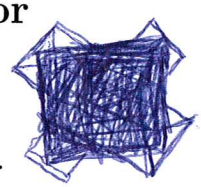
Mon. Not. R. Astron. Soc. **000**, 1–28 (2013) Printed 18 April 2013 (MN LATEX style file v2.2)

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[4,11], 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], Kevin R.V. Casteels[8], M. Jordan Raddick[9], Sugata Kaviraj[2,10], Robert C. Nichol[4,11]

[1] *School of Physics and Astronomy, University of Minnesota, USA*
[2] *Department of Physics, University of Oxford, UK*
[3] *School of Physics and Astronomy, University of Nottingham, UK*
[4] *Institute of Cosmology and Gravitation, University of Portsmouth, UK*
[5] *Institute for Astronomy, ETH, Zürich, Switzerland*
[6] *Center for Astrophysics and Space Sciences, University of California San Diego, USA*
[7] *Astronomy Department, Adler Planetarium and Astronomy Museum, USA*
[8] *Departament d'Astronomia i Meteorologia, Universitat de Barcelona, Spain*
[9] *Department of Physics and Astronomy, Johns Hopkins University, USA*
[10] *Centre for Astrophysics Research, University of Hertfordshire, UK*
[11] *SEPnet, South East Physics Network, UK*

Accepted XXXXXXX

ABSTRACT

Morphology is a powerful probe for quantifying the dynamical history of a galaxy. 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, yet the number of galaxies available in very large samples make visual inspection of each galaxy 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. The GZ2 sample includes all galaxies in the DR7 Legacy survey with $m_r > 17$, alongside galaxies selected from the deeper imaging of SDSS Stripe 82. The original Galaxy Zoo project primarily separated galaxies only into early-types, late-types, and mergers; GZ2 classifies finer morphological features. These features include the presence of bars, bulges, and edge-on disks, as well as quantifying the relative strengths of galactic bulges and spiral arms. This paper presents the full public 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 professional astronomers, especially for T-types, strong bars, and arm curvature. Both the raw and reduced data products can be obtained in electronic format at `http://data.galaxyzoo.org`.

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 (SDSS; 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 at a manageable level of complexity only simple morphological distinctions were initially requested, essentially dividing systems into el-

⋆ E-mail: willett@physics.umn.edu

liptical, spiral and merger. Following the success of the original project, we wanted to determine if the same method could be used for a more complex classification system. This paper presents data and results from Galaxy Zoo's successor, Galaxy Zoo 2 (GZ2), comprising detailed morphologies for more than 300,000 of the largest and brightest SDSS galaxies.[1]

While the morphological distinction used in the original Galaxy Zoo (GZ1) – that which divides spiral and elliptical systems – is the most fundamental, there is a long history of finer-grained classifications. The first systematic approach to classification (Hubble 1936) included a division between barred and unbarred spirals, creating the famous "tuning fork". Further distinctions were based on the shape of early-type systems or tightness of late-type spiral arms. These finer distinctions are often believed to be 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). 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. Detailed classifications of SDSS images by experts have been done by both Fukugita et al. (2007) and Baillard et al. (2011), who determined modified Hubble types for samples of 2253 and 4458 galaxies, respectively. The largest detailed professional classification effort to date was undertaken by Nair & Abraham (2010a), who provide classifications of $\sim$ 14000 galaxies. Galaxy Zoo 2 includes more than an order of magnitude more systems, each with a large number of independent inspections. The size of this sample allows for a more complete study of small-scale morphological features and better statistics for rarer classes of objects, while multiple classifications yields an estimate of the associated uncertainty.

The use of proxies for morphology such as colour, concentration index, spectral features, surface brightness profile, structural features, spectral energy distribution or some combination of these is not an adequate substitute. Each proxy has an unknown and possibly biased relation with the morphological features under study. With a sufficiently large set of classified galaxies, however, we can fully sample the morphological diversity of the local population and quantify the relationship between morphology and the proxies discussed above.

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), 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 were made possible by the participation of hundreds of thousands of volunteer 'citizen scientists'. 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). Since then, this method has been expanded beyond simple shape classifications to supernova identification (Smith et al. 2011), exoplanet discovery (Fischer et al. 2012; Schwamb et al. 2012) and a census of bubbles associated with star formation in the Milky Way (Simpson et al. 2012), amongst many others.

Several results based on early Galaxy Zoo 2 data have already been published. Masters et al. (2011, 2012) use galaxy bar classifications to show a clear increase in bar fraction for galaxies with redder colours, lower gas fractions, and more prominent bulges. Hoyle et al. (2011) developed a separate interface to measure bar properties, showing that the bars themselves are both redder and longer in redder disk galaxies. Skibba et al. (2012) demonstrated that a significant correlation exists between barred and bulge-dominated galaxies at separations from 0.15–3 Mpc. Kaviraj et al. (2012) used GZ2 to study early-type galaxies with visible dust lanes, while Simmons et al. (2013) discovered a population of AGN host galaxies with no bulge, illustrating how black holes can grow and accrete via secular processes. Finally, Casteels et al. (2013) quantify morphological signatures of interaction (including mergers, spiral arms, and bars) for galaxy pairs in the SDSS.

This paper is organised as follows. Section 2 describes the sample selection and method for collecting morphological classifications. Section 3 outlines the data reduction process, and Section 4 describes the tables that comprise the public data release. Section 5 is a detailed comparison of GZ2 to four additional morphological catalogues that were created with SDSS imaging. Section 6 presents morphologically-sorted colour-magnitude diagrams as an example of the science that can be done with GZ2. We summarise our results in Section 7.

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. 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. Spectroscopic targets came from the SDSS Main Galaxy Sample (Strauss et al. 2002).

Several cuts were applied to the DR7 Legacy sample for selection in GZ2. The goal for these cuts was to include the nearest, brightest, and largest systems for which fine morphological features could be resolved and classified. We required a Petrosian half-light magnitude brighter than 17.0

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

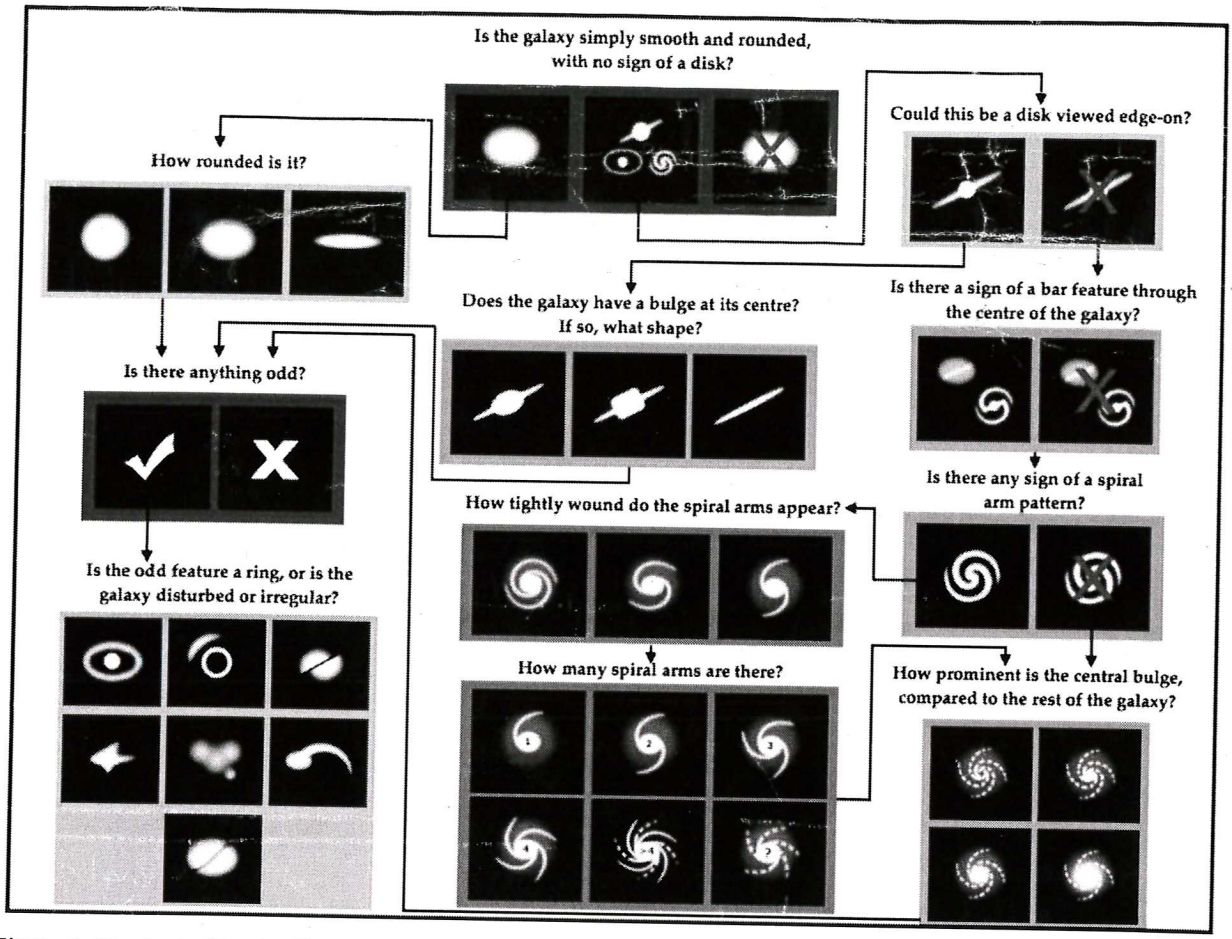


Figure 1. Flowchart of the classification tasks for GZ2, beginning at the top centre. Tasks are colour-coded by their relative depths in the decision tree. Tasks outlined in brown are asked of every galaxy. Tasks outlined in green, blue, and purple are (respectively) one, two or three steps below branching points in the decision tree. Table 2 gives a description of the responses that correspond to the icons shown here.

