Mon. Not. R. Astron. Soc. **000**, 1–28 (2012) Printed 1 March 2013 (MN LATEX style file v2.2)

Galaxy Zoo 2: detailed morphological classifications for 304,122 galaxies from the Sloan Digital Sky Survey

Kyle W. Willett [1⋆], Chris J. Lintott[2,7], Steven P. Bamford[3], Karen L. Masters[4], Brooke D. Simmons[2], Kevin Schawinski[5], Lucy Fortson[1], Robert J. Simpson[2], Ramin A. Skibba[6], Edward M. Edmondson[4], Arfon M. Smith[2,7], Robert C. Nichol[4], Kevin R.V. Casteels[3,8]

[1] *University of Minnesota, USA*
[2] *University of Oxford, UK*
[3] *University of Nottingham, UK*
[4] *University of Portsmouth, UK*
[5] *ETH, Zürich, Switzerland*
[6] *University of California San Diego, USA*
[7] *Adler Planetarium, USA*
[8] *University of Barcelona, Spain*

* We might want to say here that after learning how effective the GZ1 volunteers were at a simple classification task, we wanted to see what they could do with a more complex classification tree.

Accepted 1988 December 15. Received 1988 December 14; in original form 1988 October 11

ABSTRACT

Morphology is a powerful and unique probe for quantifying the dynamical history of a galaxy. However, automatic classifications of morphology (either by computer analysis of images or by using other physical parameters as proxies) still have drawbacks when compared to visual inspection. The number of galaxies in large samples makes this impractical for individual astronomers. Galaxy Zoo 2 (GZ2) is a citizen science project that provides morphological classifications of more than 300,000 galaxies drawn from the Sloan Digital Sky Survey. These include all galaxies in the DR7 Legacy survey down to $r \geq 17$, along with deeper classifications of galaxies in Stripe 82. The original Galaxy Zoo project primarily separated galaxies only into early- or late-types; GZ2 classifies finer morphological features. These include the presence of bars, bulges, edge-on disks, and merging galaxies, as well as quantifying the strength of multiplicity of features such as galactic bulges and spiral arms. This paper presents the full data release for the project, including measures of classification accuracy and user bias. We show that the majority of GZ2 classifications agree with those made by salaried astronomers, especially for T-types, strong bars, and arm curvature. Both raw and reduced data products are fully available and can be obtained in electronic format at `http://data.galaxyzoo.org`.

The Galaxy Zoo 2 sample includes...

galaxies selected from the deeper imaging of SDSS Stripe 82.

(Also, is this all Stripe 82 galaxies, or just some?)

5 visual inspection of each galaxy

features

Not clear what this means

Key words: catalogues, methods: data analysis, galaxies: general, galaxies: spiral, galaxies: elliptical and lenticular

1 INTRODUCTION

The Galaxy Zoo project (Lintott et al. 2008) was launched in 2007 to provide morphological classifications of nearly one million galaxies drawn from the Sloan Digital Sky Survey (York et al. 2000). This scale of effort was made possible by combining classifications from hundreds of thousands of volunteers, but in order to keep the task to a manageable size only simple morphological distinctions were initially requested, essentially dividing systems into elliptical, spiral and merger. This paper presents data and results from that project's successor, Galaxy Zoo 2 (GZ2), which collected more sophisticated morphological classifications for more than 250,000 of the brightest SDSS galaxies.[1] *

Maybe say "a manageable level of complexity"?

While the morphological distinction used in

* E-mail: willett@physics.umn.edu

[1] `http://zoo2.galaxyzoo.org`

Galaxy Zoo 1 (GZ1) – that which divides spiral and elliptical systems – is the most fundamental, there is a long history of finer-grained morphological classification. The first systematic approach to classification (Hubble 1936) included a division between barred and unbarred spirals, creating the famous 'tuning fork', and further distinctions based on the shape of early-type systems or tightness of spiral arms. These finer distinctions are correlated with physical parameters of the systems being studied; the presence of a bar, for example, may drive gas inwards and be correlated with the growth of a central bulge (a review is given in Kormendy & Kennicutt (2004) and an updated picture by Masters et al. 2011). Similarly, the presence of a central bulge is likely to indicate a history of mass assembly through significant mergers (Martig et al. (2012) and references therein) and so on. Careful classification of morphological features is thus essential if the assembly and evolution of the galaxy population is to be understood.

Whereas traditional morphological classification relied on the careful inspection of small numbers of images by experts (e.g., Sandage 1961; de Vaucouleurs et al. 1991), the sheer size of modern data sets make this approach impractical. The largest detailed professional classification effort to date was undertaken by Nair & Abraham (2010a), who provide classifications of $\sim$ 14000 systems. The present study includes an order of magnitude more systems, allowing for a more careful study of the relationships and interdependence of such small scale morphological features.

The use of proxies for morphology such as colour, concentration index, spectral features, surface brightness prole, structural features, spectral energy distribution or some combination of these is not an adequate substitute; each proxy has an unknown biased relation with the morphological features under study. The complexity of the relationship between these variables is, rather, the main reason for requiring such a large set of classifications, as one can only isolate significant numbers of red, barred, bulgeless spirals in the field (to give one example) with a sufficiently comprehensive starting set.

Despite recent advances in automated morphological classification, driven in part by the availability of large training sets from the original Galaxy Zoo (Banerji et al. 2010; Huertas-Company et al. 2011; Davis & Hayes 2013) XXX ADD MORE REFS XXX, the state of the art does not provide an adequate substitute for classification by eye. In particular, as Lintott et al. (2011) note such efforts typically use proxies for morphology as their input, and so they suffer equally from the objections raised above to the use of morphological proxies. The release of the dataset associated with this paper will be of interest to those developing such machine learning and computer vision systems.

These results have been made possible by the participation in the Galaxy Zoo project by hundreds of thousands of 'citizen scientists'. Since the original Galaxy Zoo demonstrated the utility of this method in producing both scientifically-useful catalogues and serendipitous discoveries (see Lintott et al. (2011) for a review of Galaxy Zoo 1 results), this method has been expanded beyond simple classifications to use cases which include exoplanet discovery (Fischer et al. 2012; Schwamb et al. 2012) and a census of bubbles associated with star formation (Simpson et al. 2012) amongst many others.

2 PROJECT DESCRIPTION

2.1 Sample selection

The primary sample of objects classified for Galaxy Zoo 2 comprised roughly the brightest 25% of the resolved galaxies in the SDSS North Galactic Cap region. The goal was to exclude the most distant, faintest and smallest systems within which fine morphological features would not be resolved. Our sample was restricted to the SDSS DR7 'Legacy' catalogue (Abazajian et al. 2009), and therefore excludes observations made by SDSS for other purposes, such as the SEGUE survey.

