freed, K. S., DeLong, D. M., Pickhardt, P. J. & Provenzale, J. M. (2005). Reviewing the Reviewers. Comparison of Review Quality and Reviewer Characteristics at the American Journal of Roentgenology. *American Journal of Roentgenology*, 40(6), 1731–1735.

Kumar, A., Goldweber, M., Joseph, P. A. & Wagner, P. J. (2008). Reviewing the SIGCSE Reviewing Process. *Inroads SIGCSE Bulletin*, 40(2), 84–89.

Lee, C. J., Sugimoto, C. R., Zhang, G. & Cronin, B. (2013). Bias in Peer Review. *JASIST*, 64(1), 1–17.

Loonen, M.P.J., Hage, J.J., & Kon, M. (2005). Who benefits from peer review? An analysis of the outcome of 100 requests for review by Plastic and Reconstructive Surgery. *Plastic and Reconstructive Surgery*, 116(5), 1461–1472.

Manchikanti, L., Kaye, A. D., Boswell, M. & Hirsch, J. A. (2015). Medical Journal Peer Review: Process and Bias. *Pain Physician*, 18(1), E1-E14.

Marsh, H. W., Jayasinghe, U. W. & Bond, N. W. (2008). Improving the Peer-Review Process for Grant Applications. Reliability, Validity, Bias, and Generalizability. *Am. Psychologist*, 63(3), 160–168.

Nobarany, S. (2014). Rethinking the Peer Review Process. *CSCW'14 Companion*, February 15–19, Baltimore, Maryland, USA.

PEERE 2014–2018. http://www.peere.org/

Pittinsky, T. L., Shih, M. & Ambady, N. (2000). Will a Category Cue Affect You? Category Cues, Positive Stereotypes and Reviewer Recall for Applicants. *Social Psychology of Education*, 4(1), 53–65.

Primack, R. B., Ellwood, E., Miller-Rushing, A. J., Marrs, R. & Mulligan, A. (2009). Do Gender, Nationality, or Academic Age affect Review Decisions? An Analysis of Submissions to the Journal 'Biological Conservation'. *Biological Conservation*, 142(11), 2415–2418.

Rooyen van, S., Godlee, F., Evans, S., Black, N. & Smith, R. (1999). Effect of Open Peer Review on Quality of Reviews and on Reviewers' Recommendations: A Randomized Trial. *BMJ*, 318(23), 23-27.

Ross, J.S., Gross, C.P., Desia, M.M., Hong, Y.L., Grant, A.O., Daniels, S.R., . . . Krumholz, H.M. (2006). Effect of blinded peer review on abstract acceptance. *JAMA*, 295(14), 1675–1680.

Tregenza, T. (2002). Gender Bias in the Refereeing Process? *Trends in Ecology & Evolution*, 17(8), 349-350.

Walker, R., Barros, B., Conejo, R., Neumann, K. & Telefont, M. (2015). Bias in peer review. *F1000Research*, 4(21), 1–18.
Wood, F. Q. (1997). *The Peer Review Process*. Canberra, Australia: National Board of Employment.