total of 37 possible responses (Figure 1 and Table 2). A classifier selects only one response for each task, after which they are immediately taken to the next task in the tree. Tasks 01 and 06 are the only questions that are always completed for a given object. Once a classification is complete, an image of the next galaxy is automatically displayed and the user can begin classification of a new object.

Data from the classifications were stored in a live Structured Query Language (SQL) database. In addition to the morphology classifications, the database also recorded a timestamp, user identifier, and image identifier for each classification; volunteers were required to log-in in order for their classifications to be recorded. Repetition

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, images shown to classifiers were selected from the database in a random order. We wanted to ensure, however, that each galaxy ultimately had enough classifications to accurately measure its uncertainty. 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 "stripe82_coadd_1" sample was removed from the site at this time. The main sample galaxies finished with a median of 44 classifications; the minimum was 16 classifications, and > 99.9% of the sample had at least 28. The "stripe82_coadd_2" galaxies had a median of 21 classifications and > 99.9% had at least 10 (Figure 2).

The last GZ2 classifications were collected on 2010-04-29, with the project spanning just over 14 months. The archived site continued to be maintained, but classifications were no longer recorded. The final dataset contained 16,340,298 classifications (comprising a total of 58,719,719 tasks) by 83,943 volunteers.

Sample	N_{gal}	N_{class} median	m_r [mag]
original	245,609	44	17.0
extra	28,174	41	17.0
Stripe 82 normal	21,522	45	17.77
Stripe 82 normal ($m_r < 17$)	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 ($m_r < 17$)	283,971	44	17.0

Table 1. Basic properties of the galaxy samples in GZ2, including the total number of galaxies (N_{gal}), the median number of classifications per galaxy (N_{class}), and the apparent magnitude limit.

in the r-band (after Galactic extinction correction was applied), along with a `petroR90_r`, the radius containing 90% of the r-band Petrosian aperture flux, 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. 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 which had been repeatedly imaged during the SDSS survey. The selection criteria are the same as that for the Legacy galaxies, with the exception of a fainter magnitude limit of $m_r < 17.77$. 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 $m_r \leqslant 17.0$. We verified that there are no significant differences in classifications between these samples that could be 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 $m_r > 17.0$ 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.

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. Maximising the visibility of faint features, however, resulted in more prominent background sky noise; since each pixel is typically dominated by a single band, the background is often brightly coloured by the Lupton et al. (2004) algorithm. Due to concerns that this would make it obvious that images were from deeper data and potentially affect morphological classifications, we created a second set of coadd images in which the colour of background pixels was removed. This was achieved by reducing the colour saturation of pixels outside of a "soft-edged" object mask.

The original and desaturated coadd image sets are labeled "stripe82_coadd_1" and "stripe82_coadd_2", respectively (Table 1). Analysis of the coadded images only slight differences between the two sets of classifications (see Section 4.2).

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 classifications 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. Users had the option to invert the default colour scaling on any image being classified.

Morphological classification of the galaxies proceeds via a multi-step decision tree. We define a *classification* as the total amount of information collected by completing the decision tree. Each individual step in the tree is a *task*, which consists of a *question* and a finite set of possible *responses*. The selection of a particular response is referred to as the user's *vote*.

Classification begins with a slightly modified version of the GZ1 task, with users identifying whether the galaxy is either "smooth", has "features or a disk", or is a "star or artifact". The exact order of any subsequent tasks depends 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 task would not be shown if they had selected either of the other two options.

The Galaxy Zoo 2 tree has 11 classification tasks with a

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

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

If we write the unbiased likelihood for a single task as:

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

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

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

then dividing (6) by (7) yields:

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

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

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

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

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

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

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

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

$$C_{j,i} = log_{10}\left(\frac{f_i}{f_j}\right)_{z=0} - \log_{10}\left(\frac{f_i}{f_j}\right)_{z=z'}. \quad (14)$$

So the correction $C_{j,i}$ for any bin is simply the difference between f_i/f_j at the desired redshift and 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 + apparent 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 data only from galaxies with sufficient numbers of responses to determine their morphology; this threshold is different for each task (Table 3).

The 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 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. These bin ranges and step sizes are chosen to maximize the phase space covered by the bias correction. Only bins with at least 20 galaxies are considered. The value of each bin in the cube is the sum of the 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 $\binom{n}{2}$ correction terms 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 5 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. This ratio is reasonably well-fit with an analytic function taken from Bamford et al. (2009):

$$\frac{f_j}{f_i}[R_{50}, M_r] = \frac{s_6}{1 + exp[(\alpha - M_r)/\beta]} + s_7 \quad (15)$$

where:

$$\alpha = s_2 \times \exp[-(s_1 + s_8 {R_{50}}^{s_9})] + s_3, \quad (16)$$

$$\beta = s_4 + s_5(x_0 - s_3), \quad (17)$$

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 that had baseline ratios reasonably well fit by an expression of this form was Task 07 (the roundedness of 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 15; for these, we instead adopt a simpler fit where both M_r and R_{50} vary linearly:

$$\frac{f_j}{f_i}[R_{50}, M_r] = t_1(R_{50} - t_2) + t_3(M_r - t_4) + t_5, \quad (18)$$

and $\{t_1, t_2, t_3, t_4, t_5\}$ are the parameters to be minimized. We fit Equation 18 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 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 4 illustrates the comparison between the mean raw and debiased vote fractions as a function of redshift. The debiased results (*thick lines*) are generally flat over $0.01 < z < 0.085$, where $L^\star$ galaxies ($M_r \sim -20.44$; Blanton et al. 2003) are within the detection 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) for the same selection criteria. The bar fraction in disk galaxies is roughly 0.35, slightly higher than the value found by using thresholded GZ2 data in Masters et al. (2011).