Several cuts were applied to the DR7 Legacy sample for selection in GZ2. We require a Petrosian magnitude brighter than 17.0 in the r-band (after Galactic extinction correction was applied), along with a r-band Petrosian radius greater than 3 arcsec. Galaxies which had a spectroscopic redshift in the DR7 catalogue outside the range $0.0005 < z < 0.25$ were removed; however, galaxies without reported redshifts were kept. Finally, objects which are flagged by the SDSS pipeline as SATURATED, BRIGHT or BLENDED without an accompanying NODEBLEND flag were also excluded. The 245,609 galaxies satisfying these criteria are referred to as the "original" sample.

An error in the original query meant that the "original" sample initially missed some objects on launch, specifically those flagged as both BLENDED and CHILD. These galaxies, which are typically slightly brighter, larger and bluer than the general population, were added to the site on 2009-09-02. The rate at which these images were shown to users was tuned so that the average number of classifications would catch up to those in the "original" sample. These additional 28,174 galaxies are referred to as the "extra" sample.

In addition to the sample from the Legacy survey, we later added images from Stripe 82, a section along the celestial equator in the Southern Galactic Cap. The selection criteria are the ~~sample~~ as that for the Legacy galaxies, with the exception of a fainter magnitude limit of $r < 17.77$. For the Stripe 82 sample only, we included multiple images of individual galaxies: one set of images at single-depth exposures, and two sets of co-added images with multiple exposures. Coadded images combined 47 (south) or 55 (north) separate scans of the region, resulting in an object detection limit approximately two magnitudes lower than in normal imaging (Annis et al. 2011).

The primary sample for GZ2 analysis consists of the combined "original", "extra", and the Stripe 82 normal-depth images with $r \leqslant 17.0$. We have verified that there are no significant differences in classifications between these samples that could have been caused, for example, by a time-dependent bias. This is hereafter referred to as the GZ2 **main sample** (Table 1). Data from both the Stripe 82 normal-depth images with $r > 17.0$ and the two sets of coadded images are included as separate data products.

2.2 Image creation

Images of galaxies from the Legacy and Stripe 82 normal depth surveys were generated from the SDSS ImgCutout web service (Nieto-Santisteban, Szalay & Gray 2004). Each image is a *gri* colour composite 424×424 pixels in size, scaled to (0.02×petror90_r) arcsec/pixel.

Table 1. GZ2 sample properties

Sample	$N_{galaxies}$	$N_{class.}$ median	m_r depth [mag]
original	245,609	44	17.0
extra	28,174	41	17.0
Stripe 82 normal	21,522	45	17.77
Stripe 82 normal (mag-limited)	10,188	45	17.0
Stripe 82 coadd 1	30,346	18	17.77
Stripe 82 coadd 2	30,339	21	17.77
main (original + extra + S82 maglim)	283,971	44	17.0

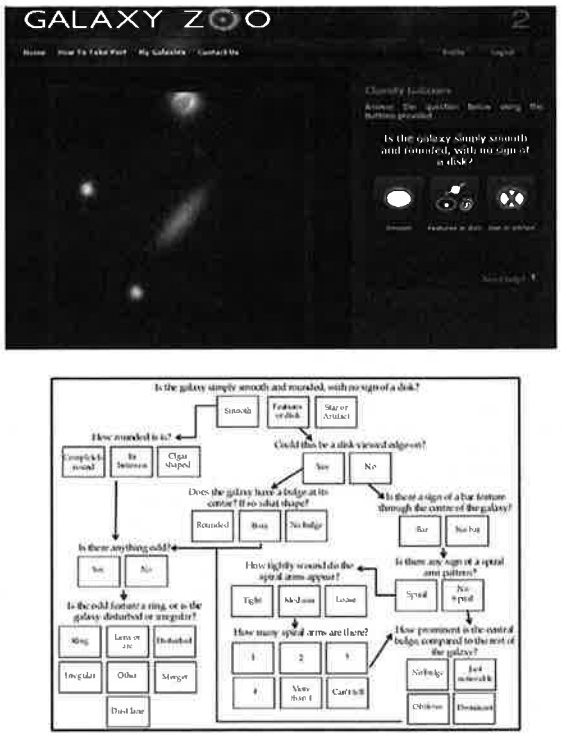


Figure 1. *Top:* Front page of the web interface for Galaxy Zoo 2, displaying Task 01. *Bottom:* Flowchart of the 11 classification tasks for GZ2, beginning at the top centre.

Coadded images from Stripe 82 were generated from the corrected SDSS FITS frames in g, r and i. Frames were stitched together using Montage[2] and converted to a colour image using a slightly modified version of the asinh stretch code (Lupton et al. 2004), with parameters adjusted to try to replicate normal SDSS colour balance. The parameterisation of the stretch function used is:

$$f(x) = \mathrm{asinh}(\alpha Q x)/Q \qquad (1)$$

where $Q = 3.5$ and $\alpha = 0.06$. The colour scaling is [1.000,1.176,1.818] in g, r and i, respectively.

The first set of coadded images were visually very different from the normal SDSS images. The background sky noise was high and colour saturated in many individual pixels. Since we were concerned that this could affect morphological classifications, we created a second set of coadd images with a desaturated background. The two coadded sets are labeled "stripe82_coadd_1" and "stripe82_coadd_2", respectively (Table 1).

2.3 Decision tree

Data for Galaxy Zoo 2 was collected via a web-based interface. Users of the interface needed to register with a username for their clicks to be recorded, but were not required to complete any tutorials. They were then shown a *gri* colour composite image of a galaxy for classification (Figure 1). Users had the option to invert the default colour scaling on any image being classified.

Classification of the galaxies proceeds via a multi-step tree. Each classification begins with a slightly modified version of the original Galaxy Zoo task, with users identifying whether the galaxy is either "smooth", has "features or a disk", or is a "star or artifact" in the image. Subsequent questions depend on the user's previous responses. For example, if the user clicks on the "smooth" button, they are subsequently asked to classify the roundness of the galaxy; this question will not be asked if they select either of the other two options.

The Galaxy Zoo 2 tree has 11 classification tasks with a total of 37 possible responses (Table 2). A classifier selects only one option for each task, after which they are immediately taken to the next step in the tree. Task 01 is the only question answered for all objects in the sample. Once a classification was completed, an image of the next galaxy is automatically displayed and the user can begin on a new object.

Data from the classifications was stored in a live Structured Query Language (SQL) database. In addition to the morphology classifications, we also registered the timestamp, user identification, and galaxy identification for each asset in the database.

Galaxy Zoo 2 was launched on 2009-02-16 with the "original" sample of 245,609 images. The "extra" galaxies from the Legacy survey were added on 2009-09-02. The normal-depth and first coadded Stripe 82 images were mostly added on 2009-09-02, with an additional $\sim$ 7700 of the coadded images added on 2010-09-24. Finally, the second version of the coadded images were added to the site on 2009-11-04.

