Autism online: A comparison of word usage in bloggers with and without Autism Spectrum Disorders

A. Taylor Newton  
Department of Psychology  
University of Denver  
anewton@du.edu

Adam D. I. Kramer  
Department of Psychology  
University of Oregon  
adik@uoregon.edu

Daniel N. McIntosh  
Department of Psychology  
University of Denver  
daniel.mcintosh@du.edu

ABSTRACT
The Internet has become a place of refuge for individuals with autism spectrum disorders (ASD). In particular, weblogs are a popular option for personal expression via the Internet. Perhaps this means of communication is well suited to bypassing deficits in social interaction and communication that characterize ASD. Using the Linguistic Inquiry and Word Count (LIWC) dictionaries [10], we compared blogs of individuals with ASD to the writing of neurotypical (NT) bloggers. We found that rates of word usage were nearly identical in the two groups with one exception — there was more variation in the use of social words in ASD compared to NT blogs. This similarity in language between ASD and NT authors suggests that communication deficits routinely found in people with ASD may be due to the social context in which their communication skills are tested, and that the affordances of asynchronous computer-mediated communication may offer alternative means of testing and expression.

Author Keywords
Autism Spectrum Disorders, blogs, comparative text analysis, word usage, LIWC, unobtrusive methodology

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION
“The Internet has been to the autistic community what sign language has been to the deaf community; a channel of communication that allows them to speak for themselves.”
-Justin Muggleton, a blogger with ASD, from A Second Life with Autism, broadcast July 28, 2007, BBC.

Marked by deficits in social interaction and communication, autism spectrum disorders (ASDs) are a range of pervasive developmental disabilities including autism, Asperger syndrome, Rett syndrome, and pervasive developmental disorder not otherwise specified. With a rising prevalence endemic to developed countries and an etiology not yet established, research efforts have aimed to define the parameters of this mysterious set of disorders [9]. There is increasing evidence that deficits appear primarily when rapid processing of socio-emotional cues is required [3]. For example, when you smile at someone, he or she tends to smile quickly (within 1000ms). This automatic matching, considered an important aspect of social and emotional processing, is impaired in children and adults with ASD [2,7]. Further, atypical non-verbal communication behaviors (e.g., eye contact, facial expression, body gestures) in ASD are identified as central to deficits in social interaction and communication [1]. Despite growing understanding of these face-to-face processes in ASD, almost nothing is known about communication in ASD in less proximal contexts, in which time and distance enable more effortful and less stressful strategies. People with ASD appear to use such alternative strategies for social-emotional perception [12]. Thus, socially distal contexts, made possible by asynchronous and semi-synchronous communication modalities, may level the playing field by not requiring rapid, automatic processing and typical non-verbal communication behaviors. Further, because ASDs are almost universally diagnosed using proximal face-to-face assessments such as the Autism Diagnostic Observation Schedule, perhaps the extent of social and communicative impairments in individuals with ASD is over-generalized as only proximal states have been exhaustively examined.

If it is true that socially distal contexts are less demanding on socio-emotional processing and communication skills, there should be little difference in such contexts between individuals with ASD compared to neurotypical (NT) individuals. As we were interested in examining the "communicational deficit" symptomatic of ASD, we posit that this symptom is an artifact of diagnostic testing that requires proximal communication. Because computer-mediated communication (CMC) is not bound by this requirement, our purpose in this paper is to compare ASD and NT individuals in the context of a distal CMC channel. We predicted that when the playing field is leveled, such that there is time for effortful socio-emotional processing...
Strategies and non-verbal communication behaviors are unnecessary (due to distance), communication differences between NT and ASD groups would be diminished.

**Socially Distal Communication**
The Internet affords communication mediated by virtual space and, in some forms, time. For example, in a chat room, participants can engage one another socially in what is nearly real time, but are not in the same physical space and therefore not communicating face-to-face. As a means of communication and social interaction, blogs separate the author from the reader in both space (physical distance between computers) and time (delay between posting an entry and potential responses), making blogging a distal communication modality.

