Why Don't More Non-North-American Papers Get Accepted to CHI?

Ellen A. Isaacs and John C. Tang

The observation has been made over the years that CHI appears to accept papers from North American authors at a higher rate than it accepts papers from authors from other parts of the world. Some have suggested that the difference is primarily due to the fact that non-native English speakers tend to have more trouble writing in English. Others have claimed that North Americans value a style of research that differs from research conducted in other part of the world. So far, this type of discussion has continued with very little data to test these theories. In past years, some basic statistics have been collected and they have provided preliminary indications that North American papers are accepted at a higher rate than non-North American papers. However, nothing had been done so far to understand what causes the difference in acceptance rates.

We decided to not only test the existence of the numerical bias, but also to do a content analysis of reasons for the higher rejection rate of non-North American papers. To do this, we collected the reviews of papers rejected from the CHI '95 conference and evaluated the reasons reviewers gave for rejecting the papers. We wanted to learn whether certain negative characteristics were mentioned in reviews of non-North American papers more than in North American paper reviews.

Our hope was that, if we could identify a pattern in the review comments that would explain the different acceptance rates, the CHI papers committee could evaluate whether those comments were based on justifiable reasons or unintended cultural biases. If the latter, the committee could work to eliminate those differences and include more high quality work that would have been excluded. If the former, the committee could inform authors more specifically about the types of characteristics it looks for in the papers it accepts. Of course some combination of both approaches would be possible as well.

It should be noted that a content analysis of reviewers' comments assumes that those comments accurately reflect the reasons the papers were rejected. We know in some cases that this is not the case. For example, sometimes reviewers explicitly stated that the paper had serious English problems, but that those problems didn't affect their assessments (presumably assuming the authors could get copy editing help if the paper was accepted). In other cases, reviewers may have had trouble articulating exactly what bothered them about a paper, and so might have given a poorer rating than their comments would justify. However, we believe that it is reasonable to work on the assumption that the comments are a rough approximation of the reviewers' rationale, especially since it is the only evidence we have of their thought process when assigning a number to a review. We simply note that there may be other factors at work that might not be picked up by this analysis.

In this report, we describe how we carried out our analysis, explain our findings, and discuss some preliminary ideas about actions that might increase the participation of non-North Americans at CHI, should that be accepted as a goal.

Method

CHI '95 received 228 submissions, 66 of which were accepted, leaving 162 rejected submissions. Each paper received evaluations from between four and nine reviewers, each of which give the paper a 1-5 rating, where 5 is a strong recommendation to accept the paper. Each paper is assigned to an associate chair (AC), who writes a "meta-review." ACs are intended to sum-

marize the reviewers' comments, perhaps weighting those comments differently depending on their judgement of the seriousness of the criticisms. ACs are also free to add their own opinions to the meta-reviews.

The analysis focused on the reviews of the rejected papers. In attempt to limit the amount of work, we excluded the 14 papers that scored an average of 2.0 or lower. This left us with reviews of 148 papers. Authors' names were stripped from the reviews, so the coders would not be able to guess the nationality of the authors.

We developed a category scheme that attempted to capture the vast majority of reviewer criticisms. We developed the categories by going through a small set of the papers and generating a list of all the criticisms made. After we found that reading new reviews no longer added new categories, we grouped the categories into related problem areas. These problem areas fell into three overall categories, which we called *Content*, *Argument*, and *Writing*, defined as follows:

Content: Problems with the topic the authors chose to study or the way they chose to study it. These problems could *not* be fixed simply by revising the paper in some way; doing so would require redoing some or all of the work.

Argument: Problems with the way the authors chose to write up the paper. These problems could be fixed if the authors reconsidered their analysis, their focus, their arguments, etc.

Writing: Problems with the writing and presentation. These problems could be fixed by having a good copy editor help rewrite the paper.

