Mon. Not. R. Astron. Soc. **000**, 1–27 (2012) Printed 20 February 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], Steven P. Bamford[3], Karen L. Masters[4], Brooke D. Simmons[2], Kevin Schawinski[5], Lucy Fortson[1], Robert J. Simpson[2], Ramin A. Skibba[6], Edward M. Edmondson[4], Arfon M. Smith[2,7]

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

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

ABSTRACT
Galaxy Zoo 2 data release paper.

Key words: galaxies

1 INTRODUCTION

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

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

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

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

⋆ E-mail: willett@physics.umn.edu
[1] http://zoo2.galaxyzoo.org

in the field (to give one example) with a sufficiently comprehensive starting set.

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

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

2 PROJECT DESCRIPTION

2.1 Sample selection

Objects classified for Galaxy Zoo 2 included around 250,000 of the brightest resolved galaxies from the SDSS North Galactic Cap region. The goal was to exclude the most distant, faintest and smallest systems within which fine morphological features would not be resolved. The 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.

Three further cuts were applied to the DR7 Legacy sample. A Petrosian magnitude brighter than 17.0 in the r-band (after Galactic extinction correction was applied) is required, along with a r-band Petrosian radius greater than 3 arcsec. Galaxies which had a spectroscopic redshift in the DR7 catalogue outside the range $0.0005 < z < 0.25$ were also excluded; 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 are also excluded. Galaxy Zoo 2 was launched with the resulting sample of 245,609 images on 2009-02-16.

An error in the original query meant that the initial Galaxy Zoo 2 sample ("original" sample) was missing some objects ("extra" sample) on launch (see §**??**), specifically those flagged as both BLENDED and CHILD. These objects, which are typically slightly brighter, larger and bluer than the general population, were added to the site on 2009-09-02. The rate at which images from the respective samples were shown to users was tuned so that the "extra" sample would catch-up to the same number of classifications as the "original" 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. These are the only images for which we include multiple images of individual galaxies: both normal single-depth and co-added. The 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 single-depth imaging. This 'stripe82_coadd' sample was added after launch, together with matching galaxies from the normal-depth Stripe 82 imaging ("stripe82" sample). These were mostly added on 2009-09-02 (but note that $\sim$ 7700 of the "stripe82_coadd" sample were not added until 2010-09-24).

Table 1. GZ2 sample properties

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

The initial Stripe 82 coadd images were visually very different from the normal SDSS images. In order to rectify this, another set of Stripe 82 coadd images were produced and added into the site on 2009-11-04. In the tables these two sets of images are indicated by sample names "stripe82_coadd_1" and "stripe82_coadd_2".

Need paragraph from SB or EE on the details for making the coadded images.

For most of the duration of Galaxy Zoo 2, all the images were shown to classifiers in a random order. However, we desired to have all galaxies classified at least a minimum number of times. Therefore, in the final period of Galaxy Zoo 2, accompanied by a competition with a running tally (dubbed the Zoonometer), objects with low numbers of classifications were preferentially shown in an attempt to get them up to a minimum of 40 classifications for the "original", "extra" and "stripe82" samples, and 20 for the "stripe82_coadd_2" sample. The "stripe82_coadd_1" sample was removed from the site at this time. This effort had mixed success, since separate classifications of the Galaxy Wars project also contributed to the count. The main sample galaxies finished with a median of 44 classifications, with 27% having fewer than 40; the "stripe82_coadd_2" galaxies had a median of 21 classifications and 26% of them having fewer than 20 (Table 1).

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

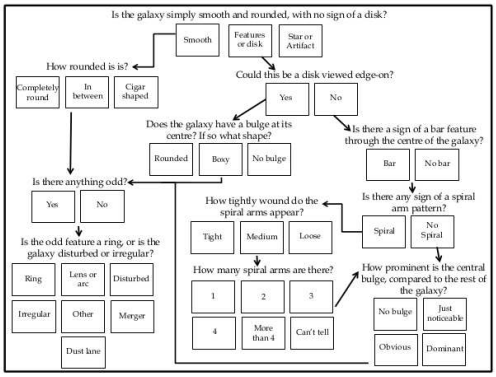

Figure 1. *Top:* Front page of the web interface for Galaxy Zoo 2, displaying Task 01. *Bottom:* Flowchart of the 11 classification tasks for GZ2, beginning at the top center.

2.2 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 classifications to be recorded, but were not required to complete any tutorials. They were then shown a *gri* colour composite image of a galaxy generated from the SDSS ImgCutout web service (Nieto-Santisteban et al. 2004), with the image randomly chosen from our sample database. The image scale is XX arcsec/pixel (add more details here).

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

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

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

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

Table 2. GZ2 classification tree

Task	Question	Responses	Next task
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 center 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

3 DATA REDUCTION

3.1 Multiple classifications

In a small percentage of cases, an individual user may classify the same object more than once. Since we wish to treat each click as an independent measurement, we removed multiple classifications of the same object by a given user from

the data, keeping only the last submitted classification. Such repeat classifications only occurred for a small proportion of objects ($\sim 1\%$), and an even smaller proportion ($\sim 0.01\%$) significantly enough to potentially alter their classifications.

3.2 Consistency and invididual user weighting

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

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

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

$$\kappa_i = \begin{cases} f_r & \text{if click corresponds to this answer,} \\ (1 - f_r) & \text{if click does not correspond.} \end{cases} \tag{2}$$

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

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

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

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

$$w = \text{power}((\kappa/0.6), 8.5) \tag{3}$$

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 2 also shows the lowest-weighted users have on average classified only a handful of objects.

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

3.3 Classification bias

The weighted vote fractions in the data are also adjusted for what we term *classification bias*. The overall effect is a change in observed morphology fractions as a function of redshift, as seen in the original Galaxy Zoo data. The presumed cause is that more distant galaxies are, on average, are both smaller and dimmer as they appear in the cutout images; as a result, finer morphological features are more difficult to identify.

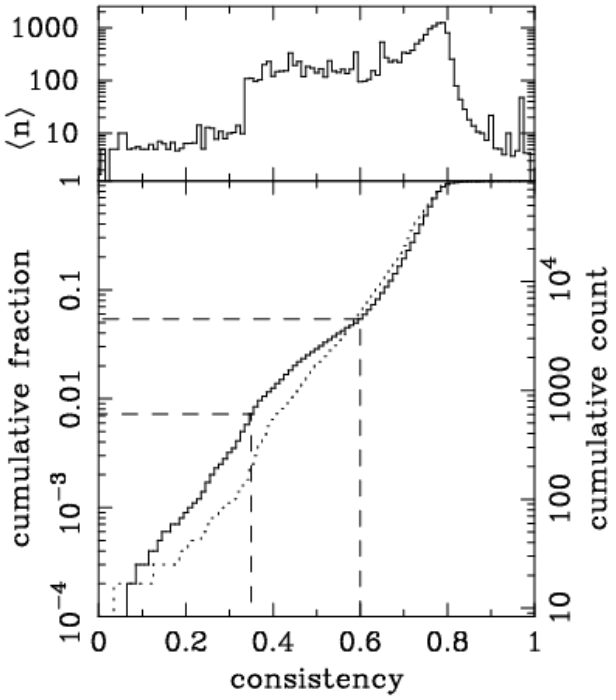

Figure 2. User consistency for Task XX. Top: total number of galaxies classified per user as a function of their consistency. Bottom: Cumulative fraction distribution of consistency. The dotted line shows the first iteration of weighting, and the solid line the third iteration. The second iteration is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

Figure 3 demonstrates the classification bias for several of the Galaxy Zoo 2 classification tasks. The average weighted vote fraction for each response is shown as a function of redshift; votes for finer morphological features (such identification of disk galaxies, spiral structure, or galactic bars) decrease at higher redshift. Part of this effect is due to the nature of a luminosity-limited sample; high-redshift galaxies must be more luminous to be detected in the SDSS and are thus more likely to be giant red ellipticals. Even if the sample is magnitude-limited, however, the vote fractions show similar changes at high redshift.