For most of the duration of Galaxy Zoo 2, all images were shown to classifiers in a random order. However, we desired to have all galaxies classified a minimum number of times. Therefore, in the final period of Galaxy Zoo 2, accompanied by a competition with a running tally (dubbed the Zoonometer), objects with low numbers of classifications were shown to users at a higher rate. The goal was a minimum of 40 classifications for the "original", "extra" and "stripe82" samples, and 20 for the "stripe82_coadd_2" sample. The "stripe82_coadd_1" sample was removed from the site at this time. The main sample galaxies finished with a median of 44 classifications, with 27% having fewer than 40; the "stripe82_coadd_2" galaxies had a median of 21 classifications and 26% of them had fewer than 20 (Table 1).

[2] `http://montage.ipac.caltech.edu`

Expand this caption for people skimming tables rather than reading the full text

Table 2. GZ2 classification tree

Task	Question	Responses	Next
01	*Is the galaxy simply smooth and rounded, with no sign of a disk?*	smooth features or disk star or artifact	→ 07 → 02 **end**
02	*Could this be a disk viewed edge-on?*	yes no	→ 09 → 03
03	*Is there a sign of a bar feature through the centre of the galaxy?*	yes no	→ 04 → 04
04	*Is there any sign of a spiral arm pattern?*	yes no	→ 10 → 05
05	*How prominent is the central bulge, compared with the rest of the galaxy?*	no bulge just noticeable obvious dominant	→ 06 → 06 → 06 → 06
06	*Is there anything odd?*	yes no	→ 08 **end**
07	*How rounded is it?*	completely round in between cigar-shaped	→ 06 → 06 → 06
08	*Is the odd feature a ring, or is the galaxy disturbed or irregular?*	ring lens or arc disturbed irregular other merger dust lane	**end** **end** **end** **end** **end** **end** **end**
09	*Does the galaxy have a bulge at its centre? If so, what shape?*	rounded boxy no bulge	→ 06 → 06 → 06
10	*How tightly wound do the spiral arms appear?*	tight medium loose	→ 11 → 11 → 11
11	*How many spiral arms are there?*	1 2 3 4 more than four can't tell	→ 05 → 05 → 05 → 05 → 05 → 05

The last GZ2 classifications were collected on 2010-04-29, spanning a period of just over 14 months. The final dataset contained 16,340,298 classifications (comprising a total of 58,719,719 questions) by 83,943 ~~participants.~~ volunteers

3 DATA REDUCTION

3.1 Multiple classifications

In a small percentage of cases, an individual user may classify the same object more than once. Since we wish to treat each click as an independent measurement, we removed multiple classifications of the same object by a given user from the data, keeping only the last submitted classification. Such repeats only occurred for a small proportion of objects (~ 1%), with only a tiny proportion (~ 0.01%) occurring enough to significantly alter the final vote fractions.

Not clear what this sentence means, and it sounds important

3.2 Consistency and invididual user weighting

The next step in reducing the data is to remove the influence of unreliable users. To do so we applied an iterative weighting scheme. First, we calculated the vote fraction ($f_r = n_r/n_{task}$) for every answer for every task for every object, weighting each user's vote equally. Here, n_r is the number of clicks for a given answer and n_{task} is the total number of clicks for that task. Individual clicks are then compared to the vote fraction to calculate its consistency κ:

Is consistency a property of users, or tasks, or galaxies??

$$\kappa = \frac{1}{N_r}\sum_i \kappa_i, \quad (2)$$

Clicks from individual users

where N_r is the total number of possible reponses for a task and:

Does this sum over users or tasks?

$$\kappa_i = \begin{cases} f_r & \text{if click corresponds to this answer,} \\ (1 - f_r) & \text{if click does not correspond.} \end{cases} \quad (3)$$

For example, if a question has three possible answers, and the galaxy corresponds best to answer a, then the vote fractions for answers (a, b, c) might be $(0.7, 0.2, 0.1)$.

- If an individual selected answer a, then $\kappa = (0.7 + (1 - 0.2) + (1 - 0.1))/3 = 0.8$
- If an individual selected answer b, then $\kappa = ((1 - 0.7) + 0.2 + (1 - 0.1))/3 = 0.467$
- If an individual selected answer c, then $\kappa = ((1 - 0.7) + (1 - 0.2) + 0.1)/3 = 0.4$

Clicks which agree with the majority thus have high values of consistency, whereas clicks which disagree have low values.

Based on the distribution of results for the initial iteration of κ (Figure 2), we chose a weighting function that down-weighted users in the tail of low consistency:

$$w = \text{power}((\kappa/0.6), 8.5) \quad (4)$$

For this function, $w = 1$ for ~ 95% of users and $w < 0.01$ for only ~ 1% of users. The vast majority of users are thus treated equally: there is no up-weighting of the most consistent users. The top panel of Figure 2 also shows the lowest-weighted users have on average classified only a handful of objects.

How does the top panel of Fig 2 show this?

After computing κ for all tasks, the vote fractions were recalculated using the new user weights. We repeated this process a third time to ensure convergence. For each task, this produces both a weighted number of votes and a weighted vote fraction for each task.

3.3 Classification bias

The weighted vote fractions in the data are also adjusted for what we term *classification bias*. The overall effect is a change in observed morphology fractions as a function of redshift, a trend seen in the original Galaxy Zoo 1 data (REF). The presumed cause is that more distant galaxies are, on average, both smaller and dimmer as they appear in the

* we should say something here to address the criticism that says, "just because you agree with the majority doesn't mean you're right"

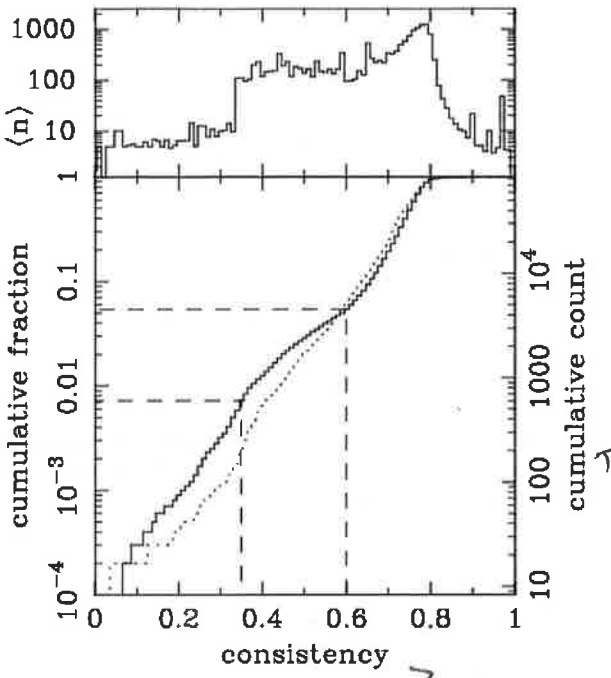


Figure 2. User consistency for Task XX. Top: total number of galaxies classified per user as a function of their consistency. Bottom: Cumulative fraction distribution of consistency. The dotted line shows the first iteration of weighting, and the solid line the third iteration. The second iteration is not shown, but is almost identical to the third.

cutout images; as a result, finer morphological features are more difficult to identify.