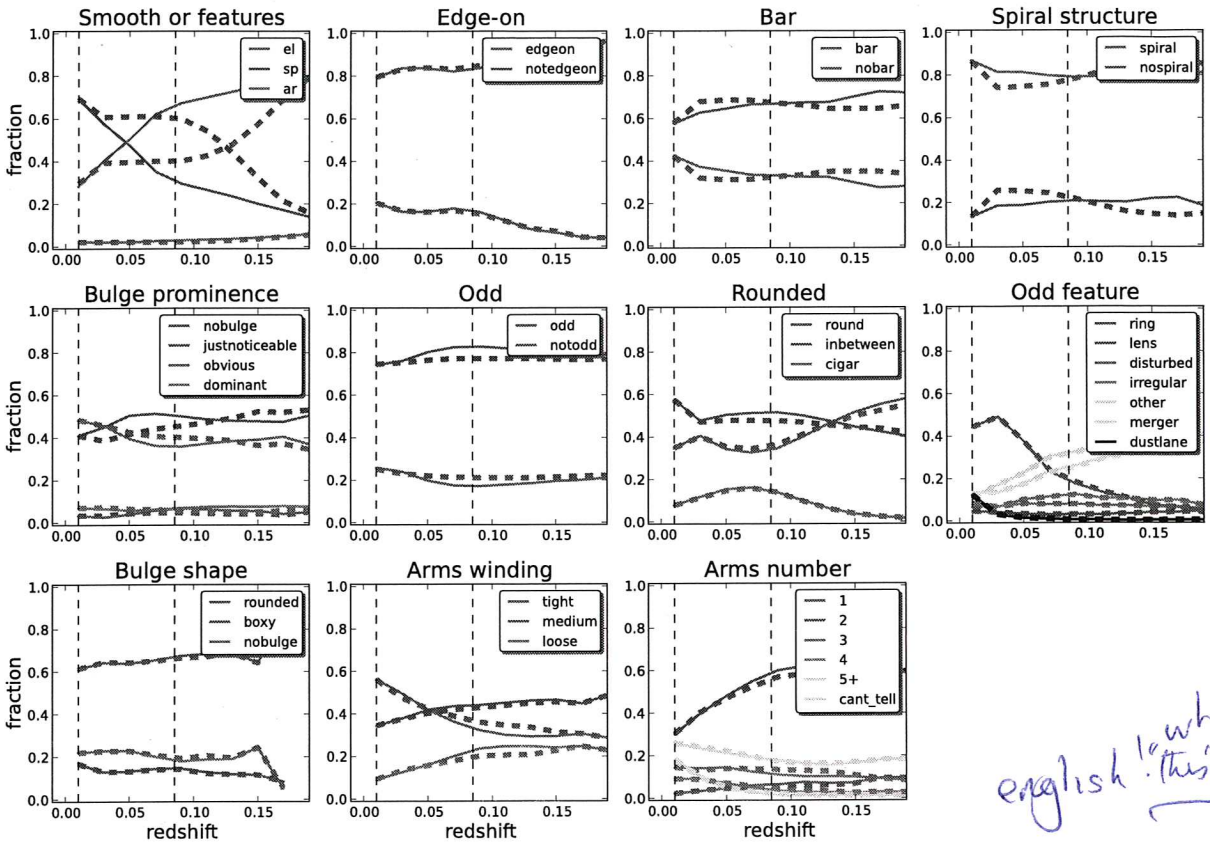


Figure 4. Type fractions as a function of redshift for the classification tasks in GZ2. Solid (thin) lines show the vote fractions, while the thick (dashed) lines show the debiased vote fractions adjusted for classification bias. This is a luminosity-limited sample for $M_r < -20.89$. The data for each task is plotted only for galaxies with enough votes to characterize the response distribution (Table 3). 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).

Task	Previous tasks	Vote fraction $n_t \geqslant 10$	Vote fraction $n_t \geqslant 20$
01	–	–	–
02	01	0.227	0.430
03	01,02	0.519	0.715
04	01,02	0.519	0.715
05	01,02	0.519	0.715
06	–	–	–
07	01	0.263	0.469
08	06	0.223	0.420
09	01,02	0.326	0.602
10	01,02,04	0.402	0.619
11	01,02,04	0.402	0.619

Table 3. Thresholds for determining well-sampled galaxies in GZ2

ual morphologies. The correction derived from well-classified galaxies is then applied to the vote fractions for *all* galaxies in the sample.

The second major difference is that the adjustment of the GZ1 vote fractions assumed that the single task was essentially binary. Since almost every vote in GZ1 was for a response of either "elliptical" or "spiral" (either anticlockwise, clockwise, or edge-on), 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 other options do not significantly affect the classification bias for the most popular responses. This is not possible for GZ2, as many tasks have more than two possible responses and represent a continuum of relative feature strength, rather than a binary choice.

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 mix of morphologies. We quantify this using the ratio of vote fractions (f_i/f_j) for responses i and j. 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:

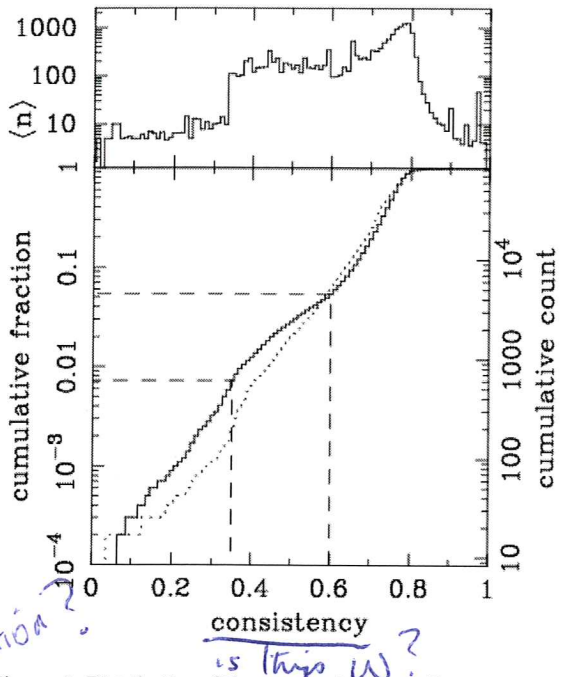


Figure 3. Distribution of the user consistency κ. Top: mean number of galaxies classified per user as a function of their consistency. Bottom: Cumulative fraction distribution of consistency. The dotted line shows the first iteration of weighting, and the solid line the third iteration. The second iteration is not shown, but is almost identical to the third. Dashed lines indicate where the user weighting function takes values of 0.01 and 1.

results for the initial iteration (Figure 3), we chose a weighting function that down-weighted users in the tail of low consistency:

$$w = \min\left(1.0, (\bar{\kappa}/0.6)^{8.5}\right). \quad (4)$$

For this function, $w = 1$ for $\sim 95\%$ of users and $w < 0.01$ for only $\sim 1\%$ 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 3 also shows that the lowest-weighted users on average classified only a handful (< 10) of objects. This effect demonstrates either learning during classification, or the systematic loss of inconsistent users during their career as classifiers; further work on user behaviour is needed to distinguish between the two possibilities.

After computing κ, 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. These are used exclusively hereafter, and for brevity we typically drop the term "weighted".

3.3 Classification bias

We also adjust the vote fractions for what we term *classification bias*. The overall effect of this bias is a change in observed morphology fractions as a function of redshift *independent of any true evolution in galaxy properties*, a trend also seen in the Galaxy Zoo 1 data (Bamford et al. 2009). The SDSS survey is expected to be shallow enough to justify an assumption of no evolution, and so the presumed cause is that more distant galaxies, on average, are both smaller and dimmer in the cutout images; as a result, finer morphological features are more difficult to identify. This effect is not limited to crowd-sourced classifications; expert classifications must also suffer from bias to some degree, although their more limited statistics make it difficult to quantify.

Figure 4 demonstrates the classification bias for several of the Galaxy Zoo 2 classification tasks. The average vote fraction for each response 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) decreases 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 the observed trends in type fractions at high redshifts is due to the nature of a magnitude-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 clear evidence of the classification bias even in luminosity-limited samples (between the dashed vertical lines in Figure 4). Since this bias contaminates any potential studies of galaxy demographics over the sample volume, it must be corrected to the fullest possible extent.

Bamford et al. (2009) corrected for classification bias in the GZ1 data, but only for the elliptical and combined spiral variables. Their approach was to bin the galaxies as a function of absolute magnitude (M_r), the physical Petrosian half-light radius (R_{50}), and redshift. They then computed the average elliptical-to-spiral ratio for each (M_r,R_{50}) bin in the lowest redshift slice with significant numbers of galaxies; 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 derived a correction for each (M_r,R_{50},z) bin and then adjusted the vote fractions for the individual galaxies in each bin. The validity of this approach is justified in part since debiased vote fractions result in a consistent morphology-density relation over a range of redshifts (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 response for each vote. This means that all tasks, with the exception of the first, depend on responses to previous tasks in the decision 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 vote fraction for this example task only addresses the total bar fraction *among galaxies that a user has classified as disks and are not edge-on*, and not as a function of the general population.

For a galaxy to be used in deriving a correction, we therefore require both a minimum weighted vote fraction for the preceding response(s) and a minimum number of votes for the task in question (Table 3). While this threshold increases the number of bins with large variances, it is critical for reproducing accurate baseline measurements of individ-

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

Table 2. The GZ2 decision tree, comprising 11 tasks and 37 responses. The 'Task' number is an abbreviation only and does *not* necessarily represent the order of the task within the decision tree. The texts in 'Question' and 'Responses' are displayed to volunteers during classification, along with the icons in Figure 1. 'Next' gives the subsequent task for the chosen response.

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 vote 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. Repeat classifications occurred for only $\sim 1\%$ of all galaxies. The removal of the repeats only altered the final vote fractions

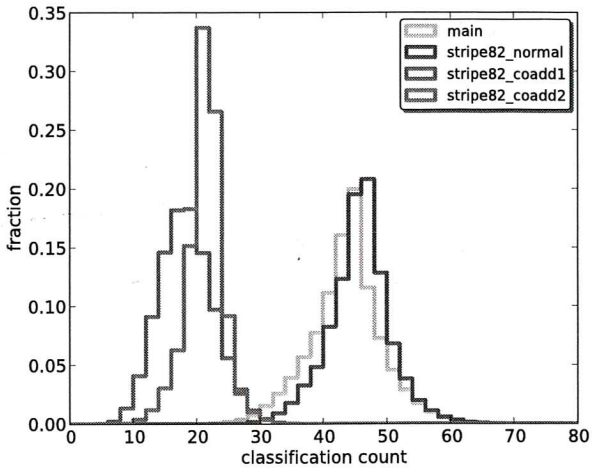


Figure 2. Distribution of the number of classifications for the sub-samples within GZ2.