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

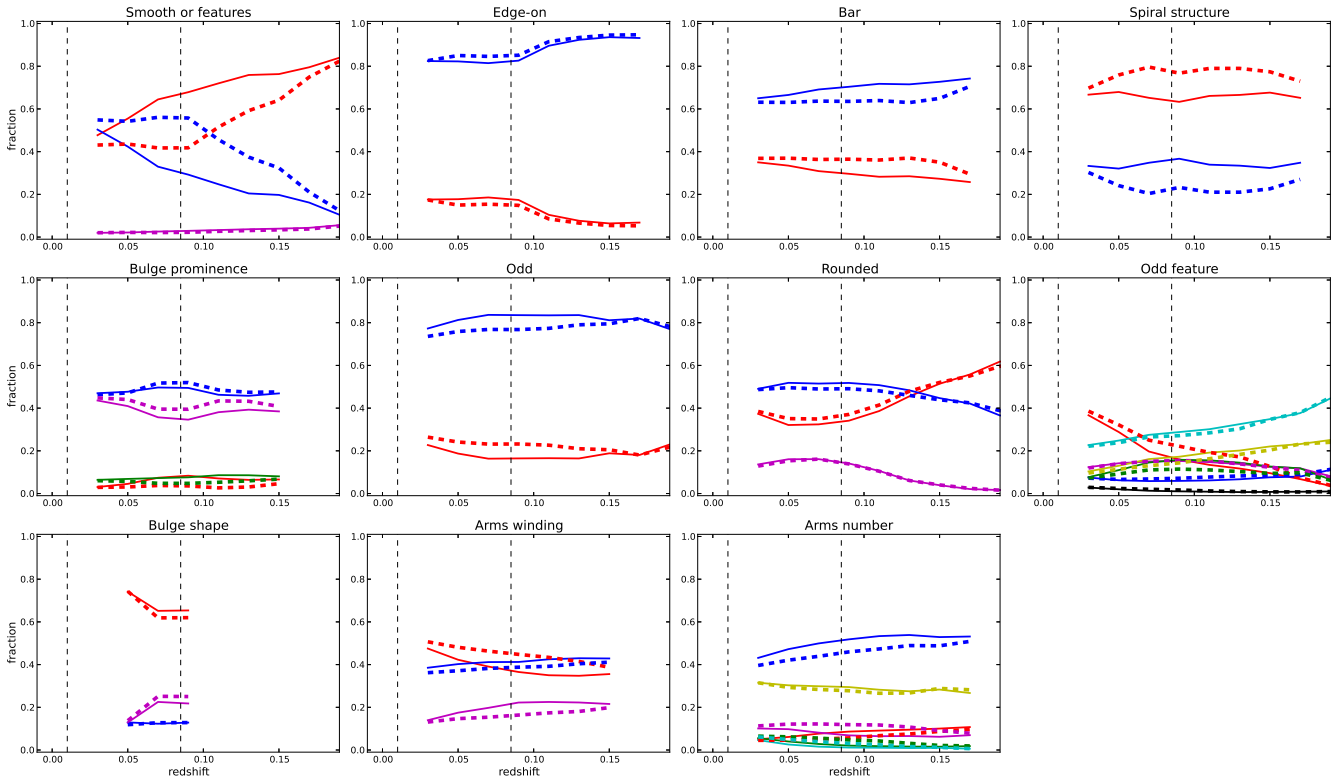

Figure 3. Type fractions for the classification tasks in GZ2. Solid (thin) lines show the weighted vote fractions, while the thick (dashed) lines show the debiased vote fractions that have been adjusted for classification bias. This is a magnitude-limited sample for $M_r < -20.17$. Vertical dashed lines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

ify and extend this technique for the Galaxy Zoo 2 classifications.

There are two major differences between the GZ1 and GZ2 data. First, GZ2 has a decision tree, rather than a single question and answer for each click on an image. This means that all tasks, with the exception of the first, depend on answers to previous classifications in the tree. For example, the bar question is only asked if the user answers "features or disk" for Task 01 and "no" to Task 02 ("edge-on disk"). Thus, the value of the weighted vote fraction for this task only addresses the total bar fraction *among face-on disk galaxies*.

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

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

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

$$\frac{p_i}{p_j} = \frac{f_i}{f_j} \times K_{j,i}. \quad (4)$$

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

$$p_i = \frac{1}{1/p_i}, \quad (5)$$

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

unity:

$$p_i + p_j + p_k + \dots = 1. \quad (6)$$

Multiplying (5) by (6) yields:

$$p_i = \frac{1}{1/p_i} \times \frac{1}{p_i + p_j + p_k + \dots} \quad (7)$$

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

$$p_i = \frac{1}{\sum\limits_{j \neq i} (p_j/p_i) + 1} \quad (9)$$

$$p_i = \frac{1}{\sum\limits_{j \neq i} K_{j,i}(f_j/f_i) + 1}. \quad (10)$$

The corrections for each pair of tasks can be directly determined from the data. At the lowest sampled redshift bin ($z \simeq 0$), $\frac{p_i}{p_j} = \frac{f_i}{f_j}$ and $K_{j,i} = 1$. From Equation 4:

$$\left(\frac{f_i}{f_j}\right)_{z=0} = \left(\frac{f_i}{f_j}\right)_{z=z'} \times K_{j,i} \quad (11)$$

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

$$(13)$$

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

$$C_{j,i} = \log\left[\left(\frac{f_i}{f_j}\right)_{z=z'} / \left(\frac{f_i}{f_j}\right)_{z=0}\right] \quad (14)$$

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

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

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

The weighted vote fractions for each task response are binned in three dimensions: the absolute magnitude M_r, the Petrosian r-band half-light radius R_{50}, and redshift z. Bins range for M_r range from -24 to -16 in steps of 0.25 mag, for R_{50} from 0 to 15 kpc in steps of 0.5 kpc, and for z from 0.01 to 0.26 in steps of 0.01. The bin ranges and step sizes are chosen to maximize the phase space covered by the bias correction, while also retaining enough galaxies in each bin to establish its morphology distribution. The value of each bin in the cube is the sum of the weighted vote fractions for that response. For each pair of responses (i, j) to a question, we compute $\log(f_j/f_i)$ in every (M_r,R_{50},z) bin. The local baseline relation is established by selecting the value in the non-empty bin(s) for the lowest-redshift slice at a given (M_r,R_{50}).

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

The baseline relations for the GZ2 tasks are shown in Figure 4. Each pair of responses has a markedly different shape of the baseline ratio. The baseline ratio for the "smooth" and "features/disk" responses to Task 01 is functionally very similar to the GZ1 relation (Figure A5 in Bamford et al. 2009), as expected. It is reasonably well-fit with an analytic function of the form:

$$\frac{f_j}{f_i}[R_{50}, M_R] = \frac{f}{1 + exp[(x_0 - M_R)/x_1]} + g \quad (16)$$

where:

$$x_0 = b^{-\left(a + hR_{50}{}^i\right)} + c \quad (17)$$

$$x_1 = d + e(x_0 - c) \quad (18)$$

and where $\{a, b, c, d, e, f, g, h, i\}$ are minimized to fit the data. The only other task had baseline ratios reasonably well fit by the same function was Task 07 (rounded smooth galaxies). None of the other tasks are well-fit by this function; for these, we instead adopt a simpler fit where both M_r and R_{50} vary linearly:

$$\frac{f_j}{f_i}[R_{50}, M_R] = t_1(R_{50} - t_2) + t_3(M_R - t_4) + t_5, \quad (19)$$

and $\{t_1, t_2, t_3, t_4, t_5\}$ are the parameters to be minimized. We fit Equation 19 to all other tasks where the number of bins is sufficient to get a reasonable fit. For pairs of responses with few sampled bins, we take the direct difference between the local ratio and the measured ratio at higher redshift. Galaxies falling in bins that are not well-sampled are assigned a correction of $C_{i,j} = 0$ for that term; this is necessary to avoid overfitting based on only a few noisy bins.

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

4 THE CATALOG

The data release for Galaxy Zoo 2 consists of XX tables, abridged portions of which appear in this paper. Table 3 contains classification data for the 283,971 galaxies in the main sample. Each galaxy is identified by its unique SDSS DR7 objID, as well as its original sample designation

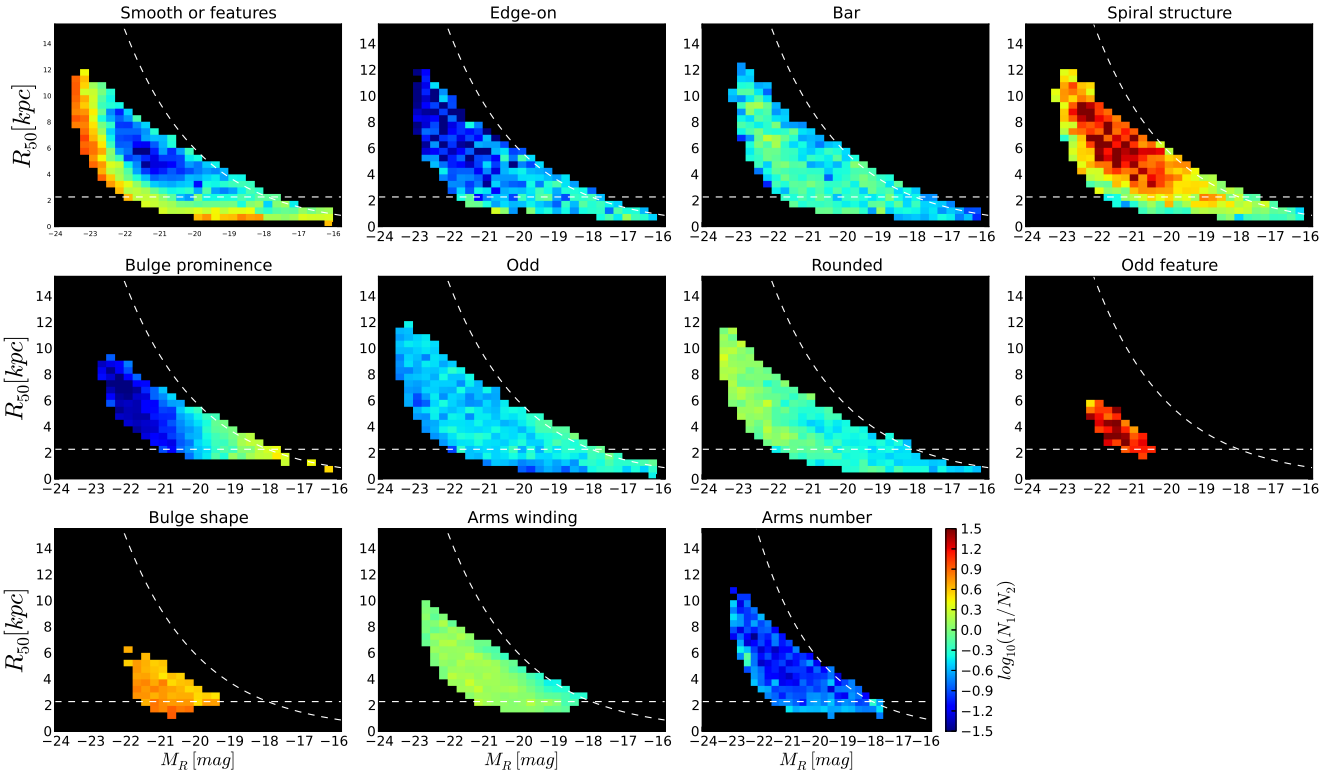

Figure 4. Local morphology ratios for morphology classifications in GZ2. The ratio of the binned vote fractions for morphologies is for the first two responses in the decision tree (Table 2) for each task; there may be as many as 21 such pairs for tasks with more than two options. The dashed horizontal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