Figure 3 demonstrates the classification bias for several of the Galaxy Zoo 2 classification tasks. The average weighted vote fraction for each response (*thin lines*) is shown as a function of redshift; the fraction of votes for finer morphological features (such as identification of disk galaxies, spiral structure, or galactic bars) decrease at higher redshift. The trend is strongest for the initial classification of smooth and feature/disk galaxies, but almost all tasks exhibit some level of change. Part of this effect is due to the nature of a luminosity-limited sample; high-redshift galaxies must be more luminous to be detected in the SDSS and are thus more likely to be giant red ellipticals. However, we see evidence of the classification bias even in magnitude-limited samples. Since this bias contaminates any potential studies of galaxy demographics over the entire volume of the sample, it must be corrected to the fullest possible extent.

Bamford et al. (2009) corrected for classification bias in the original Galaxy Zoo data, but only for the elliptical and combined spiral variables. Their approach was to bin the galaxies a function of absolute magnitude (M_r), the physical Petrosian half-light radius (R_{50}), and redshift. They then measure the average elliptical-to-spiral ratio for each (M_r,R_{50}) bin in the lowest available redshift slice; this yields a local baseline relation which gives the (presumably) unbiased morphology as a function of the galaxies' *physical*, rather than *observed* parameters. From the local relation, they derive a correction for each (M_r,R_{50},z) bin and then adjust the vote fractions for the individual galaxies in each bin. The validity of this approach is justified in part by the agreement of these debiased probabilities with a monotonic morphology-density relation (Bamford et al. 2009). We modify and extend this technique for the Galaxy Zoo 2 classifications as described below.

There are two major differences between the GZ1 and GZ2 data. First, GZ2 has a decision tree, rather than a single question and answer for each click on an image. This means that all tasks, with the exception of the first, depend on answers to previous classifications in the tree. For example, the bar question is only asked if the user classifies a galaxy as having "features or disk" and as "not edge-on". Thus, the value of the weighted vote fraction for this example task only addresses the total bar fraction *among face-on disk galaxies*, and not as a function of the general population.

Our approach is to examine only biases within the context of the individual classification tasks. The corrections used to debias each task are derived based only galaxies with sufficient votes to characterize that feature. We employ a combination of threshold on the weighted vote fraction for preceding tasks as well as a lower limit on the total number of votes for a galaxy to be used in deriving a correction. While this increases the number of noisy bins, it is critical for reproducing accurate baseline measurements of individual morphologies. The adjustment derived from well-classified galaxies is then applied to the vote fractions for *all* galaxies in the sample.

The second major issue is the adjustment of the GZ1 vote fractions assumed that the single task was essentially binary. Since almost every vote in GZ1 was either for "elliptical" or "spiral" (either anticlockwise or clockwise), they were able to use that ratio as the sole metric of the morphology. No systematic debiasing was done for the other GZ1 response options ("star/don't know", "merger", or "edge on/unclear"), and the method of adjusting the vote fractions assumes that these do not significantly affect the classification bias for the most popular responses.

Vote fractions for each galaxy are adjusted for classification bias using the following method. The method relies on the assumption that for a galaxy of a given physical brightness and size, a sample of other galaxies with similar brightnesses and sizes will (statistically) share the same average morphologies for a given task. We represent this as the ratio of vote fractions (f_i/f_j) for responses i and j. Finally, we assume that the true (that is, unbiased) ratio of likelihoods for each task (p_i/p_j) is related to the measured ratio via a single multiplicative constant:

$$\frac{p_i}{p_j} = \frac{f_i}{f_j} \times K_{j,i}. \quad (5)$$

In this case, the adjusted likelihood for a single task is written as:

$$p_i = \frac{1}{1/p_i}, \quad (6)$$

and the sum of all the likelihoods for a given task must be unity:

$$p_i + p_j + p_k + \cdots = 1. \quad (7)$$

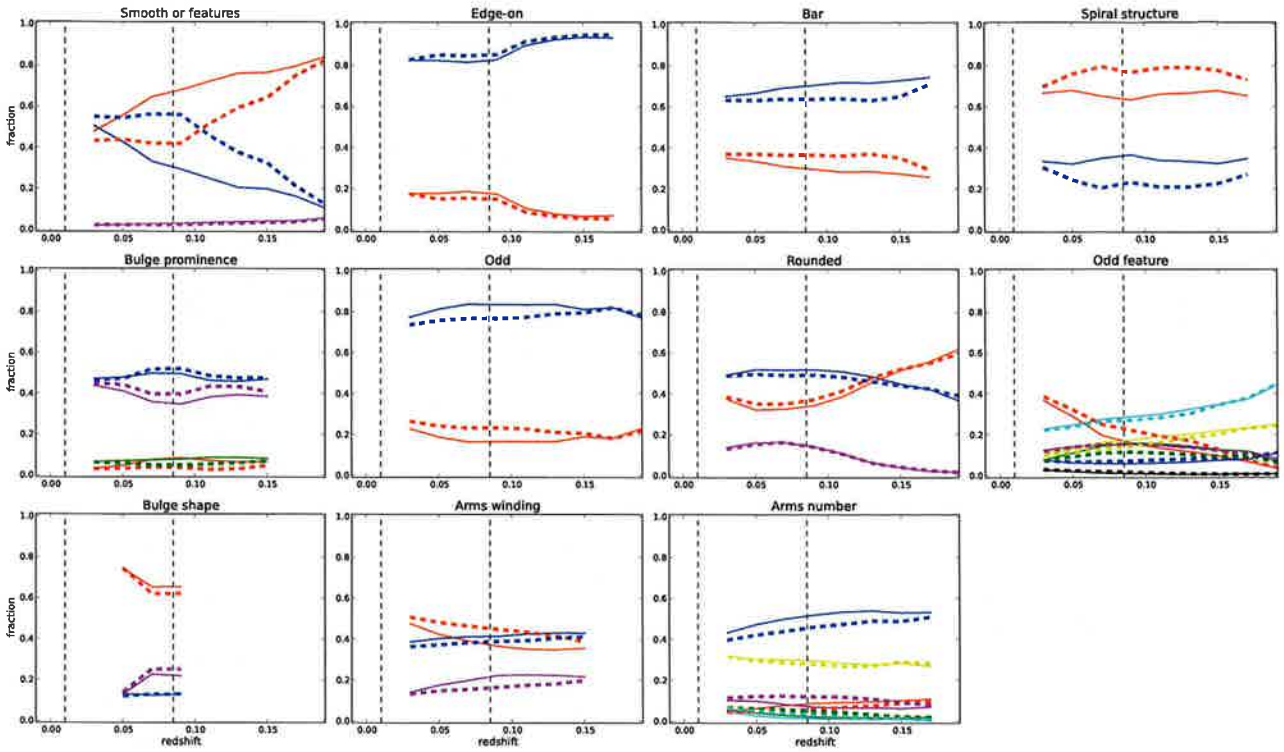


Figure 3. Type fractions for the classification tasks in GZ2. Solid (thin) lines show the weighted vote fractions, while the thick (dashed) lines show the debiased vote fractions that have been adjusted for classification bias. This is a magnitude-limited sample for $M_r < -20.17$. Vertical dashed lines show the redshift at $z = 0.01$ (the lower limit of the correction) and $z = 0.085$ (the redshift at which the absolute magnitude limit reaches the sensitivity of the SDSS).