There were 12 Content criticisms, 14 Argument criticisms and 9 Writing criticisms. These criticisms were grouped into subcategories as listed and defined in tables 1, 2 and 3.

To code the reviews, the coder read through each statement of each review and decided whether the statement contained any criticisms and if so, which ones. Once a criticism had been counted for a particular reviewer's comments on a paper, that criticism would not be counted again, regardless of how many times the reviewer mentioned it. In other words, the coding reflects which criticisms were made of each paper by each reviewer, not how many times they cited that problem. We made this decision because we felt that it would be likely to reveal a pattern if one existed.

We chose not to use the number of times a reviewer mentioned a problem as a measure of the severity of the problem because we felt it would introduce too much variability, and if anything, camouflage any possible patterns. Reviewers are likely to vary greatly in this respect, and coders also would be likely to disagree on the definition of a second mention of a problem compared with a clarification of the description of the problem. (For example, consider the comment, "This paper doesn't flow well, it meanders from topic to topic without making connections between the sections." Is that two statements about a problem with the flow, or one statement with a clarification?)

Table 1: Content Criticisms

Not New or Significant	
Not New	Not new, better stuff exists
Won't Stimulate	Won't stimulate research
Not Significant	Didn't learn from it, not significant
Premature	Work is premature, not ready for publication
Improper Methodology	
No Evaluation	System not used, evaluated, tested
Poor Method	Inadequate testing method
Wrong Problem	
Unrealistic	Problem wasn't realistic
Generalize	Hard to generalize to other problems
Poor Idea/ Design	System is a bad idea, not useful to users, poor design
Narrow	Problem too narrow
Wrong Conference	
Not CHI	Not relevant to HCI
Engineering	Engineering-focused, not design-focused

The bulk of the reviews were coded by one person. A second person coded a small subset of the reviews. When comparing the coding of the two reviewers, we found that the first coder identically coded 84% of the second reviewers' codes, missed 9% and disagreed with 7%. However, the second reviewer tended to note fewer problems, and so identically coded only 58% of the first reviewers' codes, missed 37% and disagreed with 5%. In the end, since most of the discrepancies involved misses rather than discrepancies, and since the first person coded 95% of the reviews, we feel reasonably confident that most of the problems were identified appropriately.

Once all the papers were coded, we determined the number of times each paper was cited for each criticism, normalized for the number of reviews for each paper. (The total number of times a criticism was cited was divided by the number of reviewers, and those numbers were used in the analysis.) We then grouped the papers by region (North America, Europe, Asia and Other). Since there were only four rejected papers in the "Other" category (three from Australia and one from Brazil) they were excluded from most of the analyses.

Table 2: Argument Criticisms

Relevance	
Rationale	Insufficient rationale for system, features not well motivated
Related Work	Not enough connection with previous/ related work
Applied	Doesn't explain how concepts/theory can be applied
Incomplete	
Undeveloped	Ideas not well developed/spelled out, incomplete
Superficial	Analysis too general, superficial
Unaddressed	Obvious or important issue/problem not addressed
Example	Needs a good example to help explain the point
Poorly Argued	
Data Support	Data insufficient to support argument
No Data	No data to support claims, no stats
Unsupported	Arguments not well supported/well articulated
Inaccurate	Inaccurate claim
Poor Focus	
Broad	Tries to cover too much, scope too broad
Poor/Wrong Focus	Lack of focus, wrong focus
Confusing	Disorganized, confusing, poor structure

Analyses of variance were done to compare the number of times papers from each region were cited with each type of problem. This analysis was designed to determine whether any region was cited with any type of problem more frequently than other regions. In addition, we conducted an analysis of variance on the numeric scores given to papers from each region. This analysis was designed to determine whether the papers from any region were rated significantly different than those from other regions. We also evaluated the acceptance rate for papers by region.

Finally, we coded the associate chairs' comments, but did not include those results with the reviewers' results because their job is to summarize the reviewers' comments, not necessarily introduce their own. We did a separate analysis on the ACs' comments.