(either original, extra or Stripe 82). N_{class} is the total number of users who have classfied the galaxy, while N_{votes} gives the total number of clicks summed over all classifications and all responses. For each of the 37 morphological classes, we give six parameters: the raw number of votes (eg, `t01_smooth_or_features_a01_smooth_count`), the number of votes weighted for consistency `t01_smooth_or_features_a01_smooth_weight`, the raw vote fraction `t01_smooth_or_features_a01_smooth_fraction`, the vote fraction weighted for consistency `t01_smooth_or_features_a01_smooth_weighted_fraction`, the debiased likelihood `t01_smooth_or_features_a01_smooth_debiased`, which is the weighted fraction adjusted for classification bias (see Section 3.3), and a boolean flag `t01_smooth_or_features_a01_smooth_flag` that is set if the galaxy is included in a clean, debiased sample.

The flags for each morphological parameter are determined by applying three criteria: the first is the requirement that more than 50% of votes for preceding task(s) must eventually select for the task being flagged. For example, to select clean barred galaxies, we require both $p_{features/disk} \geqslant 0.5$ and $p_{notedge-on} \geqslant 0.5$. Secondly, the object must exceed a minimum number of total votes (ranging from 10-30) for that task. Finally, the debiased vote fraction for the response must exceed a threshold; this is 0.5 for Tasks 02 and 03, and 0.8 for all other tasks. The latter option is the same as chosen for GZ1.

Table 4 has the same format as Table 3.

Tables **??** and **??** contain data for the co-added images in Stripe 82.

Table **??** contains the Stripe 82 galaxies with $r > 17.0$.

Finally, although not reproduced in this paper, the repository at `data.galaxyzoo.org` contains a pre-matched table containing SDSS metadata for the spectroscopic galaxies in the GZ2 main sample. This contains some of the most commonly used DR7 parameters including SDSS exposure information, position, photometry, size, and redshift. This has the same rows as Table **??** and is provided as a resource for members of the community who wish to compare the morphological data against external parameters.

Catalog will contain:

- raw vote fractions for each response
- debiased vote fractions for each response
- "clean" flags for solid detections on each response
- number of weighted votes for each response
- Chris' other metrics on classification confidence? (see GZ1, Table 4)
- Galaxy Wars?
- Voronoi tessellation bins?
- The matched SDSS data?

4.1 Stripe 82

Another possible method for debiasing the full GZ2 sample is to use the data from Stripe 82 to establish the unbiased

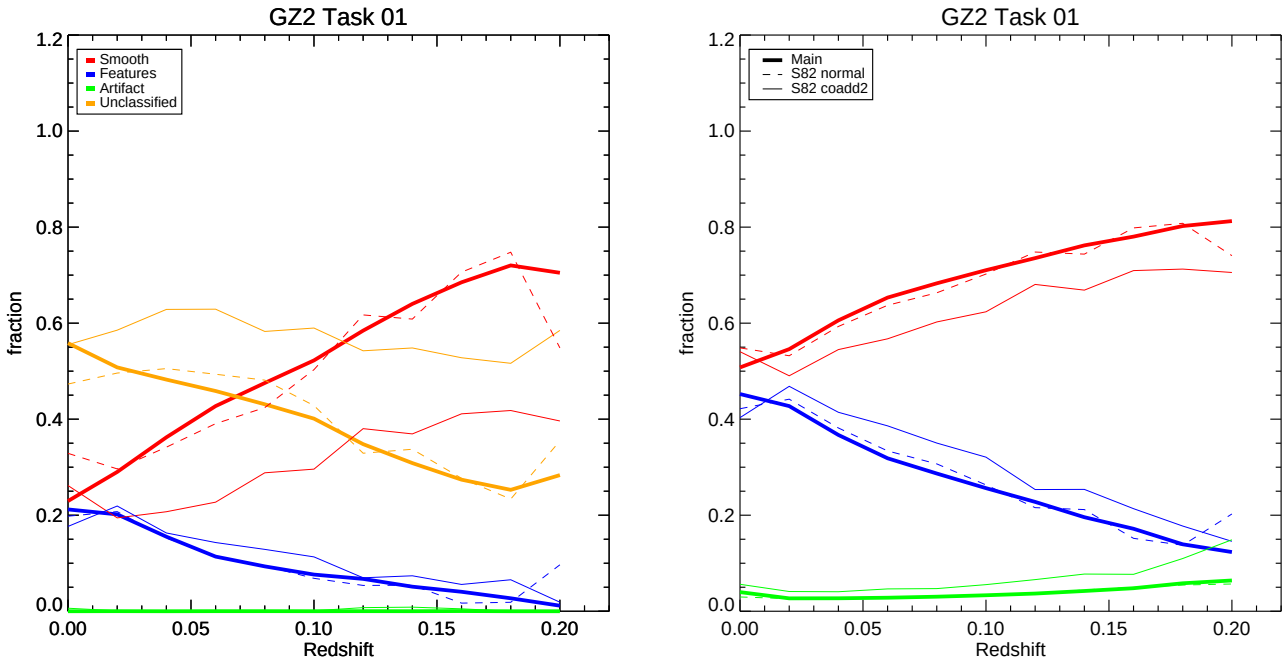

Figure 5. GZ2 weighted vote fractions for Task 01 (*smooth, features/disk, or star/artifact?*) as a function of spectroscopic redshift. The left graph shows the fraction of galaxies for which a category exceeded a threshold of 0.8. Galaxies which had no answer above the threshold are labeled as "unclassified".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

level. Normal-depth images for Stripe 82 are selected in precisely the same manner as the main GZ2 sample, but with a deeper magnitude limit of $r_{petro} < 17.77$. The fiber size and redshift range are identical, resulting in a comparison sample of roughly 10% the size of the GZ2 original + extra samples.

From the gz2sample – gz2table match (325651 rows):

- GZ original + extra: no redshifts for 38820/273783 galaxies = 14.2%
- GZ s82 normal depth: no redshifts for 3735/21522 galaxies = 17.4%
- GZ s82 coadd1 depth: no redshifts for 10581/30346 galaxies = 34.9%
- GZ s82 coadd2 depth: no redshifts for 10578/30339 galaxies = 34.9%

Galaxies in Stripe 82 have both "normal-depth" images as well as deeper images created by co-adding exposures from the runs that SDSS did in the region prior to DR7 (Annis et al. 2011). With increased sensitivity to fainter magnitudes, this should mean that galaxies at higher redshifts will have their morphological classifications altered, presumably in the direction of higher accuracy. The spectroscopic incompleteness fraction is similar for the GZ2 original + extra (14%) and Stripe 82 normal depth (17%) samples, but is twice as high for the galaxies in the Stripe 82 coadded data (35%).

The hypothesis is that the change in vote fraction as a function of galaxy metadata (z, R_{50}, m_R, μ) would have the same effect on classifying GZ2 main sample galaxies at higher redshift in the same R_{50}, m_R, μ bins.

Firstly, we need to verify that the distribution of votes for the main sample is similar enough to the Stripe 82 normal-depth (with $r < 17.0$) that the same bias correction will apply for both. Table 5 shows the tasks in the GZ2 question tree and the mean weighted vote fraction for each response. The distributions for both the main sample and Stripe 82 galaxies are quite similar, with the difference in the mean varying by $< 10\%$ for almost all responses. The only exceptions to these are for responses that target rare objects (and thus are subject to higher variance for low-number statistics), such as dust lanes, rings, and high-multiplicity spiral arms.

The right panel of Figure 5 shows that the weighted vote fractions also behave similarly as a function of redshift, particularly in the $0.01 < z < 0.08$ range covered by the GZ1 debiasing technique. The agreement is generally good between the Stripe 82 normal depth and the GZ2 main sample; this is not the case for the coadded Stripe 82 data, however. For Task 01, fewer galaxies are classified as robustly smooth (above the 0.8 threshold), moving instead to the "unclassified" category. Similarly, the coadded data shows higher fractions of galaxies with bars (Task 03) and for possessing visible spiral structure (Task 04). A possible cause for this is that the new image pipeline in the coadded data allows viewers to see faint features or disks. The average seeing in the coadded data improves from $1.4''$ to $1.1''$ (Annis et al. 2011), which may need to be corrected for in the Stripe 82 comparisons.