Anecdotally, members of the ASD community have noted that the distal nature of communication via the Internet has created a comfort zone. As one ASD blogger noted, “You’re on your own computer, in your own room, your own space” (as cited in [13]). In contrast to the anti-social manner in which individuals with ASD are often depicted, the Internet, a medium without the pressures of face-to-face interaction has allowed individuals with ASD to seek out social interactions and communication with others. In particular, blogging has attracted many members of the ASD community, despite a deficit in written expression in everyday speech or writing can help to predict narcissism. The advent of the Internet has dramatically increased the utility of text analysis techniques, as communication is largely available in immediately analyzable text logs (which would previously require extensive transcription). With this ease also arise new opportunities for comparative text analysis, especially with populations that are either difficult to reach or with whom it is difficult to communicate, such as individuals with ASD. Blogs, with their broad appeal to many groups and personal, diary-type focus, are ideal for gaining insight into the characteristics of their authors. Further, word usage in blogs has already been normed in a population sample [5], using the Linguistic Inquiry and Word Count (LIWC) software [10], further aiding comparative text analysis efforts. The LIWC standard dictionary comprises about 2,300 English words that capture about 80% of the words used in everyday conversation in English. These words fall into 72 categories, such as positive and negative emotional words pronouns, and words about work, school, and money.

These dictionaries were recently used to analyze correlations among LIWC word categories for 1.8 million blogs from the Blogger framework [5]. A five-factor structure of the way words are used in blogs emerged: Blogs vary in terms of 1) how “ranty” the blog is (including anger and swear words), 2) how “melancholy” it is (e.g., sad words), 3) how “social” it is (e.g., leisure, and TV words), 4) how “work”-related the blog is (e.g., work, school, and money words) and 5) how “metaphysical” the blog is (e.g., religion and death words).

**Our Study**
In this paper, we ask the question “Do bloggers with ASD show different patterns of word usage than do NT bloggers?” If so, differences would indicate that individuals with ASD are using language in a way that may reflect characteristic differences compared to NT individuals, as the current research on ASD would suggest. However, similar word usage in ASD and NT blogs would suggest that the impairments in social interaction and communication are due to or enhanced by the socially proximal contexts in which these disorders are diagnosed and studied. An absence of communication differences in these more distal contexts would indicate the need to specify further the parameters of ASD impairments.

**Method**
We found 57 blogs written by individuals self-identified as having ASD using widely available search engines (e.g., Google, Yahoo, Technorati, Blogger search) and listings on websites with content about autism. Blogs were included in analysis if they met the following criteria: the author appeared to be 18 years older, the blog had a personal diary theme, the blog was not password protected, and the blog contained at least one (1) entry a month beginning May 2007 or earlier (some had posts as early as October 2003). We did not restrict our sample to Blogger blogs as in [5], though we believe that the manner in which words are used likely does not vary much across blogging sites compared to how much it varies within blogging sites.

Forty (40) ASD blogs met all criteria, with a mean of 418.5 posts per author (SD=580; median=128). We included all blog entries made prior to data analysis (September 2007), representing each blog as a single row, with word category counts equal to the sum of words across all posts, with an additional variable representing the number of posts.

Words were analyzed using LIWC dictionaries [10] and the TAWC word counting program [4], mimicking the analysis strategy used in our normed NT comparison set [5].

**Results**
To determine whether ASD and NT bloggers use different rates of words, we used the five-factor factor structure for

1 As Kramer & Rodden did not require one post per month, our mean and standard deviation are understandably higher than theirs (M=27.1, SD=270), though as number of posts is directly modeled in their factor structure, this is not an issue.
words are used to discuss social activities or experiences, plus a negative loading for number of posts on leisure, home, sports, TV, music, and money categories, factor, which included word loadings for school, job, leisure, home, sports, TV, music, and money categories, plus a negative loading for number of posts [5]. As these words are used to discuss social activities or experiences, this result may stem from either a lack of interest in social engagement or from a lack of participation in social activities (thus leaving the blogger few social topics to blog about). Both possibilities are consistent with the diagnostic criteria for autism, indicating that scores on the social factor may covary with autistic severity, leading to a much larger variance than the population sample. This suggests that higher variability on this factor for those with ASD can be interpreted as indicating construct validity for the five-factor structure.