(thus changing the morphological classification) for $\lesssim 0.01\%$ of the sample.

3.2 Consistency and invididual user weighting

The next step in reducing the data is to reduce the influence of unreliable classifiers. To do so we applied an iterative weighting scheme, similar to that used for GZ1, but adjusted to account for questions for which more than two answers are possible. First, we calculated the vote fraction ($f_r = n_r/n_t$) for every response to every task for every galaxy, weighting each user's vote equally. Here, n_r is the number of votes for a given response and n_t is the total number of votes for that task. Each vote is compared to the vote fraction to calculate a user's consistency κ:

$$\kappa = \frac{1}{N_r} \sum_i \kappa_i, \tag{2}$$

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

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

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

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

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

Each user was assigned a consistency ($\bar{\kappa}$) by taking the mean consistency of every response. From the distribution of

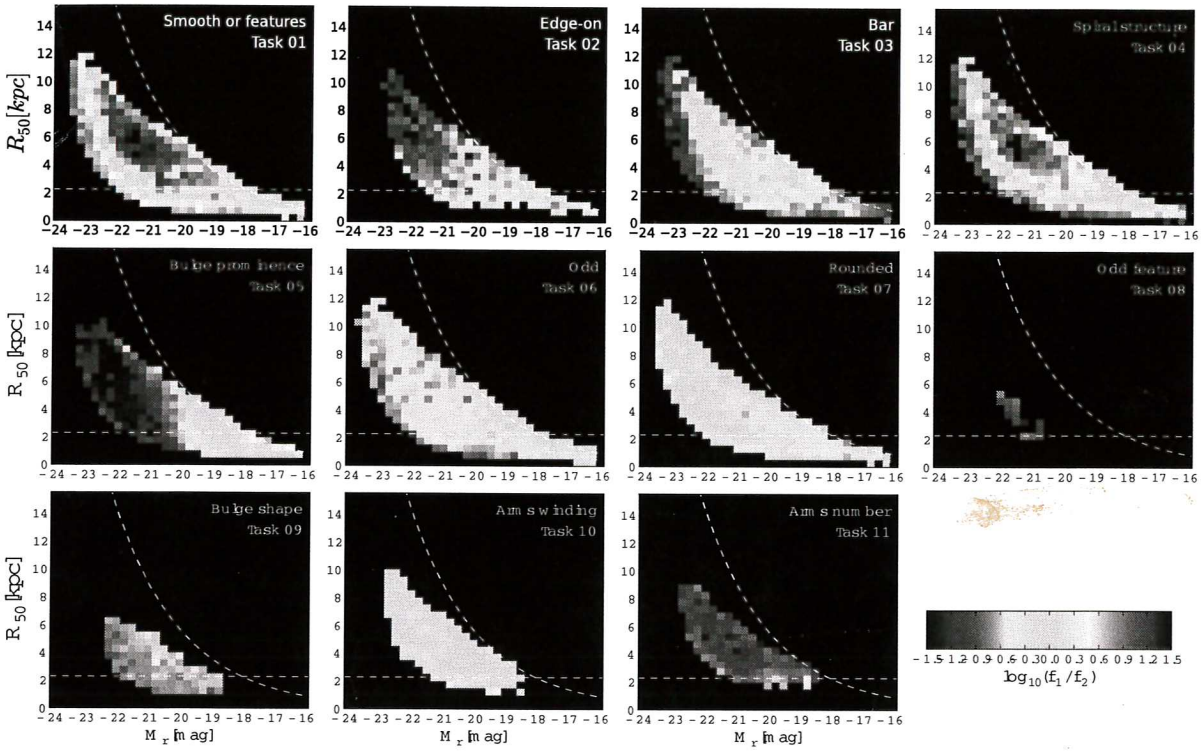


Figure 5. Local morphology ratios for GZ2 classifications; these are used to derive the corrections that adjust data for classification bias (§3.3). The ratio of the binned vote fractions is for the first two responses in the decision tree (Table 2) for each task; there may be as many as 21 such pairs per task, depending on the number of unique responses. 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}.

4 THE CATALOGUE

Other possible inclusions for catalogue:

- Metrics on classification confidence (Table 04, Lintott et al. 2011) *(not generated yet)*

4.1 Main sample

The data release for Galaxy Zoo 2 consists of four tables, abridged portions of which appear in this paper. Table 4 contains classification data for the 243,500 galaxies in the main sample with spectroscopic redshifts. 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 classified the galaxy, while N_{votes} gives the total number of votes 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 vote 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 (as described below).

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 galaxies from which a clean barred sample can be identified, 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 due to 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. Note that 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 5 shows the GZ2 classifications for the 42,462 main sample galaxies without spectroscopic redshifts. To compute the debiased likelihoods, we used the morphology corrections obtained for galaxies in the spectroscopic main sample. We then used the photometric redshift provided by the SDSS (Csabai et al. 2003) to derive M_r and R_{50} and select the appropriate correction bin. The mean error in the redshift of the photometric sample (from the SDSS photo-z) 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 3 bins can potentially produce a very large change in the debiased vote fractions.

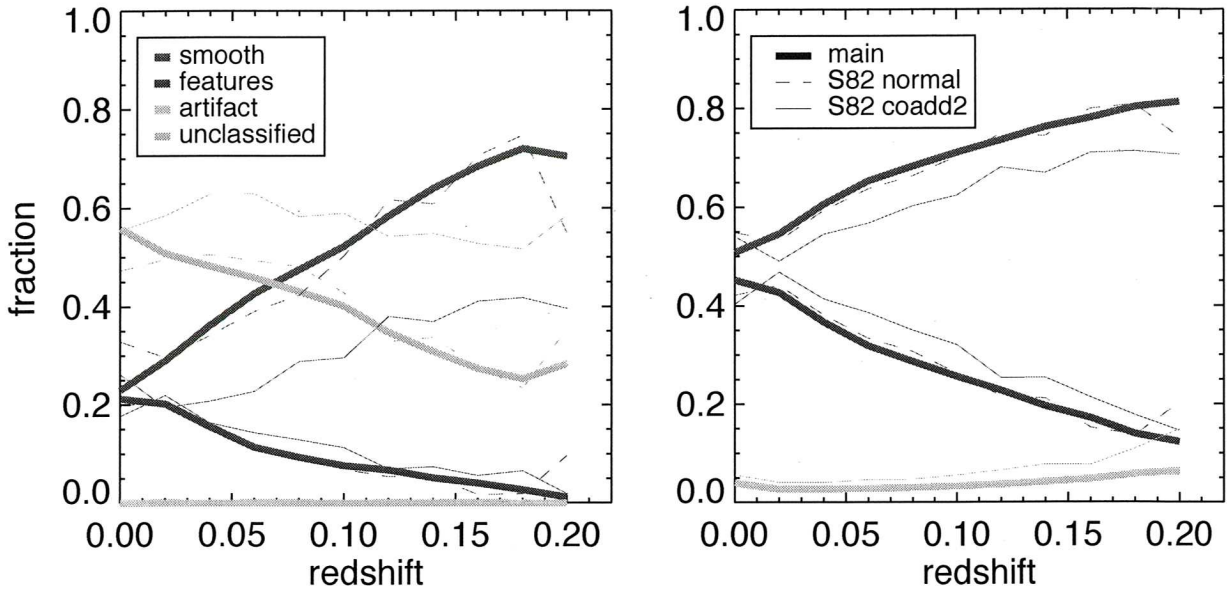


Figure 6. GZ2 vote fractions for Task 01 (*smooth, features/disk, or star/artifact?*) as a function of spectroscopic redshift. The left graph shows the fraction of galaxies for which the vote fraction exceeded 0.8. Galaxies which did not have a response above the threshold are labeled as "unclassified". The right shows the mean of the vote fractions, 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 $m_r < 17.0$, the same magnitude limit applied to the GZ2 main sample.

Since the redshift can have a strong effect on classification bias, we separate galaxies with spectroscopic and photometric redshifts, and do not recommend that the debiased data be combined for analysis. For use cases where the main driver is the number of galaxies, however, it may be possible to combine the raw vote fractions to create the largest possible sample.

4.2 Stripe 82

The distribution of votes for the main sample is similar enough to that of the Stripe 82 normal-depth imaging (with $m_r < 17.0$) that the same bias correction applies for both. The distributions of vote fractions 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 this similarity 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.