What needs to be done:

• Verify that the normal-depth Stripe 82 vote fractions are similar to the original + extra samples. **Yes.**

• See whether co-added images follow a different distribution, esp. as function of z, R_{50}, M_r, μ. **They do, even accounting for the magnitude cut.**

• Decide which of the two co-added samples is more reliable for debiasing efforts. **The only difference between the two samples is a moderate increase in spiral fraction for disk galaxies in coadd1 data. The results of K-S tests on the weighted vote fractions are confusing. Very similar distributions (by eye) have very low probabilities according to the K-S test. Perhaps quote the various moments instead.**

• How far out in redshift can we correct the normal data using coadd? Need enough galaxies per bin in M_r, R_{50}, z space *for each task* in the Stripe 82 data to establish the baseline. **20 galaxies per bin at** $\Delta M_r = 0.4$ mag, $\Delta \mathrm{R}_{50} = 0.8$ kpc, $\Delta \mathrm{z} = 0.01$ **covers a comparable amount of parameter space compared to the GZ1 results of B09.**

• Figure out how the vote fractions are going to be adjusted once a bias has been identified. **Have attempted to correct results for binary tasks based on the B09 method.**

• Apply corrections to the vote fractions in the GZ2 main sample. **Done; works well over redshift range (**$0.01 < z < 0.09$**). See Figure 7**.

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

The biggest systematic difference is for Task 05, Answer 11 (prominence of the bulge is "just noticeable"), for which the mean weighted fraction in coadd2 data is $\sim 35\%$ higher than from coadd1 data. This is an opposite (but not equal) effect than Answer 12 (obvious bulge), for which the coadd1 data is $\sim 13\%$ higher; this may indicate a general shift in votes toward a more prominent bulge. A similar but smaller effect is seen in classification of bulge shapes for edge-on disks (Task 09), where votes for "no bulge" in coadd1 data go to "rounded bulge" in coadd2. The specific cause for these effects as it relates to the image quality is unknown.

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

The main arguments against using the Stripe 82 data are twofold. First, it must be shown that the increased depth of the coadded images is sufficient to correct the main sample data over a sufficiently large redshift range. This could be accomplished by using the deep Stripe 82 images as the assumed correction and then applying Equations **??** to Task 01. The range over which the type fractions are flat as a function of redshift then determines the fraction of the total main sample to which these can be applied. If the number of galaxies included are a small fraction, then this method is not practically useful for debiasing the GZ2 sample, since we want robust classifications for as many galaxies as possible.

The second argument against the Stripe 82 data is that the correction derived is likely subject to larger amounts of uncertainty due to the fact that it only has $\sim 10\%$ of the number of galaxies as the main sample. This means either smaller coverage for the correction in (z, M_r, R_{50}) space and/or larger bins to achieve a minimum sampling. I can also show this in plots.

We applied the bias corrections derived from the main sample to the galaxies at normal depth in Stripe 82 (Figure 7). The correction flattens the redshift effect in all tasks, similar to the main sample data. The variance along redshift bins is somewhat higher – formal error bars will need to be computed to see if there is any statistical difference, or whether the result is consistent with a smaller total sample of galaxies.

5 COMPARISON OF GZ2 TO OTHER CLASSIFICATION METHODS

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

The main purpose of this section is to perform preliminary comparisons between the weighted, but not debiased data from GZ2 available in early 2012. This analysis will need to be repeated to account for biases in the data; hopefully with the groundwork laid here, it will be faster to repeat it later.

5.1 Galaxy Zoo 1 vs. Galaxy Zoo 2

As a check of the classification consistency, we can compare the results from GZ2 to those in GZ1 (Lintott et al. 2011). The galaxies in GZ2 are a subset of those in GZ1, with 248,883 matches between the samples. Task 01 in GZ2 is almost identical to the interface of GZ1. GZ1 allowed for selection of "merger" and "don't know" options in addition to the first three; and asks for galaxies with "features or disk" rather than only for spiral structure.

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

There are 83,956 galaxies identified as "clean" spirals in GZ1. The agreement with the "features or disk" response in GZ2, however, is significantly lower. Only 31.6% of the GZ1

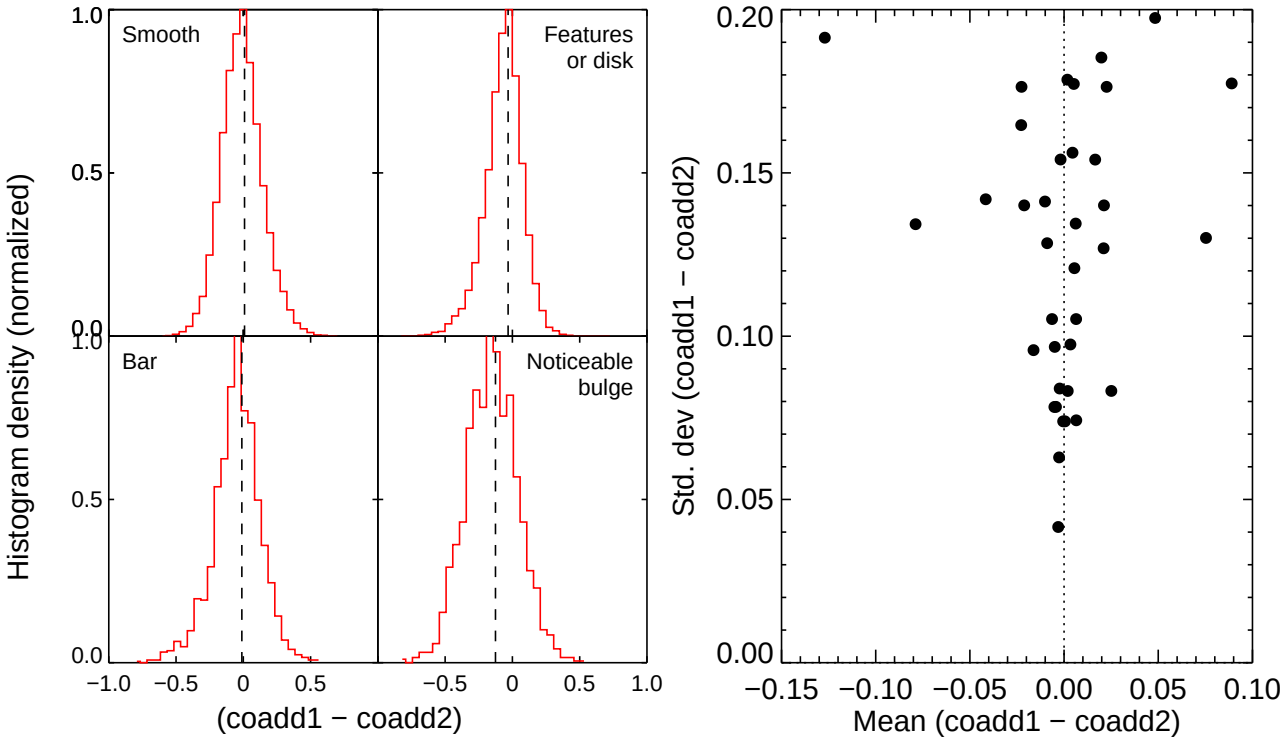

Figure 6. Comparison of GZ classifications from the different coadd techniques for Stripe 82. Left: Distribution of the difference in weighted vote fractions for galaxies that appear in both the coadd1 and coadd2 samples. Each panel shows selected answers in the tree galaxies with at least 10 responses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

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).

Since the GZ1 vote fractions were specifically directed to galaxies with spiral arms, the results should also be compared to GZ2 Task 04, which specifically asks for spiral structure. The agreement with the GZ1 clean spirals is higher than for Task 01, but still well short of that for smooth/elliptical galaxies. 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.

Add trumpet plot; GZ2 is more conservative than GZ1.

Figure 8 shows 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 being 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 centered 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 also had an option to identify mergers; this was a rare response, comprising less than a percent of the total type fraction at all redshifts (Bamford et al. 2009). Darg et al. (2010) found that a weighted 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 particularly high when considering that other reponses to that question (such as "disturbed", "irregular", or "other") could conceivably be applied to merging systems. The agreement between GZ1 and GZ2 is surprisingly high, but indicates that mergers were reliably identified in both despite the additional options in the GZ2 question tree.