Table 3: Writing Criticisms

Description	
Poor Description	System description unclear
Unclear Study	Unclear how system was used/studied/ tested
Poor Writing	
Poor Writing	Unclear/poor/rough/awkward writing
Improper English	Strange/poor use of English, improper grammar
Jargon	Too many technical terms undefined, too much jargon
Too Detailed	Too much detail
Wordy	Wordy
Informal	Writing too informal
Figures	Hard to understand data/graphs/figures

We were provided with a list of nationalities of the papers, which had been assigned to region based on the nationality of the contact person for the paper, in most cases the first author. This is a relatively conservative definition of nationality, since some of the papers classified as European or Asian in fact may have had input from North Americans. This input presumably would reduce the chances that the paper would exhibit any typically non-North American properties, should they exist. As a result of this classification, then, it is possible that we would overlook certain properties that are common to purely non-North American papers, but we can be reasonably sure that any results we find are relatively robust.

Results

The first part of the analysis confirmed that a difference does exist in the acceptance rate of North American and non-North American papers, and the second part examines the source of that difference.

The Difference Exists

Reviewers rated non-North American papers lower than North American papers, and at least European papers were accepted at a lower rate than North American papers. There were not enough Asian papers to conclude that they were accepted at a lower rate, although the pattern looked similar to European papers.

An analysis of variance showed that North American papers score an average of 3.17 compared with 2.74 for European papers and 2.73 for Asian papers and 2.48 for the (4) Other papers ($F(3,224) = 4.83$, $p<.01$). Only the difference between the North American vs. European and Asian papers is significant. (All post-test analyses are based on a Tukey's test with an alpha level of .05.)

A Chi-squared analysis showed that North American papers are accepted at a higher rate (36%) than European (16%) or Asian (13%) papers (Chi-square = 10.29, $p<.01$). The effect appears to come from a difference between North American and European papers, since the number of Asian papers submitted (15) is too small to show an effect.

Reasons for the Difference

To understand the reasons for these differences, we examined the reasons for the rejections. The analyses show that only a few problems are cited disproportionately across region, and most of those are mentioned for European papers. European papers were more likely to be criticized for tackling problems that weren't new or significant and for being less well focused. Both European and Asian papers were more likely to be cited for writing problems and in particular, problems in the use of English.

Analyses of variance showed that there were significant differences in the number of "Argument" problems cited ($F(2,141) = 4.28$, $p <.05$), and that the difference was due to a higher incidence among European papers compared with North American papers. There were also significant differences in "Writing" problems ($F(2,141) = 6.37$, $p<.01$). In this case, post-test analysis showed that both European and Asian papers were cited with significantly more writing problems than North Americans, but there was no significant difference between Europeans and Asians. (Table 4 shows the average number of problems cited per reviewer per paper by region.) There were no significant differences in Content problems across region.

When we look at the more specific types of problems, we find that there are significant differences in the number of "Not New or Significant" problems ($F(2,141) = 3.50$, $p<.05$) and "Focus" problems ($F(2,141) = 3.75$, $p<.05$). In both cases, Europeans are cited with significantly more problems than North Americans. There were also significant differences in the "Poor Writing" category ($F(2,141) = 10.16$ $p<.001$), which again shows a three-step result in which Asians are cited with more such problems than Europeans who are cited with more problems than North Americans. (Table 5 shows the number of problems cited for each of these issues per reviewer per paper by region.) No other categories showed significant differences.

Table 4: Average number of problems of each type cited per reviewer per paper across regions

Problem	North America	Europe	Asia
Content	1.1	1.3	1.1
Argument	1.4*	1.8*	1.6
Writing	0.7*#	1.0*	1.2#

** # indicate the mean is significantly different from other regions with that mark in that category.*

Table 5: Average number of problems of each type cited per reviewer per paper across regions.

