Impressions of World of Warcraft players' personalities based on their usernames: Interobserver consensus but no accuracy

Lindsay T. Graham*, Samuel D. Gosling

The University of Texas at Austin, United States

Abstract

Millions of people interact in the context of massively-multi-player-online-role-play-games (MMORPGs). Here we address whether MMORPG usernames convey accurate information about users' personalities. Impressions based on email addresses show some accuracy but MMORPG's are different because they provide little accountability, with little expectation that users will ever meet in person. We assessed consensus and accuracy in impressions based on 1357 (153 females) World of Warcraft users. Ratings of the usernames in terms of the Big Five dimensions yielded strong consensus (mean ICC = .46) but low accuracy (mean r = .01). Lens-model analysis suggest that features of the usernames have little validity, but observers used them anyway. Discussion focuses on why results might differ from those based on email addresses.

1. Introduction

More than 11 million people subscribe to the World of Warcraft (WoW; White, 2008). Accurate estimates are hard to come by, but similarly huge numbers of subscribers are thought to be associated with other massively-multi-player-online-role-play-games (MMORPGs) such as Second Life, with an estimated 6.5 million users in 2007 (Rainbridge, 2007), and Facebook's FarmVille, which alone claims to have 80 million subscribers (Yee, Ducheneaut, Nelson, & Likarish, 2011).

One feature of MMORPGs is that players may engage in long-term interactions with other players whom they may never meet in person. In many instances, if they do meet, online meetings precede offline interactions (Nardi & Harris, 2006). Thus, in most cases, usernames are the first piece of information players gain about one another.

It is well established that individuals form impressions of others on the basis of minimal information (Ambady & Skowronska, 2008). Even information as minimal as email addresses, can yield valid data regarding what a person is like (Back, Schmukle, & Egloff, 2008; Mori & Takahira, 2005a, 2005b; Takahira & Mori, 2005). One study showed that targets' email addresses elicit greater than chance levels of interobserver agreement (mean ICC = .58) about the characteristics of the target, that email-address based impressions show some accuracy, and that the accuracy of the impressions is mediated by features (e.g., number of characters, cuteness, creativity) of the addresses (Back et al., 2008). Other fragments of information, such as one's favorite music (Rentfrow & Gosling, 2006), the content of one's Facebook page (Back et al., 2010), or one's physical appearance (Naumann, Vazire, Rentfrow, & Gosling, 2009) can also generate impressions of targets about which judges agree and show some accuracy (Graham, Sandy, & Gosling, 2011).

Would MMORPG usernames also elicit consensual, accurate impressions of users? The fact that email addresses do elicit consensus and accuracy suggests that the same effects might also be found for MMORPG usernames, especially given that MMORPG usernames may not be subject to the same external constraints (e.g., social and professional propriety, company policies) as email addresses. However, the very same lack of constraints could permit MMORPG users to present an impression that deviates from the limitations of the offline world (e.g., one's offline personality).

The fragments of information examined in previous studies of personality impressions are typically accompanied by other information that exerts some accountability on the targets and therefore predisposes individuals to broadcast honest signals. For example, an individual might convey his or her music preferences in a social setting where it would not be a simple matter to pretend to be a different gender, age, or race. Some deception would be possible in
other online forums (e.g., dating and employment websites) but the expectation of future offline interactions would mitigate against engaging in it too much. In contrast, MMORPG interactions (e.g., among members of “guilds” – teams that regularly play together) can last for months or years, but still involve no expectations of future offline interactions so it would be easy for a gamer to play as a person of a different gender, age, or race, or personality without much danger of having the deception discovered.

The present study aims to examine impressions of MMORPG users based on their usernames alone. In particular we focus on the degree to which the impressions are consensual (i.e., associated with interobserver agreement) and accurate (i.e., correspond to self-reported personality provided by the users). As in many subcultures, MMORPG users may have their own norms and a specialized vocabulary that could be opaque to non-members. Therefore, to examine possible expertise effects, we compared the impressions formed by experienced WoW users with those formed by non-users.