The right panel of Figure 6 shows that the 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 type fractions as a function of redshift for both the Stripe 82 normal depth and the rest of the GZ2 main sample are very similar; this is not the case, however, for the coadded Stripe 82 data. For Task 01, fewer galaxies are classified as robustly smooth (above the 0.8 threshold), moving instead to the "unclassified" category. Coadded data also showed 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 (from 1.4″ to 1.1″; Annis et al. 2011) and higher signal-to-noise in the coadded images.

For almost every response in the GZ2 decision tree, the data (no bias correction) show no systematic differences between classifications using the coadd1 and coadd2 images. Figure 7 shows the difference between the two vote fractions ($\Delta_{coadd} = f_{coadd1} - f_{coadd2}$). If the mean value of Δ_{coadd} for a response 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}. For most purposes, therefore, the two versions of images were not distinguished and their classifications can be combined.

The biggest systematic difference is for the response to Task 05 (bulge prominence) of the bulge being "just noticeable". The mean fraction in coadd2 data is $\sim 35\%$ higher than that in coadd1 data. This effect is opposite (but not equal) to that for an "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 not investigated further in this paper.

The comparison of the coadd1 and coadd2 data sets (which only differ in their treatment of background noise) 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%)

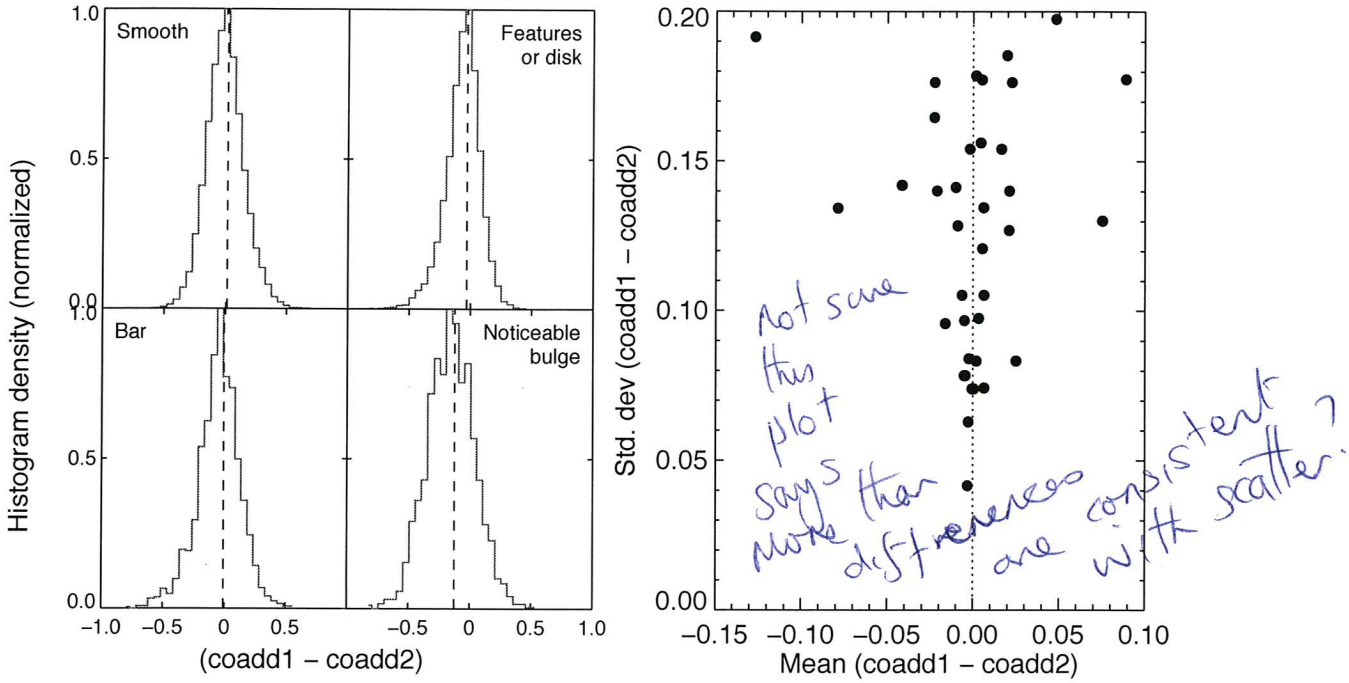


Figure 7. Comparison of GZ2 classifications from the two sets of coadded images for Stripe 82. Left: Distribution of the difference in vote fractions for galaxies that appear in both the coadd1 and coadd2 samples. Each panel shows selected responses for galaxies with at least 10 responses to the task. 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 response in GZ2 for which the mean $|\Delta_{coadd}| > 0.1$. Right: Mean values of the difference in the vote fractions for every response in the GZ2 tree.

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.

Three tables of GZ2 morphological data are presented for the Stripe 82 galaxies. Table 6 gives classifications for Stripe 82 normal-depth images with spectroscopic redshifts. Galaxies in this table with $m_r < 17.0$ also appear in Table 4; however, the corrections for classification bias here are derived based only on Stripe 82 data, and so debiased likelihoods and flags may be slightly different.

Tables 7 and 8 contains classification data for the Stripe 82 galaxies with co-added images and spectroscopic redshifts. Debiased probabilities and flags are derived from separately from each coadded data set. Since both the number of galaxies and the average number of classifications per galaxy are a small fraction of that in the main sample, though, the corrections encompass a smaller range of tasks and phase space in (M_r,R_{50},z). The increased exposure time and improved seeing, however, means that the effect of classification bias is lessened at lower redshifts; the raw vote fractions may thus be more suitable for some science cases using the coadded images.

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. These tables contains some of the most commonly used DR7 parameters including SDSS exposure information, position, photometry, size, and redshift. These are provided as a resource for members of the community who wish to compare the morphogical data against external parameters.

Tables 4–8 can also be accessed from CasJobs in the SDSS Data Release 10, expected to be available in Jul 2013.

4.4 Using the classification data

Add text providing 1–2 examples of how the GZ2 catalogue can be used. Example: selecting clean samples of barred spirals.

5 COMPARISON OF GZ2 TO OTHER CLASSIFICATION METHODS

We compare the classifications for GZ2 to four other morphological catalogues, all of which are for SDSS images and overlap (at least in part) with the GZ2 sample:

- 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 compared 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 broadly similar to the interface of GZ1, with some modifications. GZ1 had six possible responses for its task: "elliptical", "clockwise spiral", "anticlockwise spiral", "other spiral", "merger" and "star/don't know". We compared data for the GZ1 "elliptical" to GZ2 "smooth", and combined responses for the three GZ1 spiral categories to the GZ2 "features or disk".

The matched GZ1-GZ2 catalogue contains 34,480 galaxies flagged as clean ellipticals based on their debiased GZ1 likelihoods. Of those, 89.0% had GZ2 raw vote fractions for "smooth" greater than 0.8 and 99.9% greater than 0.5. Using the GZ2 debiased likelihoods, 50.4% of galaxies have vote fractions exceeding 0.8 in both samples, while 97.6% have vote fractions exceeding 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 catalogues. The 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 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. An experienced classifier, for example, would know that selecting "features or disk" is followed by additional questions, none of offer options for an uncertain classification. If the classifier is less confident in identifying a feature, it is possible they would avoid this by clicking "smooth" instead. This is a hypothesis; there is no direct evidence from the data suggesting that this has taken place, but suggest it as one possibility for explaining the discrepancy in otherwise similar classification tasks.

Since the GZ1 vote fractions were specifically directed to galaxies with spiral arms, we also compared GZ1 to the results of Task 04 in GZ2, which specifically asks for spiral structure in disk galaxies that are not edge-on. The agreement with the GZ1 clean spirals is higher than for Task 01, but still well short of that for smooth/elliptical galaxies.