Multiplying (6) by (7) yields:

$$p_i = \frac{1}{1/p_i} \times \frac{1}{p_i + p_j + p_k + \dots} \quad (8)$$

$$p_i = \frac{1}{p_i/p_i + p_j/p_i + p_k/p_i + \dots} \quad (9)$$

$$p_i = \frac{1}{\sum_{j \neq i} (p_j/p_i) + 1} \quad (10)$$

$$p_i = \frac{1}{\sum_{j \neq i} K_{j,i}(f_j/f_i) + 1}. \quad (11)$$

The corrections for each pair of tasks can be directly determined from the data. At the lowest sampled redshift bin ($z \simeq 0$), $\frac{p_i}{p_j} = \frac{f_i}{f_j}$ and $K_{j,i} = 1$. From Equation 5:

$$\left(\frac{f_i}{f_j}\right)_{z=0} = \left(\frac{f_i}{f_j}\right)_{z=z'} \times K_{j,i} \quad (12)$$

$$K_{j,i} = \left(\frac{f_i}{f_j}\right)_{z=z'} / \left(\frac{f_i}{f_j}\right)_{z=0} \quad (13)$$

(14)

This can be simplified if we define $C_{j,i} \equiv \log_{10}(K_{j,i})$:

$$C_{j,i} = \log\left[\left(\frac{f_i}{f_j}\right)_{z=z'} / \left(\frac{f_i}{f_j}\right)_{z=0}\right] \quad (15)$$

$$C_{j,i} = \log\left(\frac{f_i}{f_j}\right)_{z=z'} - \log\left(\frac{f_i}{f_j}\right)_{z=0}. \quad (16)$$

So the correction $C_{j,i}$ for any bin is simply the difference between f_i/f_j at the desired redshift and between that of a local baseline, where the ratios between vote fractions are expressed as logarithms.

The local baselines and subsequent corrections are derived from the main sample data (original + extra + magnitude-limited Stripe 82). Since determining the baseline ratio relies on absolute magnitude and physical size, we only use the 86% of galaxies in the main sample with spectroscopic redshifts. We also use only galaxies with sufficient numbers of classifications to determine the morphology ratios. This varies as a function of the task – for the questions asked of every galaxy (Tasks 01 and 06), we set the minimum number of classifications at 30. This is well below the median of 43, and includes > 97% of the sample. For other tasks with fewer total responses, this can be as low as 10 classifications per task.

The weighted vote fractions for each task response are binned in three dimensions: the absolute magnitude M_r, the Petrosian r-band half-light radius R_{50}, and redshift z. Bins range for M_r range from -24 to -16 in steps of 0.25 mag, for R_{50} from 0 to 15 kpc in steps of 0.5 kpc, and for z from

0.01 to 0.26 in steps of 0.01. The bin ranges and step sizes are chosen to maximize the phase space covered by the bias correction, while also retaining enough galaxies in each bin to establish its morphology distribution. The value of each bin in the cube is the sum of the weighted vote fractions for that response. For each pair of responses (i, j) to a question, we compute $\log(f_j/f_i)$ in every (M_r,R_{50},z) bin. The local baseline relation is established by selecting the value in the non-empty bin(s) for the lowest-redshift slice at a given (M_r,R_{50}).

Since each unique pair of responses to a question will have a different local baseline, there are $\binom{n}{2}$ corrections for a task with n responses. This reduces to the method with a single pair of variables described in Bamford et al. (2009) if $n = 2$.

The baseline morphology ratios for the GZ2 tasks are shown in Figure 4 for the first two responses in each task. To derive a correction for bins not covered at low redshift, we attempted to fit each baseline ratio with an analytic, smoothly-varying function. The baseline ratio for the "smooth" and "features/disk" responses to Task 01 is functionally very similar to the GZ1 relation (Figure A5 in Bamford et al. 2009), as expected. It is reasonably well-fit with an analytic function of the form:

$$\frac{f_j}{f_i}[R_{50}, M_R] = \frac{s_6}{1 + exp[(x_0 - M_R)/x_1]} + s_7 \qquad (17)$$

where:

$$\alpha = s_2^{-(s_1 + s_8 R_{50}{}^{s_9})} + s_3 \qquad (18)$$

$$\beta = s_4 + s_5(x_0 - s_3) \qquad (19)$$

and where $\{s_1, s_2, s_3, s_4, s_5, s_6, s_7, s_8, s_9\}$ are minimized to fit the data. The only other task had baseline ratios reasonably well fit by these parameters was Task 07 (rounded smooth galaxies). We adopted the same approach for this task and were able to fit the behavior of all three pairs of responses with the same functional form.

None of the other tasks are well-fit by a function of the form in Equation 17; for these, we instead adopt a simpler fit where both M_r and R_{50} vary linearly:

$$\frac{f_j}{f_i}[R_{50}, M_R] = t_1(R_{50} - t_2) + t_3(M_R - t_4) + t_5, \qquad (20)$$

and $\{t_1, t_2, t_3, t_4, t_5\}$ are the parameters to be minimized. We fit Equation 20 to all other tasks where the number of bins is sufficient to get a reasonable fit. Finally, for pairs of responses with only a few sampled bins, we instead used the direct difference between the local ratio and the measured ratio at higher redshift. Galaxies falling in bins that are not well-sampled are assigned a correction of $C_{i,j} = 0$ for that term; this is necessary to avoid overfitting based on only a few noisy bins.

The success of this method is generally good for most GZ2 tasks and responses. Figure 3 illustrates the comparison between the raw and debiased vote fractions. The debiased results (*thick lines*) are generally flat over a range of $0.01 < z < 0.085$, where L galaxies fall below the magnitude limit of the survey and the bins are more poorly sampled. The debiased early- and late-type fractions of 0.45 and 0.55 agree with the GZ1 type fractions derived by Bamford et al. (2009). The bar fraction in disk galaxies is roughly 0.35, which is slightly higher than the value found by using thresholded GZ2 data in Masters et al. (2011).