2. Material and methods

Drawing on the design implemented in Back et al.'s (2008) study of impressions based on email addresses, we collected data from three independent sources: observer ratings of personality based on WoW usernames, users' self-reports of personality, and codings of features (e.g., creative, sexually themed) of the usernames. This design allowed us to examine the interjudge consensus among the observers, the accuracy of their impressions, and the degree to which features of the usernames were associated with the observers’ impressions and with the WoW users’ personalities.

2.1. Users

1357 WoW players (153 females; mean age = 25.96, SD = 7.46) were recruited via an advertisement on a popular WoW community website (http://www.wowinsider.com/). Users reported their WoW usernames and completed self-reports of personality. After, users were entered into a raffle to win 25 pieces of gold (the WoW currency). Users with multiple names listed just one, assumed to be the one they identified most strongly with.

2.2. Instruments

Users completed the 44-item Big Five Inventory (BFI; John & Srivastava, 1999), which measures the Big Five personality dimensions (Extraversion, Conscientiousness, Agreeableness, Emotional Stability, and Openness to Experience). The BFI would be too long for use by the observers who were rating 1357 usernames. So observers used a very brief, but widely used and validated instrument, the Ten Item Personality Inventory (TIPI; Gosling, Rentfrow, & Swann, 2003), which also measures the Big Five. The TIPI has demonstrated high convergence with the BFI (Gosling et al., 2003).

2.3. Observer ratings

All observers were undergraduates at The University of Texas, Austin (UT). Two sets of observer ratings were obtained. The first set (1 male, 4 females) had never played WoW, but all knew the basics of how the game works. The second set of observers (3 males) had all played WoW for a substantial amount of time (i.e., 2–3.5 years).

Observers independently viewed each username and rated the personality of each target. To combat fatigue, observers rated usernames in sets of 200 and were encouraged to take breaks as needed. Observers took about 3 h to rate each set of usernames.

2.4. Features of the usernames

Characteristics for coding the usernames were derived from three sources: (1) the categories used by Back et al. (2008) to code email addresses were reviewed, and the six items that could be applied to WoW usernames (i.e., creative, cute, sexually themed, self-enhancing, fantasy) were retained; (2) eight more features generated to capture elements that seemed particularly applicable to MMORPG usernames (e.g., geeky, sci-fi sounding, aggressive); (3) one objective feature (name consisting of multiple vs. single words).

Eight UT undergraduates (5 males, 3 females), none of whom served as observers, coded each username on the 15 categories (mean alpha = .60; max. = .93, min. = .32). To reduce fatigue, coders coded usernames in sets of about 430 and were encouraged to take breaks as needed. Each set took about 2 h to code.

3. Results

3.1. Interobserver consensus

To provide estimates of both the reliability of a single judge and the reliability of the aggregated judgments, interobserver consensus was calculated using intraclass correlations (ICCs) for both single, (ICC[2,1]), and aggregate, (ICC[2,k]), ratings respectively (Shrout & Fleiss, 1979). Aggregate-measure ICCs will be higher than single-measure ICCs because the averaged ratings reduce error variance through the effects of aggregation. ICCs were computed separately for the five non-WoW-player observers and the three WoW-player observers, as well as for all eight observers combined.

Consensus was strong for all Big Five dimensions (Table 1). There were small differences between non-player and player observers, such that consensus was highest in WoW-player’s ratings of Extraversion, Emotional Stability, and Openness, but highest in non-WoW-player’s ratings of Agreeableness and Conscientiousness. When all observers were combined, consensus was quite strong across all traits with the strongest consensus for ratings of Agreeableness, Conscientiousness, and Emotional Stability.

Overall, the strong interobserver consensus in all groups of raters is of a similar magnitude to those reported for impressions based on email addresses (also reported in Table 1).