5.2 Nair & Abraham

Nair & Abraham (2010a, ; hereafter NA10) catalogued 14,034 galaxies from the SDSS DR4. The galaxy redshifts range from $0.01 < z < 0.1$, and go down to an extinction-corrected apparent magnitude limit of $g < 16$. Overlap with the GZ2 sample is substantial; joining on their objIDs, 12,480 galaxies are found in both catalogs. This comprises 89.9% of the NA10 catalog, but only 4.5% of the total GZ2 sample. The GZ2 sample is deeper ($r < 17$), spans a wider

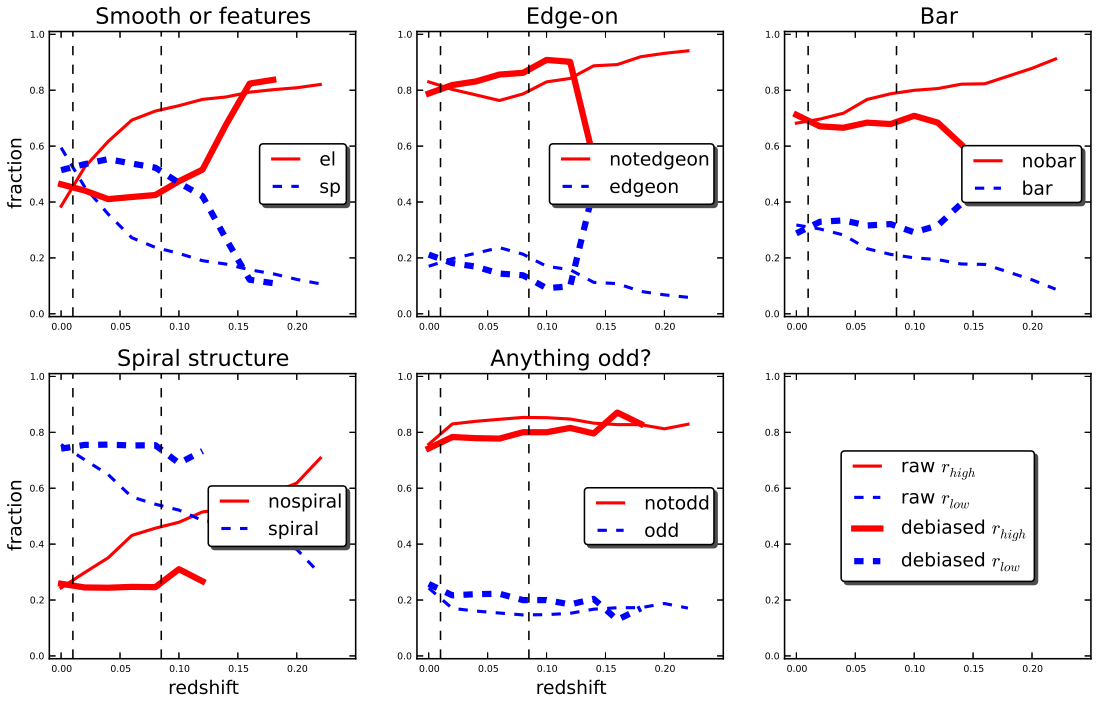

Figure 7. Type fractions for the five binary tasks in GZ2 for the normal-depth Stripe 82 data. This is a magnitude-limited sample for $M_r < -20.17$. Vertical dashed lines show the redshift at $z = 0.01$ (the lower limit of the correction) and $z = 0.085$ (the redshift at which the absolute magnitude limit reaches the sensitivity of the survey).

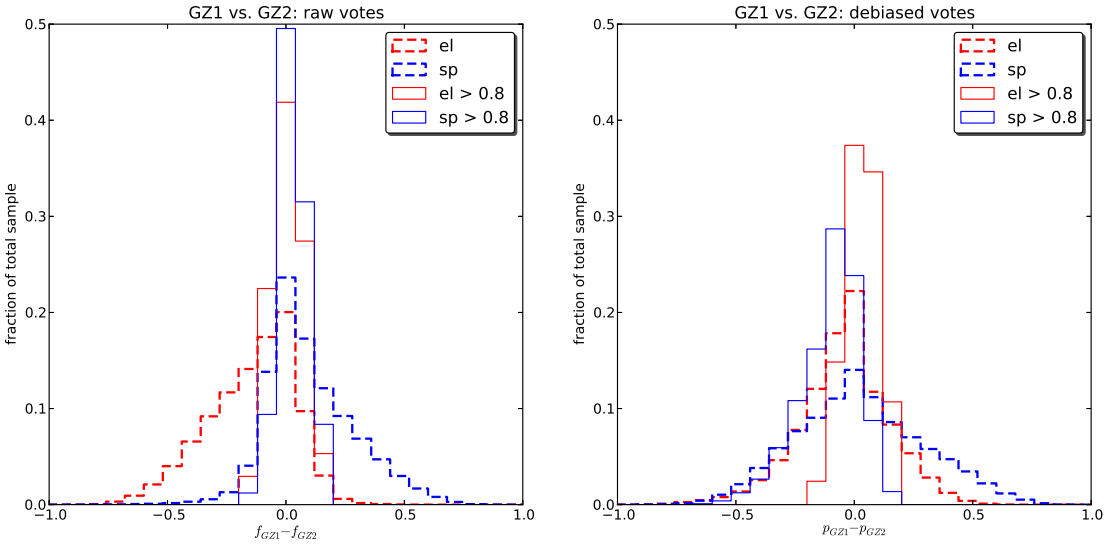

Figure 8. Differences in the vote fractions for galaxies in both the Galaxy Zoo 1 (GZ1) and Galaxy Zoo 2 (GZ2) projects. Left: Distribution of the differences in the raw weighted vote fractions. Dashed lines show data for all galaxies, while solid lines are for the subset in which f_{el} or $f_{sp} >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

redshift range ($0.0005 < z < 0.25$), and contains more recent data releases (DR7) than the NA10 catalog.

Nair & Abraham (2010a) used classifications by a single professional (P. Nair) to carry out morphological classifications. She determined RC3 T-Types for the entire sample through visual inspection of g-band images, covering each source twice. There is no discussion as to the procedure if T-Types changed between her first and second classifications.

In addition to the T-Types, she also classified various "fine structure" morphological features in each galaxy. It is not stated how many times these classifications were performed. 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)

Several of the fine structure features can be directly compared to the GZ2 data: bars, rings, and pairs/interactions.

5.2.1 Bars

NA10 detect 2537 barred galaxies, which is 18% of their total sample. For objects with T-Types later than E/S0, this rises to 25% of the sample. This is much lower than the fraction found in local galaxies ($\sim 60\%$), but similar to the fraction found by (Masters et al. 2011) for the bar fraction in disky, face-on galaxies from GZ2 (29.4%).

Two parameters can be set that reduce the number of galaxies in the overlap between the samples, but result in a cleaner cut for comparisons. The first uses the Masters et al. (2011) cut for face-on galaxies ($\log(a/b) < 0.3$ using the `expab_r` parameter in the Galaxy table). The second is to only look at galaxies with at least 10 classifications for task 03 (*bar present?*) in GZ2. The reduced face-on sample has 7,121 galaxies from the original 12,480. All trends described below hold generally for both the full overlapping samples and the cleaner face-on sub-sample.

Of the objects NA10 identify as barred, 2348/2537 (93%) are objects in GZ2. Figure 9 shows correlations for face-on galaxies between the bar fractions of the NA10 and GZ2 catalogs. Galaxies in NA10 are classified by a single person, and so the presence of a bar is indicated with a series of flags. Bars in a galaxy can be classified according to 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 9 shows that the raw GZ2 weighted bar fraction (WBF) very closely agrees with the NA10 fraction of barred galaxies for each GZ2 bin. The two quantities are not identical; the x-axis plots galaxies with *individual classifications* at whatever fraction of votes chose a bar. The y-axis shows that for galaxies in that WBF bin from GZ, the average bar classification is similar to the weighted vote totals. The distribution shows a Spearman's rho of $\rho = 0.984$, and closely follows a 1:1 relationship between the two lines. If supported by debiased GZ2 data, this means that the aggregate votes of Zooites closely reproduce the expert classification of NA10.

In the top right, Figure 9 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 extrema, although there are a large number of galaxies in GZ2 with a fraction of 0.0 than one would expect. Possession of a bar is less straightforward; while the frequency of NA10 bars does increase with GZ2 fraction, 32% of NA-barred galaxies lie below the canonical GZ2 WBF of 0.5. Conversely, there are very few false positives (5.5% of non-barred NA10 galaxies) above this value.

In the bottom left of Figure 9, the distribution of GZ2 WBF 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 WBF. There is a clear difference in GZ 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 WBFs, 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 WBF is used (Masters et al. 2011). Weakly-barred galaxies have only 13% of their GZ2 WBFs above 0.8 and 47% above 0.5.

NA10 identify three other fine-structure features related to bars: ansae, peanuts, and nuclear bars. None of the three correlate strongly with a high GZ2 WBF, with most galaxies in all three features at less than 0.5. Nuclear bars are the only feature that overlaps with the NA10 bar 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 face-on subsample of galaxies, due to the axial cut.

In the full GZ2 sample (original + extra), 87,651 galaxies had at least 10 responses to the question *"Is there a sign of a bar feature?"* – 22.1% of them had a WBF greater or equal to 0.5. In the clean face-on sample, this rises to 28.7%.This is in very good agreement with both the 29.4% fraction found in the smaller volume-limited sample of Masters et al. (2011), and with the $\sim 30\%$ for disk galaxies with $(b/a) > 0.4$ (or $\log(a/b) < 0.398$) and T-types of S0 or later of Nair & Abraham (2010b).

In the published literature, (Masters et al. 2011) mention the NA10 catalog and note that both papers agree with a bar fraction depending on morphology with a minimum near the division between the blue and red sequences. There is no detailed comparison between the galaxies in the two samples.