4 THE CATALOG

Other possible inclusions for catalog:

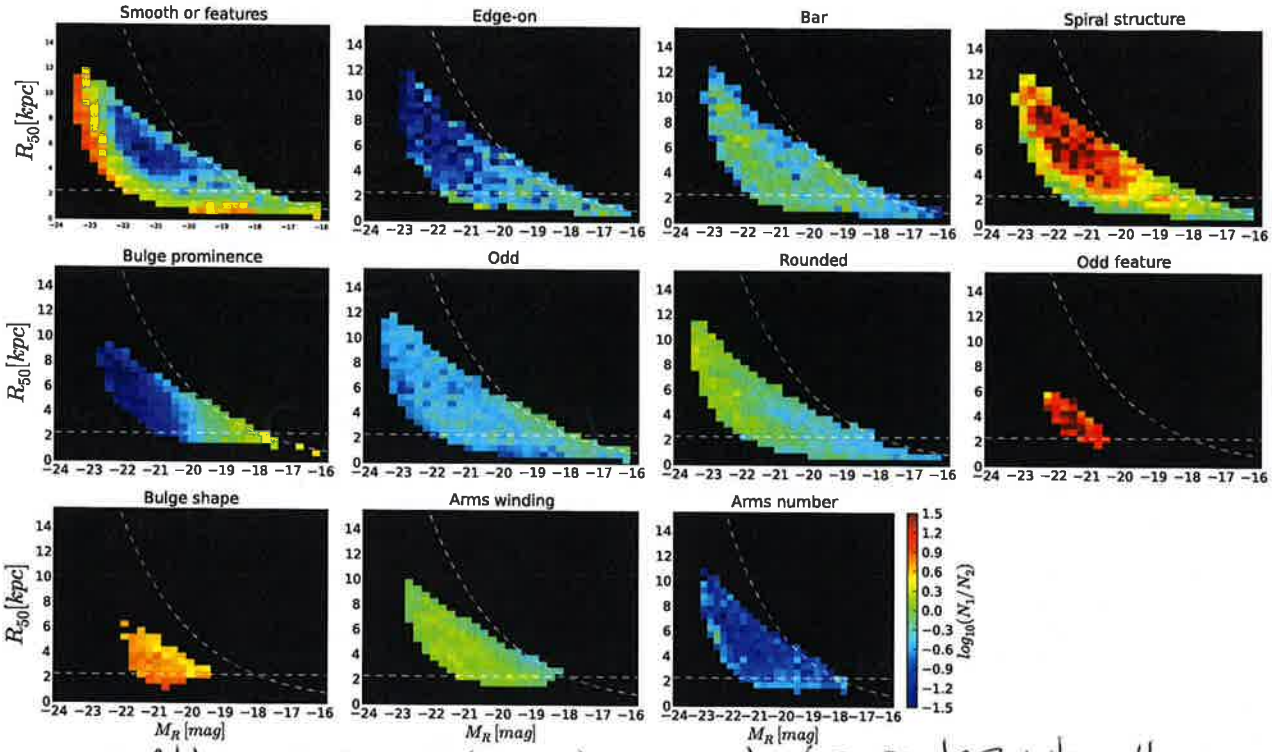
- Metrics on classification confidence (Table 04, Lintott et al. 2011)
- Galaxy Wars
- M_r,R_{50},z bins for each galaxy
- Voronoi tessellation bins
- Matched SDSS metadata

4.1 Main sample

The data release for Galaxy Zoo 2 consists of four tables, abridged portions of which appear in this paper. Table 3 contains classification data for the 283,971 galaxies in the main sample. Each galaxy is identified by its unique SDSS DR7 objID, as well as its original sample designation (either original, extra or Stripe 82 normal-depth). N_{class} is the total number of users who have classfied the galaxy, while N_{votes} gives the total number of clicks summed over all classifications and all responses. For each of the 37 morphological classes, we give six parameters: the raw number of votes for that response (eg, `t01_smooth_or_features_a01_smooth_count`), the number of votes weighted for consistency (`*_weight`), the fraction of votes for the task (`*_fraction`), the vote fraction weighted for consistency (`*_weighted_fraction`), the debiased likelihood (`*_debiased`), which is the weighted fraction adjusted for classification bias (see Section 3.3), and a boolean flag (`*_flag`) that is set if the galaxy is included in a clean, debiased sample.

Flags for each morphological parameter are determined by applying three criteria: the first is the requirement that more than 50% of votes for preceding task(s) must eventually select for the task being flagged. For example, to select clean barred galaxies, we require both $p_{features/disk} \geqslant 0.5$ and $p_{notedge-on} \geqslant 0.5$. Secondly, the object must exceed a minimum number of total votes (ranging from 5–30) for that task, in order to eliminate variance from small-number statistics. Finally, we establish a threshold value for the debiased vote fraction; this is 0.5 for Tasks 02 and 03, and 0.8 for all other tasks. GZ1 also used a debiased threshold value of 0.8, based on a correction applied to raw vote fractions at the same threshold (Bamford et al. 2009; Lintott et al. 2011).

Table 4 shows the GZ2 classifications for main sample galaxies without spectroscopic redshifts. To derive the debiased likelihoods, we used the morphology corrections derived from galaxies in the spectroscopic main sample. We then used the photometric redshift provided by the SDSS to derive M_r,R_{50} and select the appropriate correction bin. The mean redshift error in the photometric sample is $\Delta z = 0.021$ (a fractional uncertainty of 27%), compared to the spectroscopic accuracy of $\Delta z = 0.00016$ (0.3%). Since the size of the redshift bins in $C_{j,i}$ is 0.01, a shift of 2–3 bins can potentially produce a very large change in the debiased vote



Add a sentence here to remind the reader why this is important for correcting the high redshift classification bias

Figure 4. Local morphology ratios for morphology classifications in GZ2. The ratio of the binned vote fractions for morphologies is for the first two responses in the decision tree (Table 2) for each task; there may be as many as 21 such pairs for tasks with more than two options. The dashed horizontal lines give the physical scale corresponding to 1″, while the curved lines show a constant apparent surface brightness of $\mu_{50,r} = 23.0$ mag arcsec^{-2}.

between the single-view image galaxies and co-add galaxies

So what should users do about these questions?

fractions. For this reason, we separate galaxies with spectroscopic and photometric redshifts, and do not recommend that the debiased likelihoods be combined for analysis.

How many galaxies in Tables 3 and 4?

4.2 Stripe 82

classification galaxies in SDSS

Table 5 contains data for the co-added images in Stripe 82. These classifications are for the second method of co-adding exposures, and are limited to the 65% of the sample with spectroscopic redshifts. Since there are only ∼ 10% of the number of galaxies used in the main sample, there are not enough classifications to robustly derive a correction from the data itself. The debiased probabilities for this sample therefore apply corrections based on those in the GZ2 main sample. Flags for these galaxies use the same 50% criteria as thresholds for previous tasks; however, since galaxies are only classified an average of 21 times each, the number of votes per task is relaxed to 10 votes for Tasks 01 and 06 and 5 votes for all other tasks.

to that of

The distribution of votes for the main sample is similar enough to the Stripe 82 normal-depth (with $r < 17.0$) that the same bias correction applies for both. Table 6 shows the tasks in the GZ2 question tree and the mean weighted vote fraction for each response. The distributions for both the main sample and Stripe 82 galaxies are quite similar, with the difference in the mean varying by < 10% for almost all responses. The only exceptions to these are for responses that target rare objects (and thus are subject to higher variance for low-number statistics), such as dust lanes, rings, and high-multiplicity spiral arms.

imaging

this similarity

of vote fractions

The right panel of Figure 5 shows that the weighted vote fractions also behave similarly as a function of redshift, particularly in the $0.01 < z < 0.08$ range covered by the GZ1 debiasing technique. The agreement is generally good between the Stripe 82 normal depth and the GZ2 main sample; this is not the case for the coadded Stripe 82 data, however. For Task 01, fewer galaxies are classified as robustly smooth (above the 0.8 threshold), moving instead to the "unclassified" category. Coadded data showed similar higher fractions of galaxies with bars and for possessing visible spiral structure. A possible cause for this is that the new image pipeline in the coadded data allows viewers to see faint features or disks, due to improved seeing in the coadded data (from 1.4″ to 1.1″; Annis et al. 2011).