3.2. Observer accuracy

To determine the accuracy of the observers’ impressions, we computed the correlations between the aggregated ratings (from non-WoW-player observers, WoW-player observers, and a composite of both types of observers), and the users’ self-reports. In contrast to the Back et al. (2008) findings for email addresses, we found no evidence for accuracy in the observers’ impressions (see Table 1).

Could the low accuracy findings be explained by low-quality criterion measures? To address the possibility that our criterion measure of self-reports was compromised by careless or random responding or by self-enhancement, we compared the scale reliabilities and the inter-scale correlations with those found in previously reported samples where participants had no incentive to provide careless or enhanced self-ratings. Cronbach alphas of .87, .78, .82, .84, and .75 (for Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Openness, respectively) were comparable to similar samples. Further, the inter-scale correlations (absolute mean r = .20) were identical to those obtained in traditional samples (John & Srivastava, 1999). These findings suggest that the low-accuracy cannot be explained by the use of an
Table 1
Observer consensus and accuracy of personality impressions of targets based on their WoW usernames.

<table>
<thead>
<tr>
<th>Trait assessed</th>
<th>Consensus (ICC)</th>
<th>Accuracy (r)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Non-WoW observers</td>
<td>WoW observers</td>
</tr>
<tr>
<td>Extraversion</td>
<td>.07 (.28)**</td>
<td>.18 (.39)**</td>
</tr>
<tr>
<td>Agreeableness</td>
<td>.29 (.67)**</td>
<td>.30 (.56)**</td>
</tr>
<tr>
<td>Conscientiousness</td>
<td>.17 (.50)**</td>
<td>.17 (.38)**</td>
</tr>
<tr>
<td>Emotional stability</td>
<td>.07 (.28)</td>
<td>.24 (.49)**</td>
</tr>
<tr>
<td>Openness</td>
<td>.05 (.20)**</td>
<td>.17 (.38)**</td>
</tr>
</tbody>
</table>

Notes: Consensus is indexed by interclass correlations (ICCs) among 5 observers (non-WoW players), 3 observers (WoW players), and 8 observers (both non-WoW and WoW player observers combined). Single and average ICCs are presented with average ICCs shown in parentheses. Accuracy is the correlation between observers’ ratings (for non-WoW player observer, WoW player observers, and all observers combined) and targets’ self-reports (N = 1357).

* p < .05.
** p < .01.
* For comparison purposes, findings for impressions of targets based on email addresses (from Back et al., 2008) are presented in data columns 4 and 8.

Accuracy criterion compromised by careless, random, or self-enhanced self-ratings.

3.3. Username cues

To understand how observers might have formed their impressions we used a lens-model analysis (Brunswik, 1956). For each personality trait, this analysis reveals observers’ utilization of username cues, the diagnosticity of the cues, and observers’ sensitivity to the diagnosticity of the cues.

Utilization of username cues was determined by regressing the codings of the 15 observable cues onto the observer ratings for each of the Big Five dimensions separately. Regressions were computed for non-WoW-player observer ratings, WoW-player observer ratings, and a combination of all observers’ ratings of personality. Adjusted-R values (Table 2) indicate the overall extent to which observer judgments were associated with the cues. Compared to non-WoW player observers, WoW player observers showed stronger tendencies to utilize the cues, particularly when making judgments of Extraversion, Agreeableness, and Emotional Stability. For comparative purposes, we present findings from Back et al.’s (2008) study of email-based impressions, which showed cue utilization in all traits, especially for Extraversiveness and Conscientiousness.

To determine which specific cues were associated with judgments of each personality dimension, we computed correlations between the cues and observer ratings (separately for the non-WoW-player observers, WoW-player observers, and all observers combined). Observer ratings were related to all the cues assessed, but varied depending on the trait (Table 3).

Next, we examined cue validity by regressing codings of the 15 cues onto user self-reports for each dimension separately (Table 2, data column 5). Adjusted R-values, which indicate how much the cues served as valid indicators of user personality, suggest that virtually none of them were valid indicators. These results differ from email addresses (Back et al., 2008), which found cue-validity for Conscientiousness, Emotional Stability, and Openness. Moreover, we found only minimal evidence for the validity of any specific cues (Table 3).