Problem	North America	Europe	Asia
Not New/Significant	.46*	.62*	.41
Poor Focus	.16*	.28*	.26
Poor Writing	.44*	.66*	1.01*
Poor English	.04*	.20*	.40*

** indicates the mean is significantly different from other regions with asterisks in that category.*

The Not New or Significant category included the problems "not new," "didn't learn," "won't stimulate research," and "premature." Table 6 provides some examples of the types of comments included in this category. European papers were cited with these problems an average of .62 times per paper per reviewer, compared with .46 for North American papers and .41 for Asian papers ($F(2,141) = 3.50$, $p<.05$). Only the difference between Europeans and North Americans is significant.

The Poor Focus category included the problems "wrong or poor focus," "too broad," and "confusing." Table 7 provides some examples of the types of comments included in this category. European papers were cited with this problem an average of .28 times per paper per reviewer, compared with .16 for North American papers and .26 for Asian papers ($F(2,141) = 3.75$, $p<.05$). Once again, only the difference between European and North American papers was significant.

Finally, the Poor Writing category included the problems, "poor writing," "poor use of English," "wordy," "too much detail," "informal writing," "too much jargon," and "poor figures." Asian papers were cited with this problem an average of .98 times per paper per reviewer, compared with .66 for European papers, and .44 for North American papers ($F(2,141) = 10.16$, $p<.001$). This three-step difference is significant.

Table 6: Examples of comments coded in the "Not new or significant" category

Not new
"It is not particularly novel or surprising that [they found what they did]."
"However, a good deal of the work that has been accomplished and proposed, and the rationale behind it, is similar to research described previously."
"Much of the paper is a rehash of arguments that have been addressed elsewhere."
Not Significant
"I really do not see anything significant here."
"However, I do not think these lessons are particularly useful to practitioners. They aren't *wrong*. They're just fairly *obvious*."
Won't stimulate research
"I'm not sanguine that this paper will stimulate much further thought/work, since the author hasn't given the reader real grist to chew on."
"I don't think it stimulates further work because this paper fails to position it in HCI research."
Premature
"I strongly encourage the authors to continue working on this topic and look forward to reading a less preliminary account of it in the future."
"To me, this paper isn't yet ready for submission."

There has been a special interest in whether poor use of English accounts for the higher rejection rate, so we looked at just the "poor English" category. Once again, we get a significant three-step difference. Asian papers get cited for poor English an average of .40 times per paper per reviewer, .20 for European papers, and .04 for North American papers ($F(2,141) = 25.12$, $p<.001$). It also appears that this category accounts for the difference in the overall "Poor Writing" category. Table 8 provides examples of the types of comments included in this category.

To evaluate the writing problem in greater detail, we looked at the European papers and considered whether papers from English-speaking countries (i.e. the UK) were criticized for writing problems less than those from non-English-speaking countries. An analysis of variance showed that there is no significant difference in the rate at which the two groups are cited with writing problems ($F(1.59) = 1.55$, ns). However, of the nine European papers accepted to CHI, seven of them were from the UK.

Table 7: Examples of comments coded in the "Poor Focus" category

Wrong or Poor Focus
"This paper might be acceptable if it were narrowed to focus on a method for analyzing data, and if some of the issues raised above were addressed."
"The groundwork could be laid in far less space ..., leaving plenty of room to actually do something with the insight."
Too Broad
"This paper tries to cover too much material, and unfortunately ends up not covering enough in sufficient depth."
"This looks like good work, but the paper covers too many different topics for a single conference paper."
Confusing
"The structure of the paper needs to be made clearer, so that new information can be put in context."
"The relationship between the two case studies is difficult to see. The transition between them is abrupt. There is not enough integration of their diverse contents."

Table 8: Examples of comments coded in the "Poor English" category

"There are frequent occurrences of sentence fragments, grammatical errors, misspelled words, and extremely awkward phrasings."
"English is often awkward or incorrect."
"There is some general difficulty in terminology that might be helped by an editorial pass by a native speaker of English, but this is not a major problem."

