Notes about the Logistic article critique

Overall, this article is much better than the last one, although it does still have a few issues. Some things you should have noted:

The author uses the chi-square test (results in Table 2) and logistic regression models (results in Table 3)

1. He creates a logistic regression model on appropriate data: binary response variables (binge drinking, marijuana use, and illicit drug use) with binary explanatory variable (participation in a certain sport).  
   Assumptions/conditions:
   1. Randomness -- met. See p. 368 where he discusses the data set at length.
   2. Independence – possibly met? Since the data is randomly chosen, we can assume one student’s responses don’t affect another’s responses. However, if there happens to be two students in the data set who are on a team *together*, their responses could be dependent. If this just happened a handful of times, it wouldn’t bias the data too much overall. But we don’t have enough information about the final data set (the actual athletes) to know whether this is something to worry about.
   3. Linearity -- met. For example, the author is comparing binge drinking for men in football vs. men not in football. Thus, this is a binary predictor: “football player” or “not football player”. In these cases, you don’t need to worry about linearity because there are only two data points, and two data points always make a perfect line, by definition. (See p. 472 of *Stat2*.)
2. We hadn’t talked about the chi-square test when you turned this in, so I don’t expect you to discuss it specifically (although most of you have probably seen it). For what it’s worth, though, the author did use this test appropriately, for a situation with a categorical explanatory variable (“other” vs. “interest”) and categorical response variables (binge drinking or not; marijuana use or not; illicit drug use or not). Although he doesn’t state explicitly whether the condition of having expected cell counts of at least 5 is met, the overall sample sizes are large enough that I don’t think this is a problem.
3. Sampling: The larger sample (14,000 students) from which this is taken is stated to be representative. However, the final data set of athletes is quite small (2,316) and, more importantly, he ends up with some very small samples in certain sports. I am uncomfortable making a conclusion about all (collegiate, US) male hockey players based on the responses of 65 people. Also, the responses are based on a *20-page* questionnaire. Survey fatigue is a real concern and we have to worry about the veracity and representativeness of the responses.
4. For both types of tests, the author does not report the p-values, he just labels things as “significant” or “not significant” at the 0.05, 0.01, or 0.001 level. It’s better to report the actual p-values so readers can make their own conclusions based on the “strength of the evidence” against Ha.
5. I’ve got some issues with his interpretation of the odds ratios...   
   First of all, he’s not explicit about what the odds ratio is, or what it’s comparing. However, it’s reasonable to assume that any knowledgeable reader of this article is going to know what an odds ratio is. And through his interpretations, he makes it clear that the comparison groups are the other athletes that don’t play the sport of interest (like “Interest” vs “Other” from Table 2).  
   When he interprets the ORs, he uses the phrase “more likely” in a way that is misleading, because of how people interpret that phrase. Suppose the OR=1.72: we try to stay away from saying that it’s “72% more likely”, because most people would interpret that as being about *probabilities*, not odds.
6. Table 1 bugs me. These are proportions, so what does “SD” mean?
7. One more note: Several of you took issue with the author’s definition of binge drinking and/or his “re-classification” of these outcomes as binary, rather than allowing for ordered or even numerical outcomes. (For example, 20 drinks at a time is different than 5 drinks at a time, so why count these as “equal”?). First of all, this is **not** author’s definition of binge drinking: it is the **medical** definition of binge drinking, which is well-known and standardized. Secondly: What makes you think that the data provided to him included the numerical values? It’s quite likely that the survey results were already binary, and there would be no way for him to recreate the students’ “actual” (numerical) drinking habits. We are often constrained by the data we have and we must do the best we can. I imagine some of you are learning this in your individual projects!   
   As another example of this, think about the NHANES data set and the racial categories: white, Black, Mexican, Hispanic, and Other. How weird! Why these groups? There are a sizeable percentage of Asian-Americans in this country, why is that not included as an option? Why are “Mexican” and “Hispanic” separate categories? (First of all, “Mexican” is a national identity, not a race. And secondly, people of Mexican heritage would also be “Hispanic”.) But these were the categories within the data – you, as the data analyst, could not change them; all you can do is recognize their limitations. Ford may have found himself in the same position: with drinking classified as “binge” or “not binge”, and no numerical values included.