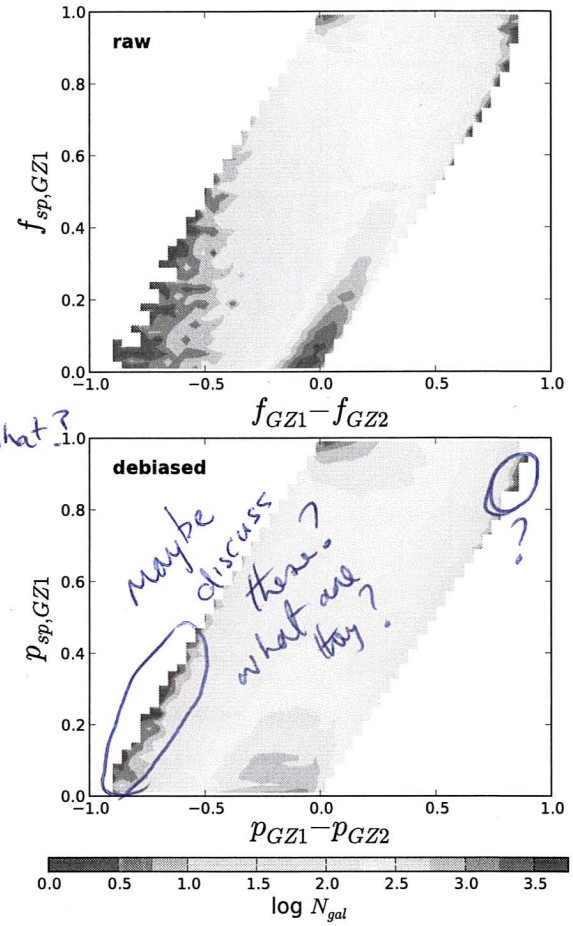


Figure 8. Comparison of spiral galaxies using classifications for "combined spiral" (GZ1) and "features or disk" (GZ2). Left: raw vote fractions. At intermediate values ($f_{sp} \sim 0.5$), GZ1 users are more likely to classify galaxies as spiral compared to GZ2. Right: debiased vote fractions. At intermediate values, GZ1 and GZ2 classifications are consistent with each other; however, there is an increased scatter in the vote fractions near $f_{sp} \simeq 0$ and $f_{sp} \simeq 1$.

Only 63.6% of galaxies have Task 04 raw vote fractions greater than 0.5 for the GZ1 clean spirals, with 66.8% for the debiased vote fractions.

Results from comparing GZ1 to the spiral structure task in GZ2 indicate the robustness of the GZ2 results. If spiral features are identified in GZ2 (having already selected for disk galaxies), then they are very likely to be similarly classified in GZ1. Conversely, if GZ1 classifications indicate a possible (but not definitive) spiral, it is less likely to appear in GZ2 Task 04, based on the stricter requirements for inclusion. The scatter in the debiased likelihoods may thus be a fair representation of the uncertainty in individual classifications.

Figure 9 shows the distribution of the difference between the vote fractions for the two projects, using the elliptical and combined spiral data for GZ1 and the Task 01 smooth and "features or disk" data for GZ2. For the raw vote fractions, galaxies showed a significant skew toward be-

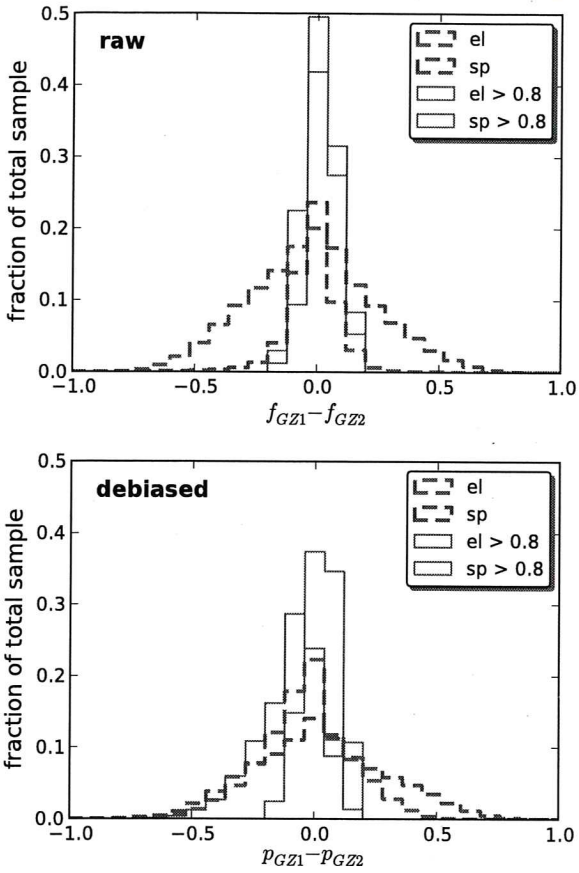


Figure 9. Differences in the vote fractions for galaxies in both the Galaxy Zoo 1 (GZ1) and Galaxy Zoo 2 (GZ2) projects. Left: Distribution of the differences in the vote fractions. Dashed lines show data for all galaxies, while solid lines are for the subset in which f_{el} or $f_{sp} > 0.8$ in both samples. Right: same plot, but using the debiased vote fractions for both samples.

ing more likely to be identified as a spiral in GZ1 than in GZ2. When restricted only to galaxies in the joint CLEAN samples ($p > 0.8$), the spread is greatly reduced and the distribution is centred around a difference of zero. The debiased vote fractions show a similar spread when comparing GZ1 and GZ2 classifications, although the skew toward spirals in GZ1 is largely removed. When using only clean galaxies and the debiased vote fractions, galaxies are more likely to be identified as spirals in GZ2.

The GZ1 interface did have one option that did not classify either early- or late-type galaxies, but rather mergers. This was a rare response in the GZ1 data, comprising less than a percent of the total type fraction at all redshifts (Bamford et al. 2009). Darg et al. (2010) found that a vote fraction of $f_{mg} > 0.6$ robustly identified merging systems in GZ1. Of the 1632 systems meeting that criteria and also classified in GZ2, more than 99% were identified as "odd" galaxies, and 77.7% had a merger fraction above 0.5 as a response to Task 08. This is partly due to early-stage merging spirals avoiding the "merger" classification, with only late-state mergers with extremely disturbed morphologies recording high vote fractions for the merger question. This agrees with the analysis of Casteels et al. (2013), who found that the merger vote fraction for close pairs in GZ2 increases strongly with decreasing projected separations.

Section on angular separation bias/crosstalk between odd questions, possibly by KRVC.

5.2 Nair & Abraham

Nair & Abraham (2010a, hereafter NA10) published a catalogue with expert morphological classifications of 14,034 galaxies from the SDSS DR4. Galaxies were selected from a redshift range of $0.01 < z < 0.1$, with an extinction-corrected apparent magnitude limit of $g < 16$. The GZ2 sample is deeper ($m_r < 17$), spans a wider redshift range ($0.0005 < z < 0.25$), and contains a more recent data release (DR7) than galaxies in NA10. 12,480 galaxies have been classified in both GZ2 and NA10; this comprises nearly all (89.9%) of the NA10 catalogue, but only 4.5% of GZ2. The overlap between the samples allows for a direct comparison of the two classification methods and schema.

Nair & Abraham (2010a) used classifications by a single astronomer (P. Nair) to quantify the galactic morphology. They determined RC3 T-types (a numerical index of a galaxy's stage along the Hubble sequence; de Vaucouleurs et al. 1991) for the entire sample through visual inspection of monochrome g-band images, covering each source twice. There is no discussion on their procedure if the perceived T-type changed between the first and second classification of an image.

In addition to the T-types, NA10 also classified various "fine structure" morphological features in each galaxy. These include:

- bars (strong, weak, intermediate, ansae, "peanut", nuclear, and/or unsure)
- rings (nuclear, inner, outer)
- lenses [regions of constant surface brightness; not gravitational lenses] (inner, outer)
- pairs of objects (close, projected, adjacent, overlapping, + flags for second object type)
- interaction (none, disturbed, warp, shells, short tail, medium tail, long tail, bridge)
- tails (number)

All references to the GZ2 vote fraction in the following sections refer to data which has been weighted for consistency, but not debiased.

5.2.1 T-types

There has been no published discussion in the literature comparing large-scale morphologies of the NA10 and Galaxy Zoo catalogues. Nair & Abraham (2010a) was published after the first GZ1 results (Lintott et al. 2008), but prior to the formal data release paper (Lintott et al. 2011). Huertas-Company et al. (2011) do compare automated classifications to both NA10 and GZ1, finding good agreement with both; this obliquely suggests that the GZ2 and NA10 classifications will also be consistent.

The left panel of Figure 10 shows the percentage of

galaxies identified as having either a disk or features from the first question in the GZ2 tree, colour-coded by their NA10 T-types. There is a clear separation in the GZ2 fractions for galaxies classified as E vs. those with T-types Sa and later. Disk galaxies, including S0's, have a median fraction of the "features or disk question" of 0.796, with a standard deviation of 0.29. This distribution is bimodal, with one peak near 0.95 and a second at 0.1. Breaking down the disk galaxies into more specific Hubble classifications, the disk galaxies with low GZ2 feature votes are found to be primarily lenticular (S0; T-type = −3 to 0) galaxies. If only galaxies with T-types Sa or later are considered, the peak at lower GZ2 vote fractions disappears. The median GZ2 vote fraction for these galaxies is 0.88, with a standard deviation of 0.23. The highest GZ2 vote fraction for an elliptical galaxy in NA10 is 0.741; therefore, any cut above this limit includes *exclusively* identified by NA10 as late-type. Even if the confidence of this decreases for the larger GZ2 sample due to the inclusion of fainter galaxies, the previous limit of 0.8 (which may be conservative) reproduces the broad morphological cuts of NA10 extremely well.