To determine observers’ “sensitivity” (the extent to which observers were weighing cues in a way that reflected their limited diagnosticity), we computed column-vector correlations between the cue-utilization correlations and the cue-validity correlations for each personality dimension (Table 2, columns 7–9). We computed a conservative measure of column-vector correlations based on the (Fisher’s z-transformed) absolute cue-utilization and cue-validity correlations. The generally low levels of sensitivity identified differ from those found for email addresses, in which observers exhibited some sensitivity for impressions of Agreeableness, Emotional Stability, and Openness (Back et al., 2008; see Table 2 for a side-by-side comparison).

4. Discussion

We examined consensus and accuracy of observer ratings based entirely on WoW players’ usernames. We found evidence for inter-observer agreement in ratings by both WoW users and non-users. However, we found no evidence for accuracy (indexed in terms of correlations between observer ratings and user self-reports). Our findings suggest that observers are making use of the cues found
in the usernames, but these cues are generally not valid indicators of the users’ personalities.

These findings raise two main questions. First, if impressions are not accurate, what factors are driving the inter-observer consensus among the raters? Second, what accounts for the discrepancy in findings between ratings based on WoW usernames and ratings based on email addresses?

Existing models of interpersonal-perception processes can inform these questions. Kenny’s (1994) Weighted Averages Model (WAM) places substantial emphasis on the role that stereotypes play in observer judgments. According to WAM, consensus among observers will increase to the extent that observers hold similar stereotypes and use them in their judgments. So, if observers share the stereotype that women are more agreeable than men and if the observers believe that User A is female and User B is male, then consensus may be promoted. It is also quite possible that other, more specialized stereotypes (e.g., “nerd”) may have been at play. The pattern of strong consensus but no accuracy suggests that observers could be basing their judgments on stereotypes of various types of MMORPG users. We did not collect the data needed to test these possibilities, but future work should examine what those stereotypes might be and whether they are indeed driving the consensus effects reported here.

One possible explanation for the observers’ lack of accuracy could be their use of invalid stereotypes. The use of stereotypes can promote or hinder the accuracy of judgments depending on whether the stereotypes have a kernel of truth to them (Lee, Jussim, & Mccauley, 1995). For example, in a study of personality judgments based on living spaces, the valid gender stereotypes promoted the accuracy of judgments of Neuroticism, but the non-valid gender stereotypes hindered the accuracy of judgments of Agreeableness (Gosling, Ko, Mannarelli, & Morris, 2002). We had no data on observers’ perceptions of user gender so we could not determine whether their gender stereotypes had a kernel of truth to them. However, we did find men were slightly more conscientious ($M = 3.47, SE = .67$) than women ($M = 3.36, SE = .64$; $t[1350] = 1.99, p = .05$) so some accuracy could have been realized by applying gender stereotypes to judgments of Conscientiousness; some cues (gender-related, masculine, feminine, aggressive, creative, cute) did signal user gender so it is possible that observers could have made use of that information to improve their judgments of conscientiousness.

Perhaps the usernames did not contain sufficient diagnostic information for the observers to distill into accurate judgments. Funder’s Realistic Accuracy Model (RAM; Funder, 1995; Funder, 1999) proposes that accurate judgments are contingent on targets emitting behavior that is diagnostic of the trait, and that behavior is not only available for observers to detect, but that it is then actually detected by observers. The cue-validity correlations in Tables 2 and 3 suggest there was virtually no diagnostic information (i.e., “good information” in RAM) available for the observers to use.

Why might the present findings differ from those found in research on email addresses? One possible explanation could lie in discrepancies between the identities of the off-line WoW user and on-line WoW character. The present study focused on targets’ personalities, not the personalities of the targets’ characters. So further work should investigate the match between players’ own personalities and those of their characters or between username-based perceptions and how the player wants to be perceived.