If we assume the accepted papers had minimal writing problems, we can conclude that although poor writing doesn't appear to be hurting papers from non-English speaking countries more than those from English-speaking countries, the latter appear to be less likely to have writing problems. In other words, poor writing is penalized regardless of country, but authors whose first language is English are less likely to have trouble writing in English.

Finally, we analyzed the data from the associate chairs to see if they were introducing a bias into the evaluation process. We found no significant differences in the problems cited by ACs.

Discussion

These findings indicate that European papers are more likely to be judged to have certain types of problems than Asian papers, in part because there were not enough Asian papers to find definitive results (which is a result in itself). The only area where Asian papers were considered systematically deficient was in their use of English, whereas European papers were criticized not only for problems with English, but also for their choice of issues and which aspect of the issues they chose to focus on.

Several possible courses of action are raised by these results. On the one hand, it is possible that these differences reflect a meaningful difference in the focus of research in North America compared with Europe. In this case, a reasonable action is to better inform Europeans about the types of problems and approaches to problems that CHI finds interesting. This approach would allow Europeans to decide whether they want to shift their focus to get their papers accepted to CHI.

On the other hand, this difference might reflect a narrow-mindedness among CHI reviewers. In this case, it would be reasonable to make a pro-active effort to educate CHI reviewers about the merits of the European focus.

In addition, given the finding that too few Asian papers were submitted to yield significant statistical results, it would be helpful to consider ways to increase submissions from Asian countries.

After a preliminary version of this report was circulated among a number of CHI organizers, a meeting was held after CHI '95 to discuss ways to address this issue for the CHI '96 conference. The CHI '96 conference chairs, technical program chairs, papers chairs, international relations chairs, some of the equivalent chairs from CHI '95, and the authors of this report were invited and provided with the report. During the discussion, the group came to a common interpretation of the data on international acceptance rates and rationales. In addition, there was an extensive discussion of those attributes CHI values in a paper. For example, the group came to realize that it expects submitters to know what CHI is about and to explain how their work is relevant to this community. The international chairs felt they came away with a better understanding of the types of papers CHI prefers and vowed to advise their country-mates on whether and how to submit their work to CHI.

Meanwhile, the committee decided to take several steps based on the analysis provided here. Some actions are designed to help non-native English speakers with writing help, others are designed to make it clearer to submitters what CHI values in a paper, and others are designed to help standardize the review process to be more responsive to the types of problems that the committee decides are important. These actions are listed here.[1]

[1] Anyone with suggestions of other possible actions are encouraged to write the authors and/or the CHI '96 papers co-chairs at chi96-papers@acm.org.

- A mentor program has been established for CHI '96 that is intended to provide early feedback to CHI submitters who ask for help. Anyone may request a mentor, but the hope is to encourage people who are new to CHI (people from other nations, students, people from other disciplines) to submit contributions that have a good chance of being accepted. Mentors may give advice on the category in which to submit a proposal, provide suggestions on the style and content of the proposal, and, if applicable, help copy edit an early draft. Although the deadline for requesting a mentor for CHI '96 has passed, information on the mentor program can be found on the Web at http://www.acm.org/sigchi/chi96/call/InvitationToSubmit.html#MENTOR. There is no guarantee that a mentored submission will be accepted, but the hope is that their chances will be increased.
- It was agreed that the technical program chairs would work with the Asian international chair to find ways to attract more Asian papers.
- The categories developed in the coding process will be used to devise a new review template. The template will help ensure that reviewers are focusing on those aspects of papers that the committee feels are most important.
- The coding categories will be used to help update an existing document that gives CHI authors specific advice about how to write a good CHI paper. This document was written with no hard data on the characteristics reviewers actually value when reviewing papers. The revised version can give a more accurate picture of the types of criticisms typically made of rejected papers.
- The CHI committee is considering writing a "values document" for its own use to give a high-level picture of what the committee considers important. The data in this study give an objective account of what CHI reviewers in fact value by their actions, and they can help the committee determine whether there are discrepancies between CHI's actions and its values.