Since there are very few objects identified as stars or artifacts in the first GZ2 question, the vote fraction for smooth galaxies is approximately $f_{smooth} = (1 - f_{features/disk})$. Elliptical galaxies (T-type = −5) have a median vote fraction of the "smooth" question of 0.86, with a standard deviation of 0.07. The GZ2 votes for the NA10 ellipticals are much more sharply peaked than the late-type galaxies, lacking the long tail seen even for very late types. This means that a cut on GZ2 votes for smooth galaxies at 0.8, for example, would include 4% late-type galaxies (20% if S0 galaxies are included).

For galaxies identified as having features that are not edge-on disks, GZ2 users then vote on whether the galaxy has visible spiral structure (Task 04). For the few NA10 elliptical galaxies that have votes for this question, ∼85% of them have GZ2 vote fractions of zero, with the remainder weakly clustered around 0.3. For NA10 late-type galaxies, the majority of disk/feature objects have high GZ2 spiral structure vote fractions. For galaxies with at least 10 votes on Task 04, 70% of Sa or later-types have a GZ2 spiral vote fraction > 0.8. This drops to 60% if S0 galaxies are included as late-type. The missing population is thus made up of galaxies that NA10 classify as having significant spiral structure, but for which GZ2 users cannot distinguish spiral arms. One might expect these galaxies to have lower magnitudes or surface brightnesses compared to the rest of the sample, thus lowering the confidence of GZ2 votes (there is no analog parameter associated with NA10 classifications). However, the apparent g and r magnitudes, as well as the absolute g-band magnitude, show no difference between galaxies above and below the 80% cutoff. Changing the value for the GZ2 vote fraction did not affect the results, so it appears that lower GZ2 vote fractions for spirals indicate intrinsically weaker (or less clearly-defined) spiral arms.

5.2.2 *Bars*

NA10 detect 2537 barred galaxies, 18% of their total. For objects with T-types later than E/S0, this rises to 25% of the sample. This is consistent with the bar fraction from (Masters et al. 2011) for disky, not edge-on galaxies from early GZ2 data (29%).

Two parameters can be set that reduce the number of galaxies in the overlap between the samples, but which result in a cleaner cut for comparisons. The first uses the Masters et al. (2011) cut for galaxies that are not edge-on (log $(a/b) < 0.3$ using the EXPAB_R parameter from SDSS). The second is to only look at galaxies with at least 10 classifications for Task 03 (*bar present?*) in GZ2, a total of 7,121 galaxies from the original 12,480. All trends described below hold generally for both the full overlapping samples and the cleaner sub-sample of disky, not edge-on galaxies. Of the objects NA10 identify as barred, 93% (2348/2537) are objects in GZ2.

Bars in NA10 can be classified according to either bar strength (weak, intermediate, strong) or by other morphological features (ansae, peanuts, or nuclear bar). A galaxy may in rare cases have both a disk-scale (strong, intermediate, or weak) and a nuclear bar. Figure 11 (*top left*) shows that the GZ2 average vote fraction for bars closely agrees with the NA10 fraction of barred galaxies for each GZ2 bin. The two quantities are not identical; the x-axis plots *individual classifications* of galaxies with varying vote fractions for the presence of a bar. The y-axis shows the ratio of barred to unbarred galaxies in NA10. The data have a Spearman's rho of $\rho = 0.984$, and closely follow a 1:1 relationship between the two lines. This is one task in which the aggregate votes of volunteers closely reproduce overall trends in expert classification.

The top right panel of Figure 11 shows the distribution of GZ2 bar votes by simply splitting the NA10 sample in two: galaxies without a bar and galaxies with a bar (of any kind). Both samples show a strong trend toward either extreme, with the strong peak near zero for non-barred galaxies indicating that GZ2 classifiers are very consistent when no bar is present. Possession of a bar is less straightforward; while the frequency of NA10 bars does increase with GZ2 fraction, 32% of NA10-barred galaxies have a GZ2 vote fraction < 0.5. GZ1 data showed similar results for both spirals (Bamford et al. 2009) and mergers (Darg et al. 2010) – a relatively small vote fraction for a feature often indicates that a feature is likely present. This also part of the justification for upweighting small but significant vote fractions in the debiasing process.

Conversely, only 5.5% of non-barred NA10 galaxies have GZ2 vote fractions above 0.5. *Suggestion from SB – look at the objects that have NA no-bar and GZ2 f > 0.5. Can we show that GZ2 is equally robust?*

In the bottom left of Figure 11, the distribution of GZ2 vote fraction as a function of NA10 bar strength is plotted. The distribution for all bars is the same as shown in the top right, increasing with GZ2 vote fraction. There is a clear difference in GZ2 classification between the three sets of bars; interestingly, all three are statistically highly distinct from each other and from the overall barred sample, according to a two-sided K-S test. The majority of both the strong and intermediate barred population have high GZ2 vote fractions, with 83% of strong bars and 56% of intermediate bars above a bar fraction of 0.8. Those numbers increase to 98% and 90%, respectively, if the criterion of 0.5 for the GZ2 vote fraction is used (Masters et al. 2011). Only 13% of weakly-

Are any of these trends (Fig 10 & 11) a function of parameters like Luminosity, stellar mass?

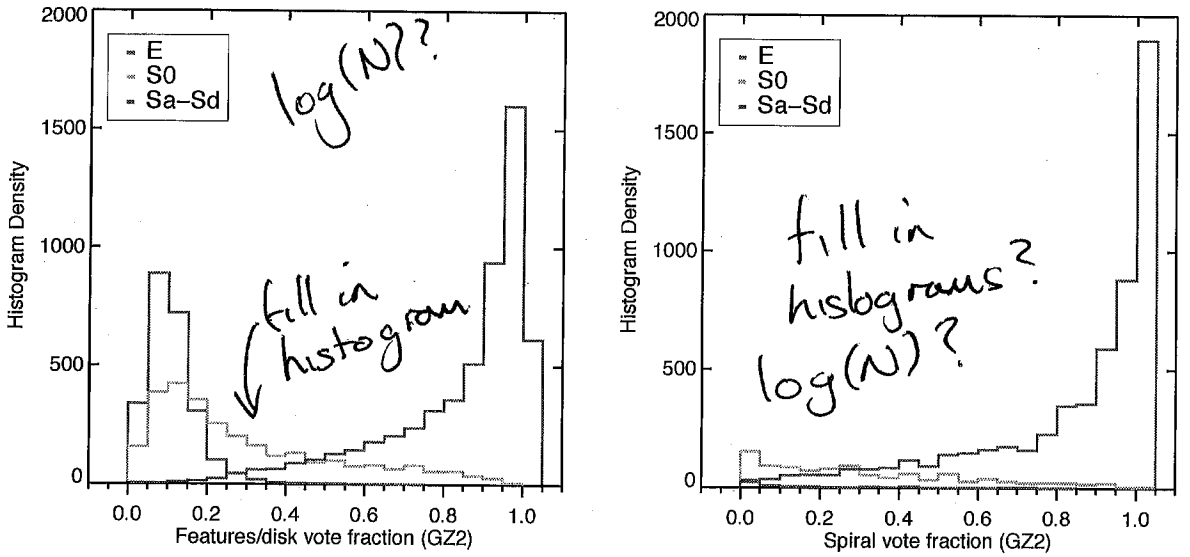


Figure 10. T-type classifications for NA10 and GZ2. Data in the left panel are for the 12,480 galaxies found in both samples; the right panel only shows the 5,683 galaxies with at least 10 responses to Task 04 (visible spiral structure) in GZ2. The distribution is of GZ2 vote fractions separated by their T-type classification from NA10.

barred galaxies have GZ2 vote fractions above 0.8, and 47% have vote fractions above 0.5.

Data from NA10 can be used to quantify a possible threshold to identify barred galaxies in GZ2 data. Based off the distribution of NA10 galaxies (Figure 11, upper right), only a tiny fraction of galaxies above a GZ2 bar fraction of 0.3 genuinely lack a bar. This threshold is more inclusive than the 0.5 used by Masters et al. (2011), but includes 97% of strong and intermediate bars and 75% of weak bars identified by NA10. Below this limit, both the NA10 and GZ2 catalogues are likely to be ~~both significantly~~ contaminated and incomplete, with the existence of a bar subject to differing opinions even among expert astronomers.