5.2.2 Rings

The NA10 catalog also classifies ring galaxies in their sample. They include three basic types of rings in the catalog based on the Buta & Combes (1996) criteria. Inner rings lie between the bulge and spiral arms or disk. Outer rings are external to the spiral arms, but are still closely linked

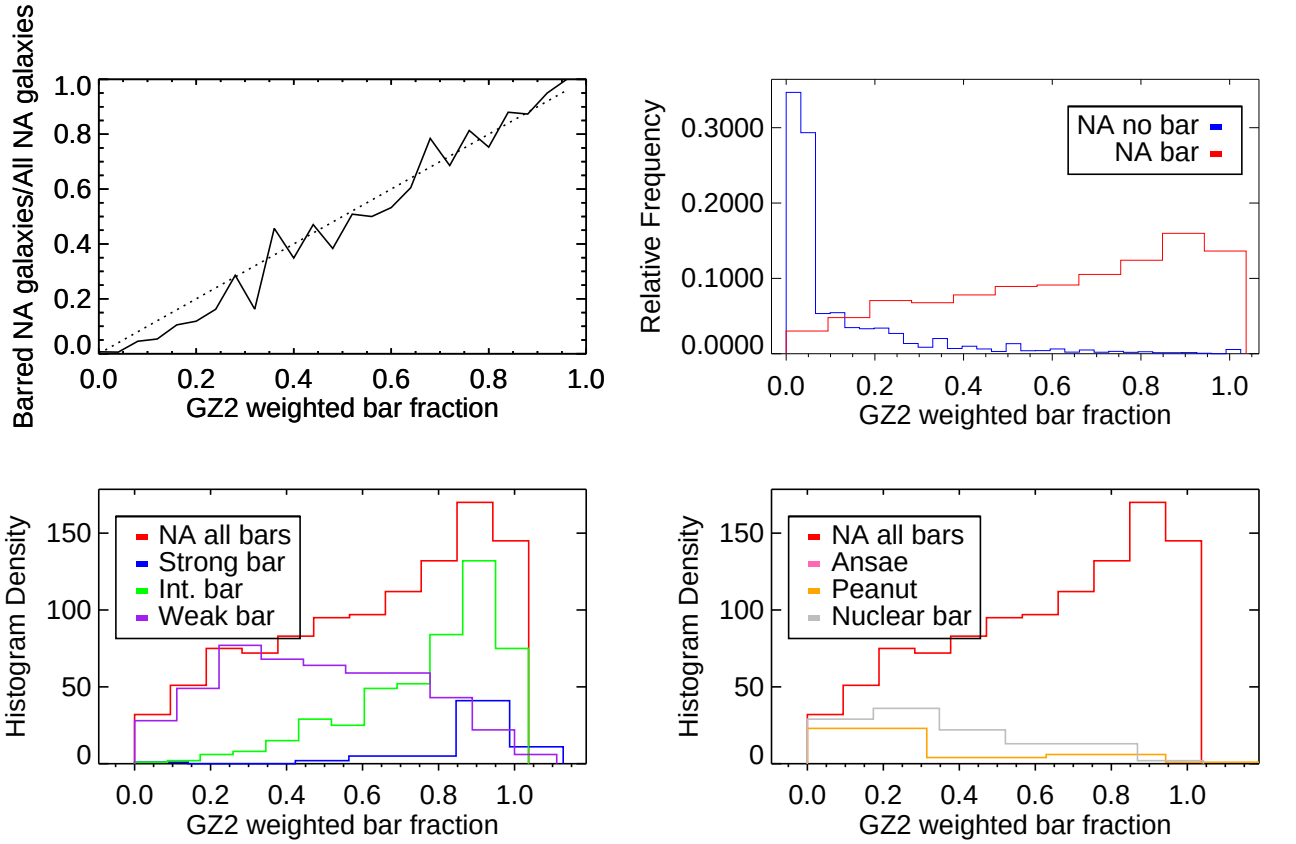

Figure 9. NA10 bar classifications compared to GZ2. Data are for the 7,121 galaxies which are face-on ($\log(a/b) < 0.3$) and with 10 or more GZ2 bar classifications.

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 answers "Yes" to the question *"Anything odd?"* Since the user then has six different options to choose from (ring, lens, disturbed, irregular, other, merger), interpreting the weighted ring fractions is less straightforward.

In the NA10 catalog, 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 $\sim 1/3$ of ringed galaxies have multiple rings flagged). In the GZ2 catalog, 8.2% of galaxies have at least 10 votes for a ring, while 22% (10%) have a weighted fraction higher than 0.5 (0.8). In both catalogs, selecting only face-on galaxies makes no significant changes in the percentage of galaxies identified as having a ring.

In the top-left of Figure 10, the distribution of the number of GZ2 votes for a ring in face-on galaxies 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 9 shows the CDF for the number of ring votes. Among all galaxies with at least 15 "Yes" votes, for example, $\sim 90\%$ of those galaxies are also identified by NA10 as having a ring; almost all of these are inner or outer rings.

In contrast, the weighted ring fraction (WRF) from the GZ2 votes is not a good match to the ring classifications of NA10. Half of all galaxies have a WRF of 0.0, indicating no votes for a ring-like structure in the image. For ringed galaxies identified by NA, the number of WRF=0.0 objects decreases dramatically, but results a generally flat distribution of ring fractions. No single cut on WRF is a good proxy for the NA10 classifications; at all WBF above 0.5, for example, only $\sim 45-65\%$ of the GZ2 ringed galaxies are identified as rings in NA10.

There is some evidence indicating that GZ2 classifications are sensitive only to certain types of rings. NA10 galaxies with nuclear rings, for example, have a large number of 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 button in the GZ2 tree intended to show an example of a ring has a center dot (presumably a galactic bulge) surrounded by a ring. This could reasonably represent either an inner or outer ring, but would presumably not be associated with the intra-bulge nuclear rings by an untrained user. The bottom-right panel of Figure 10 shows that the number of NA10 galaxies with inner and/or outer rings does rise with WRF, but is still only 55% successful at WRF> 0.8.

5.2.3 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

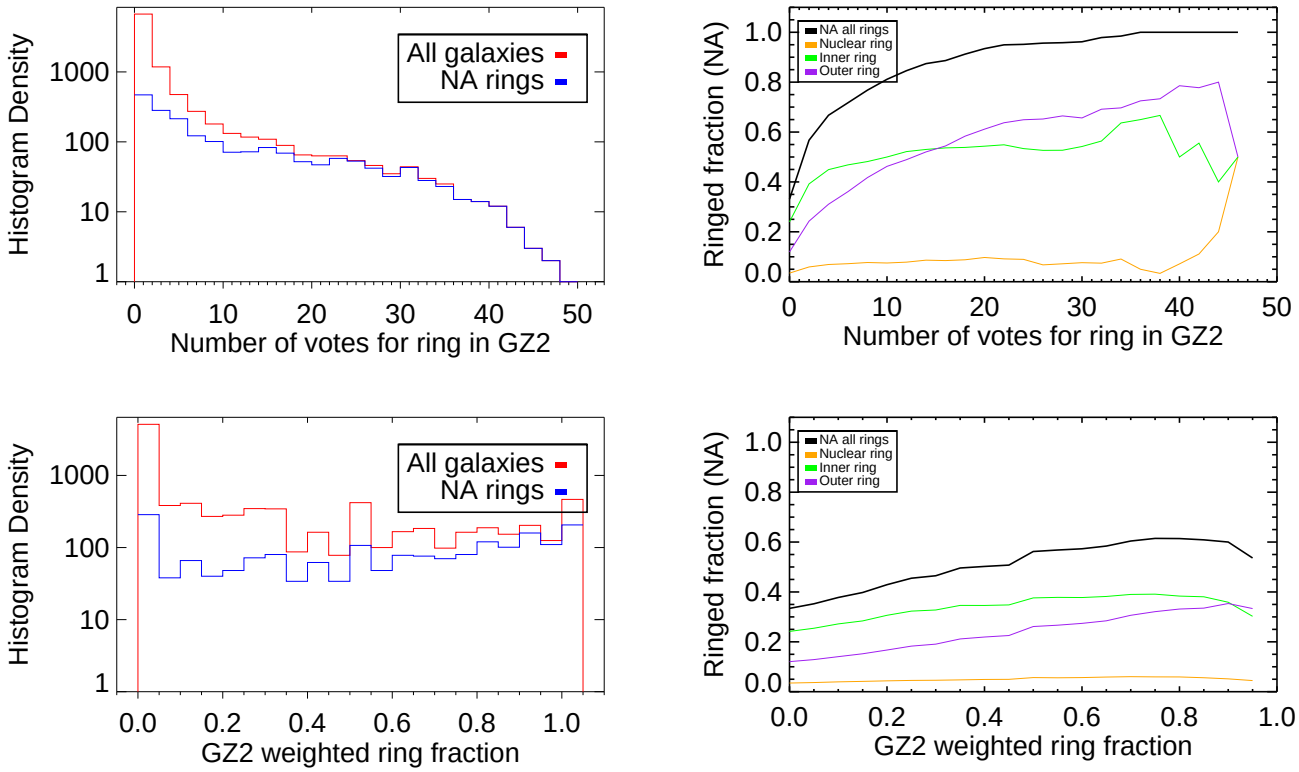

Figure 10. NA10 ring classifications compared to GZ2. Data are for the 9,746 galaxies in both samples which are face-on ($\log(a/b) < 0.3$).

bridges). If tidal tails are present, there is an additional flag couting the number of tails.