The following is a list of the top 10 problems cited by reviewers, in descending order of frequency. Each problem is followed by a real example that is typical of that category. This list gives a good indication of what CHI reviewers would like to see in paper submissions, independent of that authors' country of origin.

1. Not new (not new, better stuff exists) e.g. "Similar points have been made in a variety of publications, from popular... to more analytic..."
2. Related work (not enough connection with previous/related work) e.g. "The author really ought to look at all the methods coming from the object-oriented community and make comparison before trying to publish."
3. Unaddressed (obvious or important issue/problem not addressed) e.g. "The lack of consideration of alternative explanations of the data is a crucial weakness of the paper."
4. Unsupported (arguments not well supported/well articulated) e.g. "They say that [claim]. That's a compelling result. What evidence do they have for it?"
5. Poor description (system description unclear) e.g. "It is by no means explicit enough what has been accomplished and

what is in the planning stages. The paper is far too vague about which works *now*."

6. No evaluation (system not used/evaluated/tested) e.g. "User studies are only mentioned in passing, and they are really needed to validate this design."
7. Rationale (insufficient rationale for system/features, not well motivated) e.g. "The design is presented with little discussion of the alternatives and almost no justification of the design choices."
8. Poor writing (unclear/poor/rough/awkward writing) e.g. "The writing definitely needs improvement."
9. Poor method (inadequate testing method) e.g. "Small numbers of subjects are used, and the measures described are of unknown reliability and validity."
10. Not applied (doesn't explain how concepts/theory can be applied) e.g. "No suggestions are given about how either the emotion input or output might be used, beyond very vague appeals to the role of emotion in human-human interaction."

These findings indicate that CHI reviewers are looking for papers about problems that are different from or major advances on existing work, that are well argued and that are about ideas or systems that have been used and evaluated properly. They want authors to place their work in the context of existing work and show how it can be applied to other related problems. They expect authors to carefully describe what they did so that it can be easily understood, and they expect well-written papers.

Conclusion

Discussions about CHI's bias against non-North American papers have been going on for years, but in most cases with relatively little data to support or contradict people's many concerns. We hope that this report helps provide some needed data to the debate. It has already instigated the beginning of a reevaluation process about CHI's values and practices and we expect that this process will go on for some time. We also hope that this study demonstrates that it is feasible to do this kind of content analysis on such variable (and volatile) data. We encourage others who are conducting various multi-national or multi-disciplinary endeavors to pursue similar analyses if they are concerned about their level of inclusiveness.

The changes that have been proposed based on this report are an experiment. There are no guarantees that they will help expand the range of participation at CHI. We plan to track the results of these efforts through CHI '96 and write a follow-up report on the efficacy of these efforts. Based on that evaluation, we expect that further modifications will be made to help broaden international participation at CHI.

About the Authors

Ellen Isaacs and John Tang both work at SunSoft in the Collaborative Computing group, an advanced development group that develops media-based technology to help distributed groups collaborate. They both design user interfaces and conduct use studies to understand how to best design technology to suit users' needs.

Isaacs received her Ph.D. in cognitive psychology from Stanford, where she studied language use and collaboration in conversation. She has a particular interest in the psychological processes involved in unconscious biases against outgroups. She became active in the CHI community several years ago and is currently workshops and SIGs co-chair for CHI '96.

Tang received his Ph.D. in mechanical engineering from Stanford University. He has been active in the CHI community over the last 10 years, presenting work from Sun and from Xerox PARC before that, and participating in the CHI papers review committee.

Authors' Address

SunSoft, Inc.
2550 Garcia Ave.,
MTV19-219 Mountain View, CA 94043, USA.
isaacs@eng.sun.com; tang@eng.sun.com