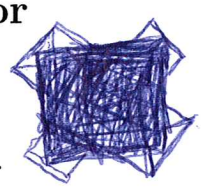
Mon. Not. R. Astron. Soc. **000**, 1–28 (2013) Printed 18 April 2013 (MN LATEX style file v2.2)

# 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 classifications 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

| 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 | 283,971 | 44 | 17.0 |
| original + extra + S82 ($m_r < 17$) | | | |

**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`

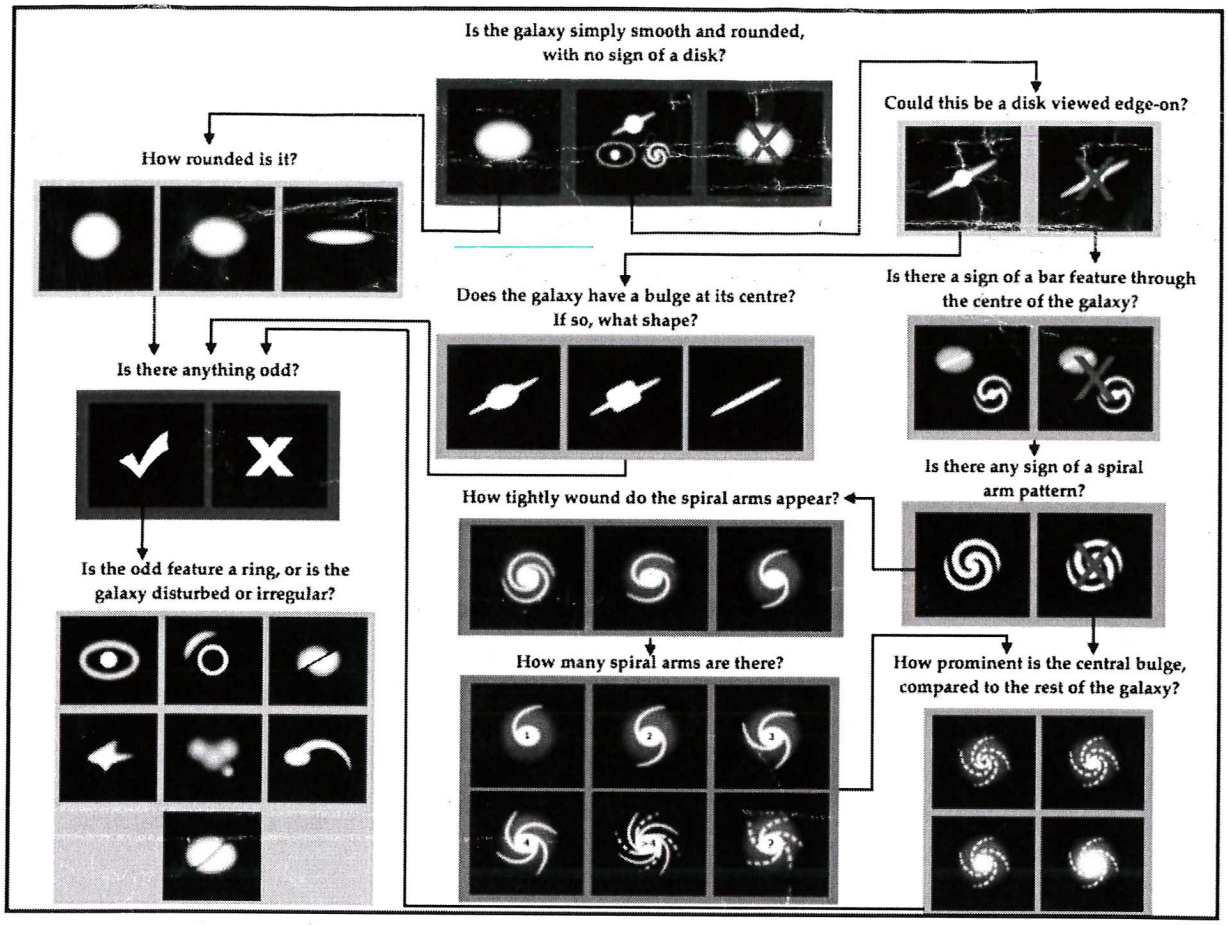



**Figure 1.** Flowchart of the classification tasks for GZ2, beginning at the top centre. Tasks are colour-coded by their relative depths in the decision tree. Tasks outlined in brown are asked of every galaxy. Tasks outlined in green, blue, and purple are (respectively) one, two or three steps below 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.

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.

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

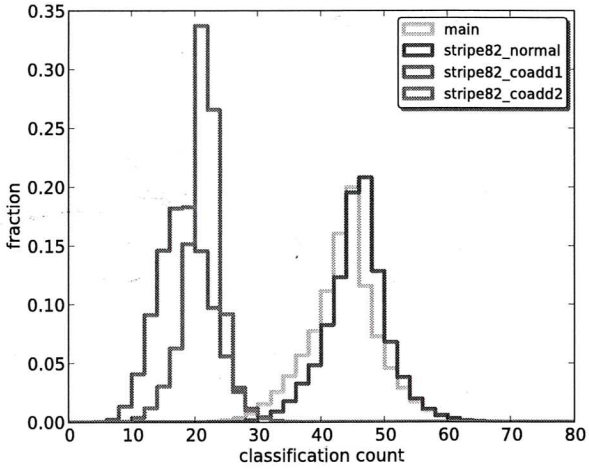



**Figure 2.** Distribution of the number of classifications for the sub-samples within GZ2.

(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 $\kappa$;