The fact that games such as WoW are fantasy-based, opens the door for creativity and flexibility in engaging in behaviors that are not common to daily life (e.g., killing opponents, commanding a starship). At the same time, many individuals devote enormous quantities of time to game-play. So these extensive immersive experiences could encourage individuals to develop new behaviors and skills they might not have acquired while interacting in the real world.

Previous work shows that people tend to present themselves the same online as they are seen in offline contexts (Back et al., 2010; Marcus, Machilek, & Schütz, 2006; Vazire & Gosling, 2004). However, these studies have been done only in contexts where targets are known by the same people both on and offline, placing

---

### Table 3

<table>
<thead>
<tr>
<th>Cue validity</th>
<th>Cues (&quot;lens&quot;)</th>
<th>Cue utilization</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>E</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>Non-WoW</td>
<td>WoW</td>
</tr>
<tr>
<td>Gender-related</td>
<td>.00</td>
<td>.01</td>
</tr>
<tr>
<td>Masculine</td>
<td>.02</td>
<td>.02</td>
</tr>
<tr>
<td>Feminine</td>
<td>-.02</td>
<td>-.04</td>
</tr>
<tr>
<td>Aggressive</td>
<td>.05</td>
<td>.00</td>
</tr>
<tr>
<td>Fantasy</td>
<td>-.02</td>
<td>-.07</td>
</tr>
<tr>
<td>Creative</td>
<td>.03</td>
<td>.00</td>
</tr>
<tr>
<td>Cute</td>
<td>.01</td>
<td>.01</td>
</tr>
<tr>
<td>Sexual</td>
<td>.01</td>
<td>-.01</td>
</tr>
<tr>
<td>Self-enhancing</td>
<td>.07</td>
<td>-.01</td>
</tr>
<tr>
<td>Funny</td>
<td>-.02</td>
<td>-.01</td>
</tr>
<tr>
<td>Geeky</td>
<td>-.03</td>
<td>-.03</td>
</tr>
<tr>
<td>Emotional</td>
<td>.03</td>
<td>-.00</td>
</tr>
<tr>
<td>Self-deprecating</td>
<td>.02</td>
<td>-.02</td>
</tr>
<tr>
<td>Sci-fi</td>
<td>-.05</td>
<td>-.04</td>
</tr>
<tr>
<td>Multiple words</td>
<td>.03</td>
<td>.03</td>
</tr>
</tbody>
</table>

Notes: E = Extraversion; A = Agreeableness; C = Conscientiousness; ES = Emotional Stability; O = Openness.

Coefficients represent Pearson correlations. Cue validity is the correlation between self-reports of the Big Five dimensions and the presence of a cue. Cue utilization is the correlation between observer ratings of personality (from non-WoW player observers, WoW player observers, and all observers combined) and the presence of a cue.

$p < .05$.

$^*$ $p < .01$.
accountability constraints on users that might not be found in MMORPG contexts. Research comparing perceptions of individuals from acquaintances entirely within and entirely outside of their virtual interactions could shed light on the factors driving the discrepancies between MMORPG findings reported here and those that have focused on other virtual contexts.

5. Conclusions

Usernames alone do not provide accurate information about MMORPG targets’ personalities, but they do provide cues that elicit agreement among observers’ impressions of the users. With daily interactions increasingly moving to online contexts, research such as the present study is needed to reveal where impressions of others can be trusted and where – as in the case of WoW – they cannot.

Acknowledgments

We would like to thank Austin Harley, Mario Alvarado, Erica Baranski, Amelia Fritz, Alex Gonzalez, Karl Jun, Taylor Lightfoot, Lucy Liu, Arian Mobasser, Mustafa Moon, Jacob Moore, Lizzie Samuel, Bryan Song, Ashley Vivlamore, Robert Wilson, Jonathan Wu, and Crystal Yang for their help with data collection.

References