What agreement?

Ny

What is the significance of this?

For almost every response in the GZ2 decision tree, the data (no bias correction) have no systematic differences between classifications using the coadd1 and coadd2 images. Figure 6 shows distributions of the differences between the two weighted vote fractions ($\Delta_{coadd} = f_{coadd1} - f_{coadd2}$). If the mean value of Δ_{coadd} for an answer is non-zero, that would indicate a systematic bias in classification due to the image processing. In GZ2, 33/37 tasks have $|\Delta_{coadd}| < 0.05$ (for galaxies with at least 10 responses to the task), with variations in the mean scattered on both sides of Δ_{coadd}.

in vote fraction

The biggest systematic difference is for Task 05, Answer 11 (prominence of the bulge is "just noticeable"), for

* I think it's important to add a short section 4.4 here giving advice to the user on what data they should use for certain kinds of studies. As an example, see the Zoo 1 data release paper (Lintott et al 2010), end of section 4, paragraph starting "Most users of the data..."

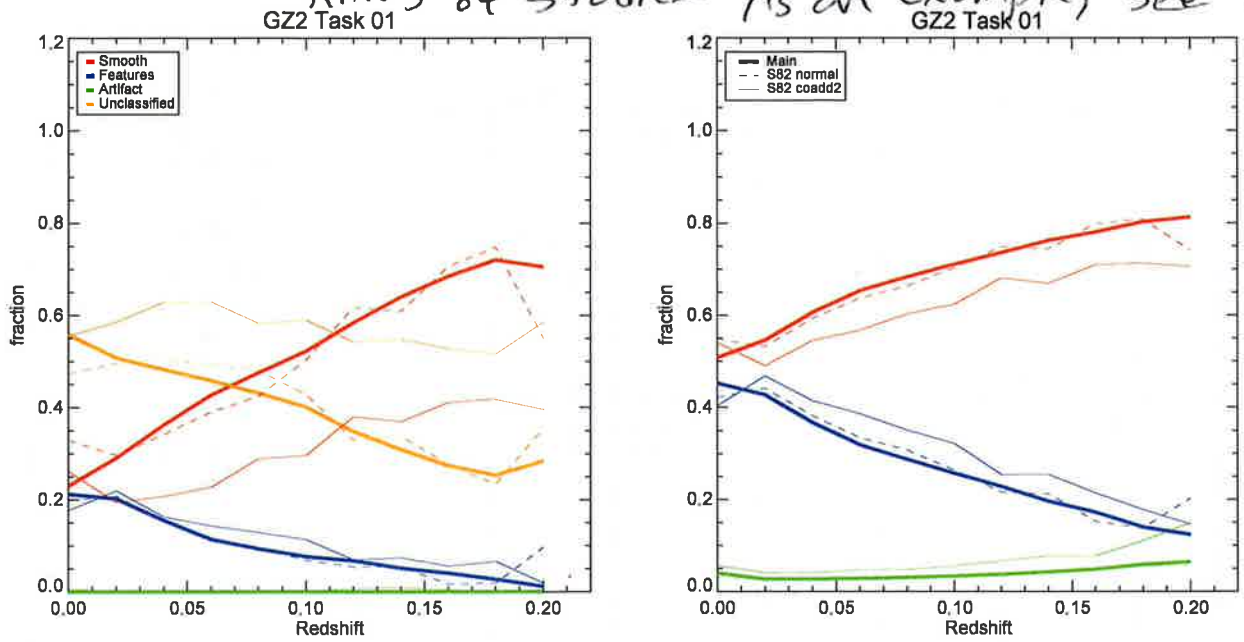


Figure 5. GZ2 weighted vote fractions for Task 01 (*smooth, features/disk, or star/artifact?*) as a function of spectroscopic redshift. The left graph shows the fraction of galaxies for which a category exceeded a threshold of 0.8. Galaxies which had no answer above the threshold are labeled as "unclassified". The right shows the mean of the vote fractions, weighted by the total number of responses to the task for each galaxy. Data are shown for the GZ2 original + extra (thick solid), Stripe 82 normal-depth (thin dotted), and Stripe 82 co-add depth (thin solid) samples. Stripe 82 data is only for galaxies with $r < 17.0$, the same magnitude limit applied to the GZ2 main sample.

which the mean weighted fraction in coadd2 data is ~ 35% higher than from coadd1 data. This is an opposite (but not equal) effect than Answer 12 (obvious bulge), for which the coadd1 data is ~ 13% higher; this may indicate a general shift in votes toward a more prominent bulge. A similar but smaller effect is seen in classification of bulge shapes for edge-on disks (Task 09), where votes for "no bulge" in coadd1 data go to "rounded bulge" in coadd2. The specific cause for these effects as it relates to the image quality is unknown.

The comparison of the coadd1 and coadd2 data sets also demonstrates the intrinsic variability in classification of a single object, even with several tens of votes. For example, in the (unbiased) vote fractions from Task 01, 6831 (32.0%) galaxies from coadd1 and 7,244 (33.9%) galaxies from coadd2 exceed the "clean" early-type threshold of $p \geqslant 0.8$. However, only 2,300 galaxies meet this threshold in *both* samples, while the union of the two yields 11,602 galaxies. The difference in numbers between the samples decreases when a higher value of p is used; a more robust jackknife sampling of the data would improve on this 1-sample jackknife.

Type fractions for both weighted and debiased vote fractions in normal-depth Stripe 82 are shown in Figure 7 for a subset of the GZ2 tasks. The correction flattens the redshift effect in all tasks, similar to the main sample data. The variance along redshift bins is somewhat higher – formal error bars will need to be computed to see if there is any statistical difference, or whether the result is consistent with a smaller total sample of galaxies.

4.3 Additional data

Although not reproduced in this paper, the repository at `http://data.galaxyzoo.org` contains pre-matched tables containing SDSS metadata for the spectroscopic galaxies in the GZ2 main sample. ~~This~~ contains some of the most commonly used DR7 parameters including SDSS exposure information, position, photometry, size, and redshift. Rows are matched to the corresponding galaxies in Tables 3–5. These are provided as a resource for members of the community who wish to compare the morphogical data against external parameters.