?? number

The lack of sensitivity to weak bars from NA10 may also be related to the design of the GZ2 interface. When users were asked if a bar is present, they were shown an icon with two examples of a barred galaxy (Figure 1). The example picture of a disk galaxy has the bar extending across the disk's full diameter, fitting the typical definition of a strong bar. With this as the only example (and no continuum of options between the two choices), GZ2 users may not have looked for bars shorter than the disk diameter, or been less confident in voting for "yes" if they did see them. Results from Hoyle et al. (2011) show that users are fully capable of identifying weak bars in other contexts; however, the construction of our decision tree means that GZ2 classifications only include examples from strong and medium bars.

are you sure?

NA10 identified three other fine-structure features related to bars: ansae, peanuts, and nuclear bars. None of the three correlate strongly with the GZ2 bar parameter, with more galaxies actually having vote fractions < 0.5 than above it. Nuclear bars are the only feature that overlaps with the NA10 bar strength classifications; out of 283 nuclear bars, 3 galaxies also have strong bars, 44 have intermediate bars, and 166 have weak bars. No ansae are detected in the GZ2 subsample of disks that are not edge-on, likely due to the axial cut.

how?

5.2.3 *Rings*

NA10 included three types of ring galaxies in their classifications, based on criteria in Buta & Combes (1996). Inner rings lie between the bulge and spiral arms or disk. Outer rings are external to the spiral arms, but are still closely linked to the spiral pattern. Nuclear rings lie in the bulge region of galaxies; no specific size scale for this is given. In GZ2, rings are classified only if the user selects "yes" for the question *"Anything odd?"* The "odd feature" task has seven responses (ring, lens, disturbed, irregular, other, merger, dust lane), of which a user can select only one; as a result, any galaxies with multiple "odd" features will have votes split among the features, with only the clearest achieving a plurality.[3] While this means that some galaxies with rings may have low vote fractions in the GZ2 classifications, those with high vote fractions are typically strong and distinct.

In the NA10 catalogue, 18.2% of galaxies have a ring. Of those, 10% are nuclear rings, 74% are inner rings, and 32% are outer rings (sum is more than 100% since ~~it is~~ a third of ringed galaxies have multiple rings flagged). In the GZ2 catalogue, 3,142 galaxies are in the clean sample of rings, but this is based on ~~only~~ a potentially small number of total votes ($N \geqslant 5$). In both catalogues, selecting only disk galaxies that are not edge-on ~~did~~ not significantly change the percentage of galaxies identified as having a ring.

does

[3] Future versions of Galaxy Zoo allow multiple responses for this task.

like what?

I feel the presentation of Section 5.2 (and maybe 5 in general) could be improved by
i) ~~more~~ use of tables to summarize comparisons
ii) some conclusions or interpretation; the text is very facts orientated
ii) improved ~~[illegible]~~ plots

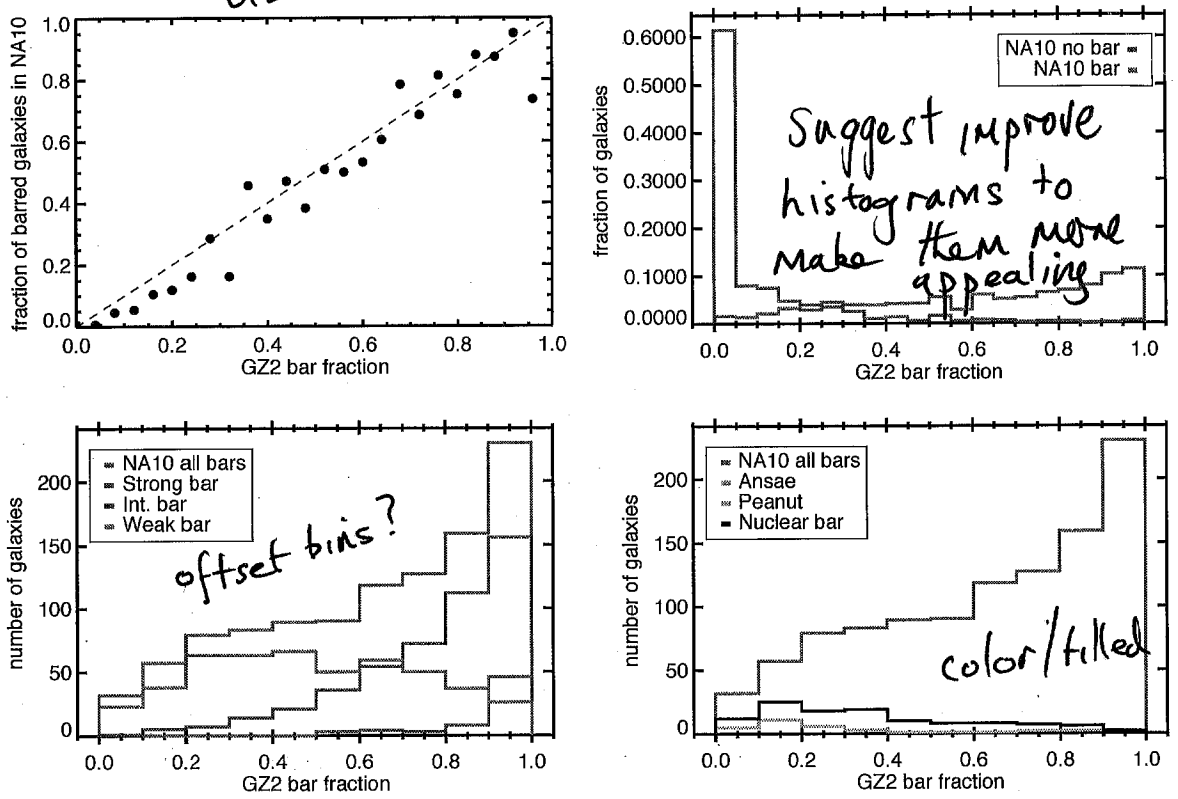


Figure 11. Galactic bar classifications for GZ2 and NA10. Data are for the 7,121 galaxies which are not edge-on (log $(a/b) < 0.3$), have 10 or more GZ2 bar classifications, and appear in both samples. Top left: mean bar fraction per galaxy in GZ2 vs. the ratio of barred to all galaxies in NA10. Dashed line shows the one-to-one relationship. Top right: distribution of the GZ2 bar vote fraction, separated by NA10 classifications. Bottom left: distribution of GZ2 bar vote fraction for the three disk-scale bar categories of NA10. Bottom right: distribution of GZ2 bar vote fraction for ansae, peanut, and nuclear bars in NA10.

In the top-left of Figure 12, the distribution of the number of GZ2 votes for a ring in disk galaxies that are not edge-on is shown, both for the total sample and for galaxies classified by NA10 as having a ring. The distributions grow closer as the number of "yes" votes increases. The top-right panel of Figure 11 shows the cumulative distribution function for the number of ring votes. Among all galaxies with at least 15 "yes" votes, for example, 90% of those galaxies are also identified by NA10 as having a ring; almost all of these are inner or outer rings.

The vote fraction for rings from GZ2 is not a good match to the ring classifications of NA10. Half of all galaxies have a vote fraction of 0, indicating no votes for a ring-like structure in the image. For ringed galaxies identified by NA10, the number of galaxies with no GZ2 votes decreases dramatically, but results a generally flat distribution of ring vote fractions. No single cut on GZ2 vote fraction is a good proxy for the NA10 classifications; at $f < 0.5$, for example, only $\sim 45-65\%$ of the GZ2 ringed galaxies are identified as rings in NA10. This may likely be the result of crosstalk among responses to Task 08.

There is some evidence indicating that GZ2 classifications are sensitive only to certain types of rings. A large fraction of NA10 galaxies with nuclear rings, for example, have many galaxies with no GZ2 ring votes. Several causes are possible: since nuclear rings are smaller, they are more difficult to discern in low surface brightness or bulge-dominated galaxies. In addition, the icon in the GZ2 tree intended to show an example of a ring has a centre dot (a galactic bulge) surrounded by a ring (Figure 1). This could reasonably represent either an inner or outer ring, but might not be associated with the intra-bulge nuclear rings by a non-expert classifier. The bottom-right panel of Figure 12 shows that the number of NA10 galaxies with inner and/or outer rings does rise with vote fraction, with a 55% success rate at $f > 0.8$.

5.2.4 *Mergers/interacting galaxies*

Galaxies in GZ2 can be labeled as a "merger" under the task *"Anything odd?"* NA10 classify possible mergers in two ways: by identifying pairs of objects in an image, and by identifying interacting galaxies. Both NA10 categories have sub-levels: paired objects are sorted by relative separation (close, projected, apparent, or overlapping pairs), and interactions by morphology (disturbed, warp, shells, tails, or bridges). If tidal tails are present, there is an additional flag couting the number of tails.

In the NA10 catalogue, 22.3% of galaxies are labeled as