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

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

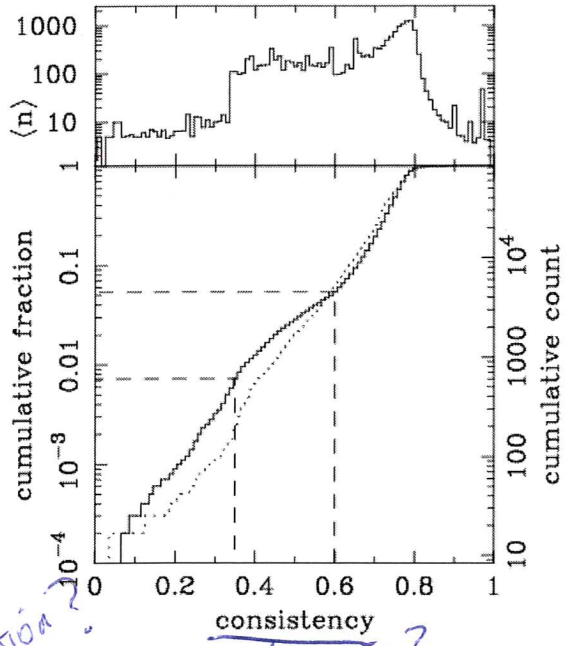



**Figure 3.** Distribution of the user consistency $\kappa$. Top: mean number of galaxies classified per user as a function of their consistency. Bottom: Cumulative fraction distribution of consistency. The dotted line shows the first iteration of weighting, and the solid line the third iteration. The 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:

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

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 $\kappa$, 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-

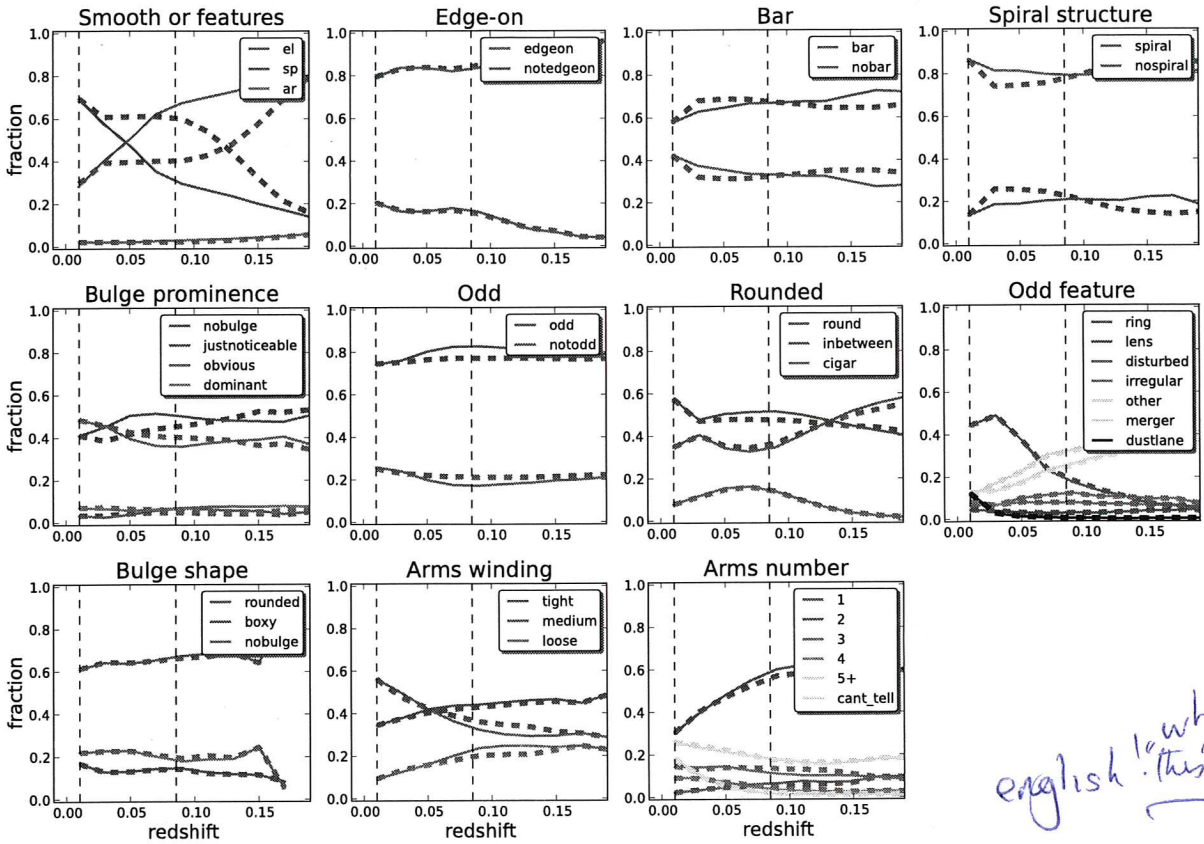



**Figure 4.** Type fractions as a function of redshift for the classification tasks in GZ2. Solid (thin) lines show the vote fractions, while the thick (dashed) lines show the debiased vote fractions adjusted for classification bias. This is a luminosity-limited sample for $M_r < -20.89$. The data for 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:

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

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).