**FUTURE DIRECTIONS**

One of the greatest strengths of this research is that it is almost entirely non-invasive. The blogs analyzed in this paper as well as those in the comparison set were published publicly and voluntarily, and as such do not suffer any experimental demand effects. However, there are limits to the amount of information that can be gained from entirely non-invasive research: Analyses based on word usage cannot, for example, differentiate between social activity and social interest in the same way that a self-report test or laboratory study can. One caveat to this argument is that many bloggers with ASD may not be willing to engage in offline research, or may participate in a biased or skeptical manner. Indeed, several of the blogs analyzed in this paper were critically discussed research into the etiology and/or treatment of ASD, suggesting that it would be inappropriate to treat these individuals as naïve participants.

In addition, fully automated procedures such as word counting may miss important qualities that a human blog reader would notice, such as humor, sarcasm, politeness, pop-culture references, or empathy. In this paper, we were interested in a comparison between a normed sample of blogs and a set of blogs by authors with ASD, and as such our analytic method was limited by the data available in the normed sample. Future research could involve human coding of linguistic qualities and word usage in ASD and NT blogs, also serving to validate factors suggested in [5].

There is also the possibility of selection bias related to the propensity of those with ASD to blog in the first place. Blogging is an inherently social communicational activity, and people with ASD who choose to blog may be more socially oriented or communicative than those with ASD who do not choose to blog. Future research should consider collecting blog data from people with ASD who do not choose to blog (e.g., by asking them to keep a public blog as part of the study), examining people with ASD who blog on externally validated characteristics (e.g., by examining their scores on traditional Autism measures), or by examining people with disorders that are not expected to show communicational or social deficits. Alternatively, NT individuals who blog consistently may be more comfortable with distal communication than average NT individuals, decreasing the apparent difference between groups.

The approach we present also suggests a broader need to examine covariates or controls for bloggers in our sample, such as interest in social activities, extraversion, ASD

<table>
<thead>
<tr>
<th>Factor</th>
<th>t</th>
<th>p</th>
<th>Cohen’s d</th>
</tr>
</thead>
<tbody>
<tr>
<td>Melancholy</td>
<td>0.9</td>
<td>0.37</td>
<td>0.14</td>
</tr>
<tr>
<td>Ranty</td>
<td>0.17</td>
<td>0.87</td>
<td>0.02</td>
</tr>
<tr>
<td>Metaphysical</td>
<td>0.04</td>
<td>0.97</td>
<td>0.004</td>
</tr>
<tr>
<td>Work</td>
<td>-0.07</td>
<td>0.94</td>
<td>-0.007</td>
</tr>
<tr>
<td>Social*</td>
<td>0.76</td>
<td>0.45</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Tests of differences between NT and ASD bloggers, with Cohen’s d (mean difference in pooled SD units) reported as an effect size. *: Test completed without making an equal variances assumption; as such, Cohen’s d is not computable.

blogs published by Kramer and Rodden [5]. To this end, we scored each of the 40 ASD bloggers on each of the five factors by using the factor loadings published in [5]. We then compared our bloggers’ scores on each factor to the scores for the NT sample.

Surprisingly, ASD blogs did not show a mean difference from NT blogs in use of any of the five linguistic dimensions, and the effect sizes were very small; see Table 1. Further analysis found only one difference: The “Sociability” factor showed nearly four times as much variance in ASD blogs compared to NT blogs (var=16.02 vs. 4.16; Levene's Test of Equal Variances’ W=64.7, p < .001), though the average number of social words did not differ, $t(39.04)=.76$, $p = .45$.

To account for possible correlation among the five factors, we also conducted a MANOVA, predicting any differences among the five word categories from blogger group, using a sample of 18,619 blogs published with the Kramer and Rodden [5] data set. Despite the large sample size, no significant difference was found between blogger groups ($F \ (5, 18653) = .71$, $p = .61$). While it is theoretically impossible to prove a null hypothesis, the accompanying small effect sizes (see Table 1) make a strong case that even if the groups do differ, they do not differ by very much at all (at most, 14% of one standard deviation). Based on these results, we feel comfortable concluding that even if these differences were truly present in ASD versus NT bloggers, the effects are far, far smaller than one would expect based on the diagnostic criteria for ASD.

**DISCUSSION**

Despite communication deficits listed in many definitions and diagnostic procedures for ASD, we found no fundamental differences between the pattern of word usage in blogs of ASD and NT individuals. Though it is theoretically impossible to reject a null hypothesis, the observed effect sizes were so small that if differences do exist, it is questionable whether they are at all meaningful in any non-theoretical context.