In the NA10 catalog, 22.3% of galaxies are labeled as paired; of these, the majority (72%) are close pairs. Interacting galaxies are a much smaller subset, comprising 7% of the NA10 sample. In the GZ2 catalog, only 3.6% (1.8%) of galaxies have a merger weighted fraction above 0.5 (0.8). If the vote totals for a possible merger are used instead, 5% (3%) of galaxies have at least 5 (10) votes for a merger. The large fraction of NA10 galaxies suggests that this is less likely to be a good proxy for a merger as identified by the GZ2 votes; interacting galaxies may be a more promising proxy.

Figure 11 shows the distributions of NA10 paired and interacting galaxies. Most galaxies in the overlapping sample have no votes for a merger; the same trends are seen for both pairs and interactions. If using the raw number of votes as a cutoff, GZ2 galaxies stabilize at an 80% match rate to the NA10 paired galaxies for 8 or more merger votes. The trend for interacting galaxies continues to increase with higher numbers of votes, only matching the 80% level at 25 or more merger votes, for which less than 100 galaxies are included.

Using the weighted merger fraction instead of the raw vote counts does not show a strong correlation with the NA10 classifications. Interestingly, both show peaks if a weighted merger fraction cutoff of 0.5 is chosen. The peak of the NA10 fraction, however, is significantly lower if matching on the weighted merger fraction; only 70% for paired galaxies and 40% for interacting galaxies. This shows no difference depending on the type of NA10 pair. Within the NA10 interacting classes, all actually show a decrease in number for a higher weighted merger fraction; the only exception is galaxies classified as bridges, the fraction of which increases slightly above a weighted merger fraction of 0.5.

Similar to the results above, there is no strong correlation between either the GZ2 weighted merger fraction or number of merger votes and the number of NA-identified tidal tails.

5.2.4 T-types

There has been no published discussion in the literature comparing large-scale morphologies of the NA10 and GZ catalogs. 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 12 shows the percentage of galaxies identified as having either a disk or features from the first question in the GZ2 tree, color-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. Late-type galaxies, including S0's, have a median weighted 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

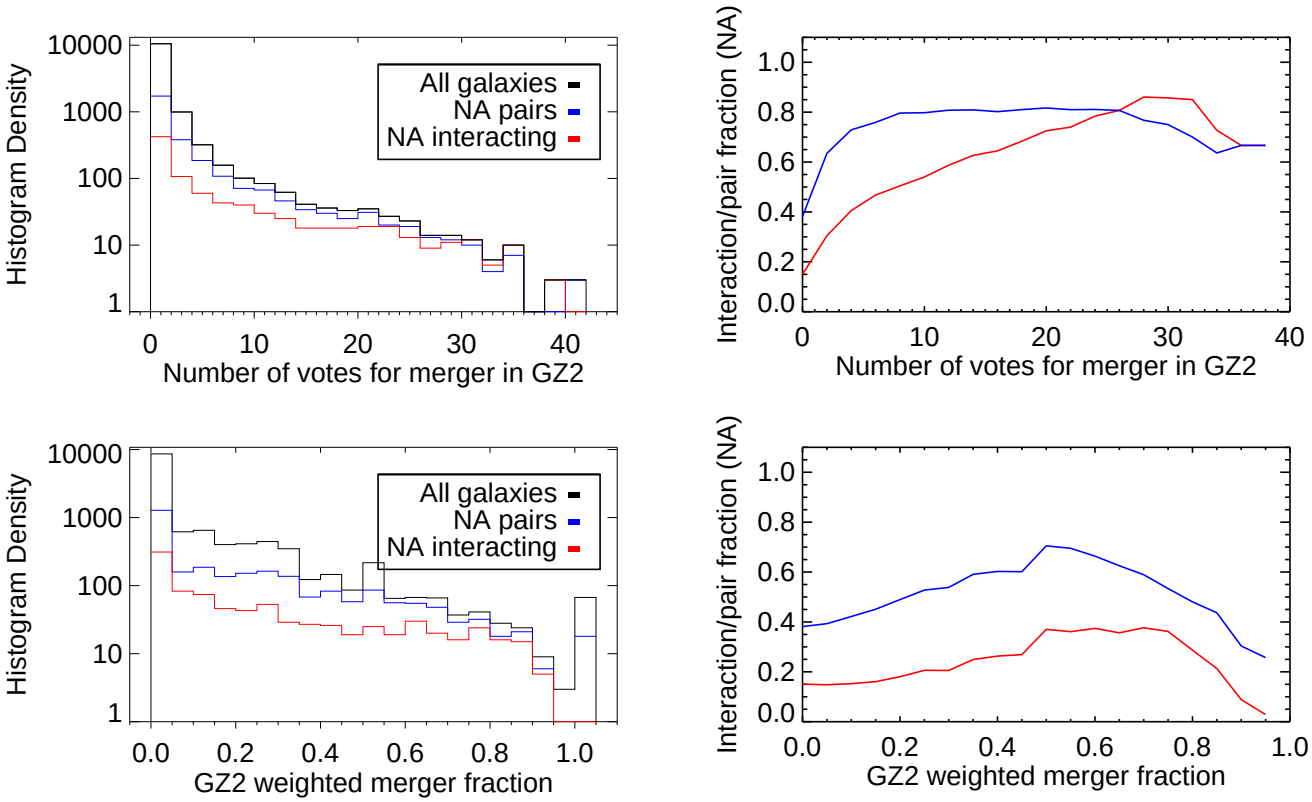

Figure 11. NA10 galaxies as a function of the votes/vote fraction for merger in GZ2. This includes all galaxies in the overlap sample (black), galaxies in pairs (blue), and interacting galaxies (red). Data are for the 12,480 galaxies found in both samples.

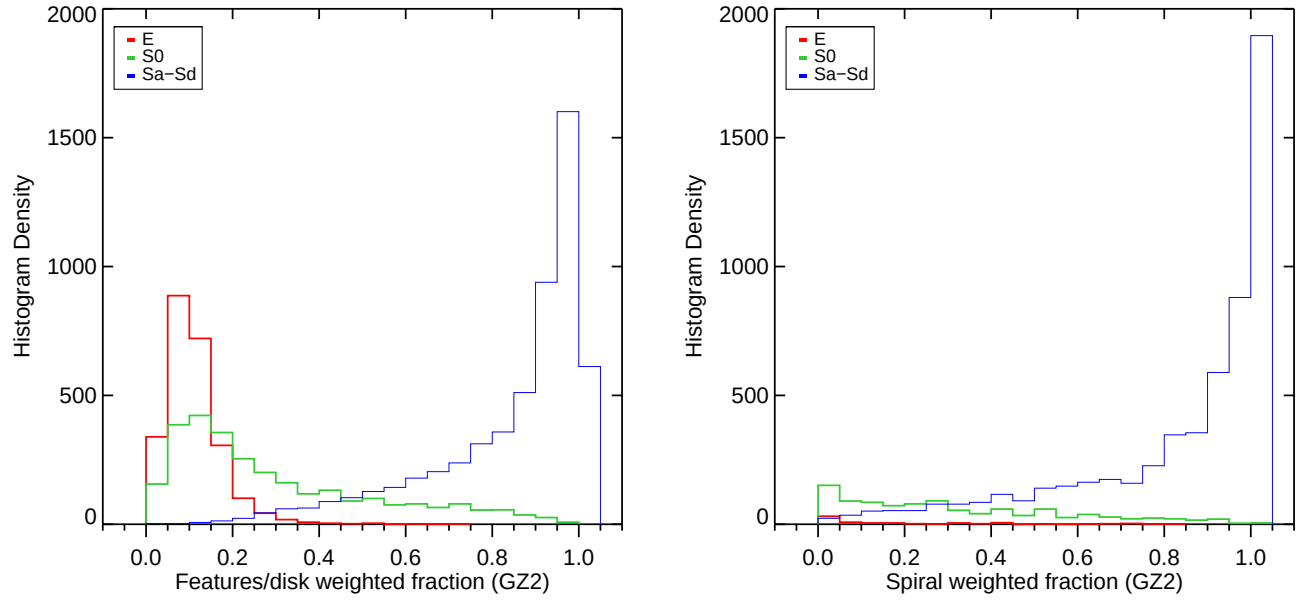

Figure 12. NA10 T-Type classifications compared to GZ2. Data on the left are for the 12,480 galaxies found in both samples; the right only shows the 5,683 galaxies with at least 10 responses to Task 04 (visible spiral structure) in GZ2.

the late-type galaxies into more specific Hubble classifications, the late-type 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 fractions disappears. The median GZ2 weighted fraction for these galaxies is 0.88, with a standard deviation of 0.23. The highest GZ2 weighted 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 weighted fraction for smooth galaxies is approximately $f_{smooth} = (1 - f_{features\_or\_disk})$. Elliptical galaxies (T-Type = -5) have a median weighted 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, $\sim 85\%$ of them have GZ2 weighted fractions of 0.0, with the remainder weakly clustered around 0.3. For NA10 late-type galaxies, the majority of disk/feature objects have high GZ2 spiral structure weighted fractions. For galaxies with at least 10 votes on Task 04 (a peak at 0.0 appears when this cut is not imposed), 70% of Sa or later-types have a GZ2 spiral vote above 0.8. This drops to 60% if S0 galaxies are included as late-type. The missing population is thus made up of galaxies with significant spiral structure by NA10, 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. Experimenting with other values for the GZ2 weighted fraction had no change on the results.