5 COMPARISON OF GZ2 TO OTHER CLASSIFICATION METHODS

- Galaxy Zoo 1 (Lintott et al. 2011)
- Nair & Abraham (2010a)
- Huertas-Company et al. (2011)
- EFIGI (Baillard et al. 2011)

5.1 Galaxy Zoo 1 vs. Galaxy Zoo 2

As a check of the classification accuracy, we compare the results from GZ2 to those in GZ1 (Lintott et al. 2011). The galaxies in GZ2 are a subset of those in GZ1, with 248,883 matches between the samples. Task 01 in GZ2 is almost identical to the interface of GZ1. GZ1 allowed for selection of "merger" and "don't know" options in addition to the first

Does this not-understood effect have implications for how users will use the data?

What is the significance of this?

Do we plan to do this computation for this paper? Or for a future paper? Or is this up to the users?

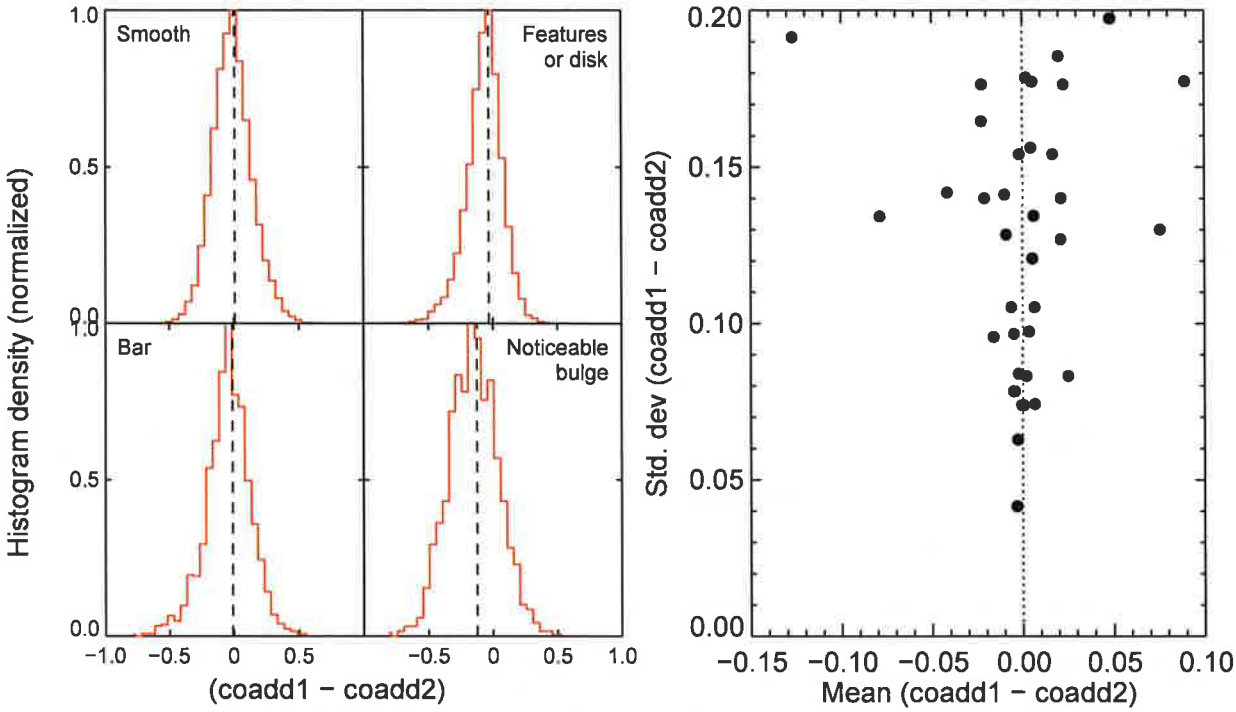


Figure 6. Comparison of GZ classifications from the different coadd techniques for Stripe 82. Left: Distribution of the difference in weighted vote fractions for galaxies that appear in both the coadd1 and coadd2 samples. Each panel shows selected answers in the tree galaxies with at least 10 responses to the task (clockwise from top left: answers 01, 02, 11, and 06). The dashed line shows the median of each distribution; a value of zero means there is no systematic difference, although the widths indicate considerable amounts of scatter for individual classifications. "Noticeable bulge" was the only answer in GZ2 for which the mean $|\Delta_{coadd}| > 0.1$. Right: mean values of the difference in the weighted vote fractions for every response in the GZ2 tree.

three; and asks for galaxies with "features or disk" rather than only for spiral structure.

The matched GZ1-GZ2 catalog contains 34,480 galaxies flagged as "clean" ellipticals based on their debiased GZ1 likelihoods. Of those, 89.0% had GZ2 raw vote fractions greater than 0.8 and 99.9% greater than 0.5. Using the GZ2 debiased likelihoods, the vote fractions match at 50.4% at a threshold of 0.8 and 97.6% at a threshold of 0.5.

There are 83,956 galaxies identified as "clean" spirals in GZ1. The agreement with the "features or disk" response in GZ2, however, is significantly lower. Only 31.6% of the GZ1 clean spirals had GZ2 raw vote fractions greater than 0.8, with 59.2% greater than 0.5. The GZ2 debiased likelihoods for the same galaxies only match at 38.1% (for 0.8) and 78.2% (for 0.5).

Figure 8 shows the difference between the vote fractions for the spiral classifications in GZ1 and features/disk classifications in GZ2 for all galaxies that appear in both catalogs. The weighted vote fractions show a tight correlation at both very low and very high values of f_{sp}, indicating that both projects agree on the strongest spirals (and corresponding ellipticals). At intermediate ($0.2 - -0.8$) values of f_{sp}, however, GZ1 has vote fractions that are consistently higher than those in GZ2, differing by up to 0.25. When using debiased likelihoods in place of the vote fractions, this effect decreases

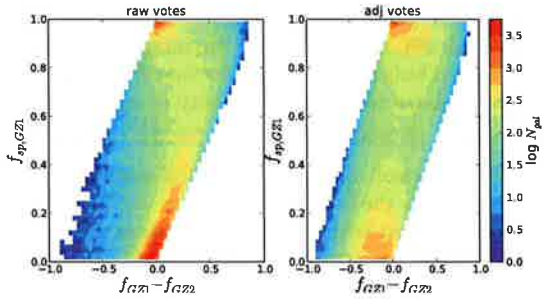


Figure 8. Differences in the vote fractions for galaxies in both the Galaxy Zoo 1 (GZ1) and Galaxy Zoo 2 (GZ2) projects. Left: Distribution of the differences in the raw weighted vote fractions. Dashed lines show data for all galaxies, while solid lines are for the subset in which f_{el} or $f_{sp} > 0.8$ in both samples. Right: same plot, but using the debiased vote fractions for both samples.

dramatically; however, the tightness of the correlation correspondingly drops at low and high f_{sp}.

Based on the vote fractions, GZ2 is significantly more conservative than GZ1 at identifying spiral structure. One possible cause for this is a bias from users who are anticipating subsequent questions about the details of any visible structures. If a user clicks "features or disk" in GZ2 then an experienced classifier may wish to avoid answering those