We did, however, find much greater variance for bloggers with ASD than for NT bloggers in word use on the “Social” factor, which included word loadings for school, job, leisure, home, sports, TV, music, and money categories, plus a negative loading for number of posts [5]. As these words are used to discuss social activities or experiences, this result may stem from either a lack of interest in social

One of the greatest strengths of this research is that it is almost entirely non-invasive. The blogs analyzed in this paper as well as those in the comparison set were published publicly and voluntarily, and as such do not suffer any experimental demand effects. However, there are limits to the amount of information that can be gained from entirely non-invasive research: Analyses based on word usage cannot, for example, differentiate between social activity and social interest in the same way that a self-report test or laboratory study can. One caveat to this argument is that many bloggers with ASD may not be willing to engage in offline research, or may participate in a biased or skeptical manner. Indeed, several of the blogs analyzed in this paper were critically discussed research into the etiology and/or treatment of ASD, suggesting that it would be inappropriate to treat these individuals as naïve participants.

In addition, fully automated procedures such as word counting may miss important qualities that a human blog reader would notice, such as humor, sarcasm, politeness, pop-culture references, or empathy. In this paper, we were interested in a comparison between a normed sample of blogs and a set of blogs by authors with ASD, and as such our analytic method was limited by the data available in the normed sample. Future research could involve human coding of linguistic qualities and word usage in ASD and NT blogs, also serving to validate factors suggested in [5].

There is also the possibility of selection bias related to the propensity of those with ASD to blog in the first place. Blogging is an inherently social communicational activity, and people with ASD who choose to blog may be more socially oriented or communicative than those with ASD who do not choose to blog. Future research should consider collecting blog data from people with ASD who do not choose to blog (e.g., by asking them to keep a public blog as part of the study), examining people with ASD who blog on externally validated characteristics (e.g., by examining their scores on traditional Autism measures), or by examining people with disorders that are not expected to show communicational or social deficits. Alternatively, NT individuals who blog consistently may be more comfortable with distal communication than average NT individuals, decreasing the apparent difference between groups.

The approach we present also suggests a broader need to examine covariates or controls for bloggers in our sample, such as interest in social activities, extraversion, ASD
severity, blogger gender, and blogger age. Though these data are not available for the comparison set, and may not be available in an unobtrusive manner, a follow-up study could provide much greater insight. Our results indicate that there is very high variability in sociability among bloggers with ASD, emphasizing the importance of not considering those with ASD to be an undifferentiated group even in terms of a defining characteristic of the disorders.

An additional future direction is to compare distal and proximal communication within individuals. Though reading and coding public blogs is an excellent way to collect data in a non-invasive manner, it does not directly compare distal and proximal communication. Without such a comparison, it is inappropriate to make strong claims about the differences between distal and proximal communication for those with ASD.

Finally, our comparison of bloggers with ASD to NT bloggers along the five-factor structure discovered in a population sample of bloggers raises the question of whether the underlying five-factor structure of word usage in blogs replicates with the ASD sample. This could be verified via a confirmatory factor analysis, though the blogger-row methodology of [5] suggests that to confirm a factor structure across 70 word category variables would require analysis of over 2100 ASD blogs [8], which may be more than exist today.

CONCLUSION
As word usage was nearly indistinguishable for ASD versus NT populations, non-synchronous computer-mediated communication channels may be especially appropriate for individuals with ASD. This similarity in language usage may suggest that communication deficits routinely found in people with ASD may be due to or enhanced by the social contexts in which their communication skills and motivation are tested. Social and communication differences between ASD and NT individuals may be moderated by the medium of communication. Moreover, performance differences in varying contexts can point to which specific social processes are impaired in ASD. The findings here are consistent with the view that it is the rapid, automatic processes that are affected, not more time consuming, effortful social processes [2,7]. Removing the requirement for rapid processing of social-emotional and non-verbal communication behaviors, differences in social communication almost disappear. As the world in which we live so often requires communication in a highly social environment, this distinction is important to note, as it suggests that communication channels such as are provided by the Internet may be able to dramatically improve the lives of people with ASD, and facilitate communication between NT and ASD individuals.

ACKNOWLEDGMENTS
We thank Dawn Adams for data collection help and Moira Burke for comments on an earlier version of this paper.

REFERENCES