5.2.5 Spiral tightness

If a disk galaxy was identified as having spiral structure, Task 10 in GZ2 asked users to classify the "tightness' of the spiral arms. This task had three options: tight, medium, or loose (accompanied with pictograms on buttons that illustrated representative pitch angles). This has direct (but not exclusive) connections to the Hubble classification of late-type galaxies; tight spirals would be Sa/Sb, medium spirals Sb/Sc, and loose spirals Sc/Sd. The agreement between the GZ2 classification can be compared to Hubble types by using the NA10 classifications.

The left side of Figure 13 shows the distribution of NA10 T-Types for galaxies based on their GZ2 weighted fractions for winding arms. This figure shows only galaxies with at least 10 votes on spiral structure; looking at all galaxies in the overlapping sample disproportionally weights the 0.0 and 1.0 weighted fraction bins. Weighted fractions for both tight and medium winding arms are relatively normally distributed, with tight spirals peaking near 0.46 and medium spirals at 0.37. Strongly-classified loose spirals are much rarer, with 75% of galaxies having a weighted fraction of less than 0.2. Almost no elliptical galaxies from the NA10 catalog are included, although there are significant numbers of S0 galaxies.

For tight spirals, galaxies with the highest weighted fractions have more earlier-type spirals than galaxies with a low vote for tight spiral winding arms. For a tight spiral weighted fraction above 0.9, 85% of galaxies are Sb or earlier. Medium-wound spirals with high weighted fractions tend to be Sb and Sc galaxies – the proportion of both types increases as a function of medium-wound weighted fraction, and constitute 84% of galaxies when the weighted fraction is greater than 0.6. Galaxies strongly classified as medium-wound are rare, however, with only 23 galaxies having a weighted fraction above 0.8. Loose spirals are dominated by Sc and Sd galaxies at high weighted fraction values, comprising more than 50% of galaxies above a loose weighted fraction of 0.7.

Overall, we see a clear trend for looser GZ2 spiral arms to correspond with later spiral T-Types from NA10 classifications. High weighted vote fractions are mostly Sa/Sb galaxies for tight winding, Sb/Sc galaxies for medium winding, and Sc/Sd galaxies for loose winding. Individual GZ2 vote fractions, however, have significant diversity even at the highest bins, and do not reliably separate the morphologies on the level of the Hubble T-Types.

5.2.6 Bulge dominance

Disk galaxies in GZ2 are also classified by the a visible dominance of a bulge (Task 05), irrespective of whether spiral structure is also identified. This task has four options: "no bulge", "just noticeable", "obvious", and "dominant" (accompanied with pictograms that illustrated bulge sizes compared to face-on spiral arms). Similar to the arm tightness task, we analyze the relationships between GZ2 classification and Hubble types from NA10.

The left side of Figure 14 shows the distribution of NA10 T-Types for galaxies based on their GZ2 weighted fractions for winding arms. This figure shows only galaxies with at least 10 votes on bulge prominence. Weighted fractions for both the "no bulge" and "dominant" responses peak strongly near zero and tail off as the vote fraction increases. Responses to the middle options, "just noticeable" and "obvious", resemble normal distributions peaking near 0.5.

"No bulge" galaxies in GZ2 are dominated by Sc and Sd spirals for non-zero weighted fractions. For weighted fractions above 0.1, 81% of galaxies are Sc or later; this rises to 100% for weighted fractions higher than 0.6. "Just noticeable" galaxies show a smooth change in T-Type distribution; low weighted fractions are dominated by S0 and Sa galaxies, while high weighted fractions are Sb-Sd. "Obvious" bulge galaxies are almost a mirror image of the "just noticeable" votes; low weighted fractions are Sb–Sd galaxies, and high

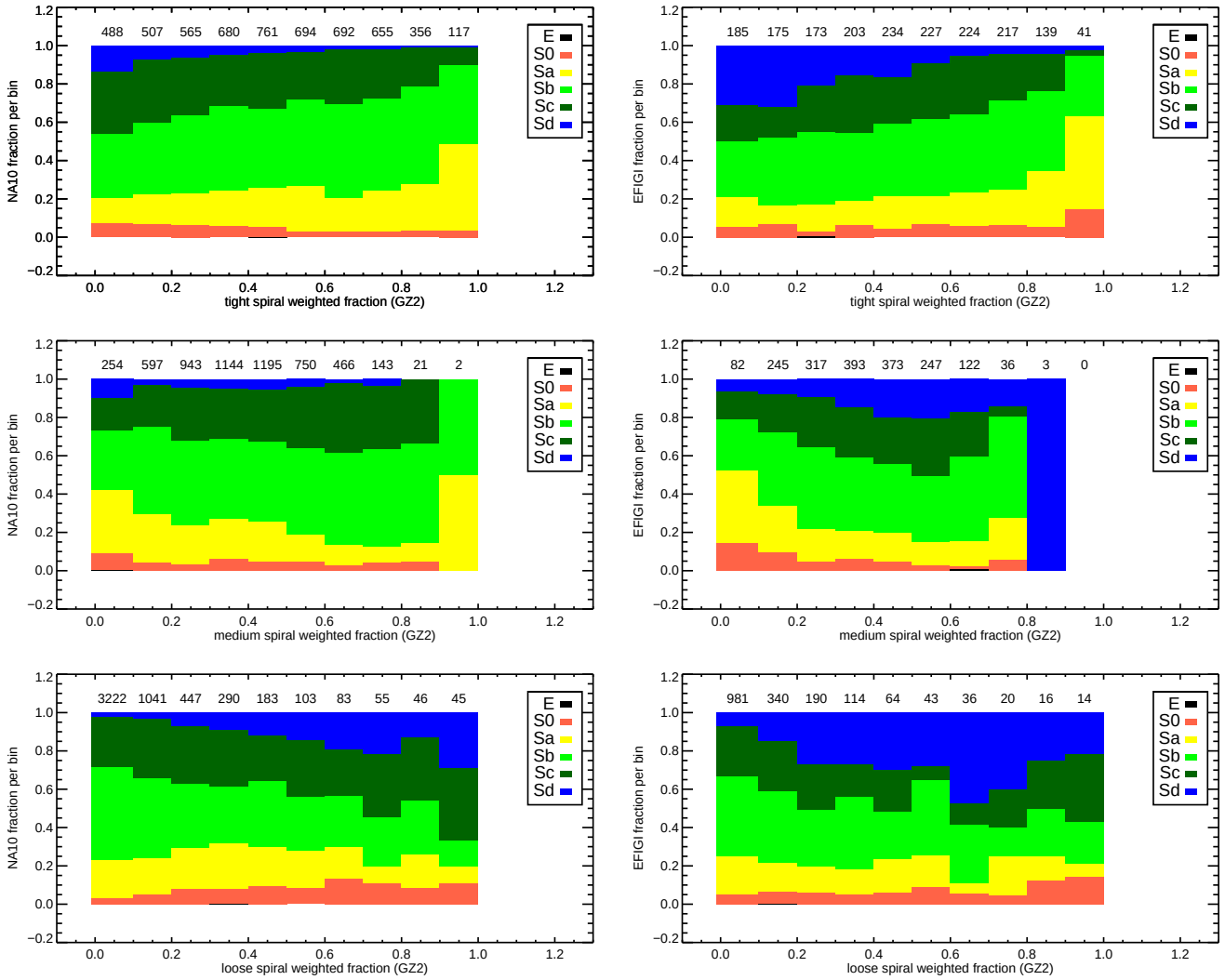

Figure 13. T-Type classifications compared to the GZ2 vote fractions for spiral tightness (Task 10). Left side is NA10 T-Types; right side is EFIGI T-Types. Data are for the 5,515 (NA10) and 1,907 (EFIGI) galaxies, respectively, with at least 10 GZ2 votes for Task 10. The number of galaxies per bin is indicated along the top of each panel.

weighted fractions are S0–Sa galaxies. Among galaxies classified as "dominant", less than 10 galaxies have weighted fractions above 0.6 (which are a surprisingly diverse mix of S0, Sa, and Sd). Most remaining galaxies have dominant weighted fractions of less than 0.1; the T-Types of the remaining galaxies between 0.1 and 0.6 mostly contain S0 and Sa spirals.

The link to T-Type is more sharply defined for bulge prominence than for spiral tightness, according to the NA10 classifications. Very clean samples of late-type (Sb–Sd) spirals can be selected using only the "no bulge" parameter; additional samples with $\sim$ 10% contamination can be selected with the "just noticeable" and "obvious" distributions. Early-type spirals and lenticulars at the same purity level can also be selected. Elliptical galaxies for which the bulge question was answered are most often "dominant", but there is no obvious separation of ellipticals from disk galaxies based on this task.

Since Hubble types are based on both bulge dominance and the distance of their spirals, we explore whether the combination of Tasks 05 and 10 from GZ2 improve the separation between T-Types. Figure 15 shows the GZ2 bulge prominence vs. spiral tightness, color-coded by their NA10 T-Types. The GZ2 variables are a weighted average of the votes for all possible task responses, where:

$$p_{bulge} = (0f_{nobulge} + 1f_{justnoticeable} + 2f_{obvious} + 3f_{dominant})/3., \quad (20)$$

and

$$p_{spiral} = (0 \times f_{tight} + 1 \times f_{medium} + 2 \times f_{loose})/2. \quad (21)$$

Using the weighted feature classifications shows a clear separation in T-Types; the majority of this, however, is for the bulge prominence category. Spiral prominence does not seem to significantly affect any of the classifications.

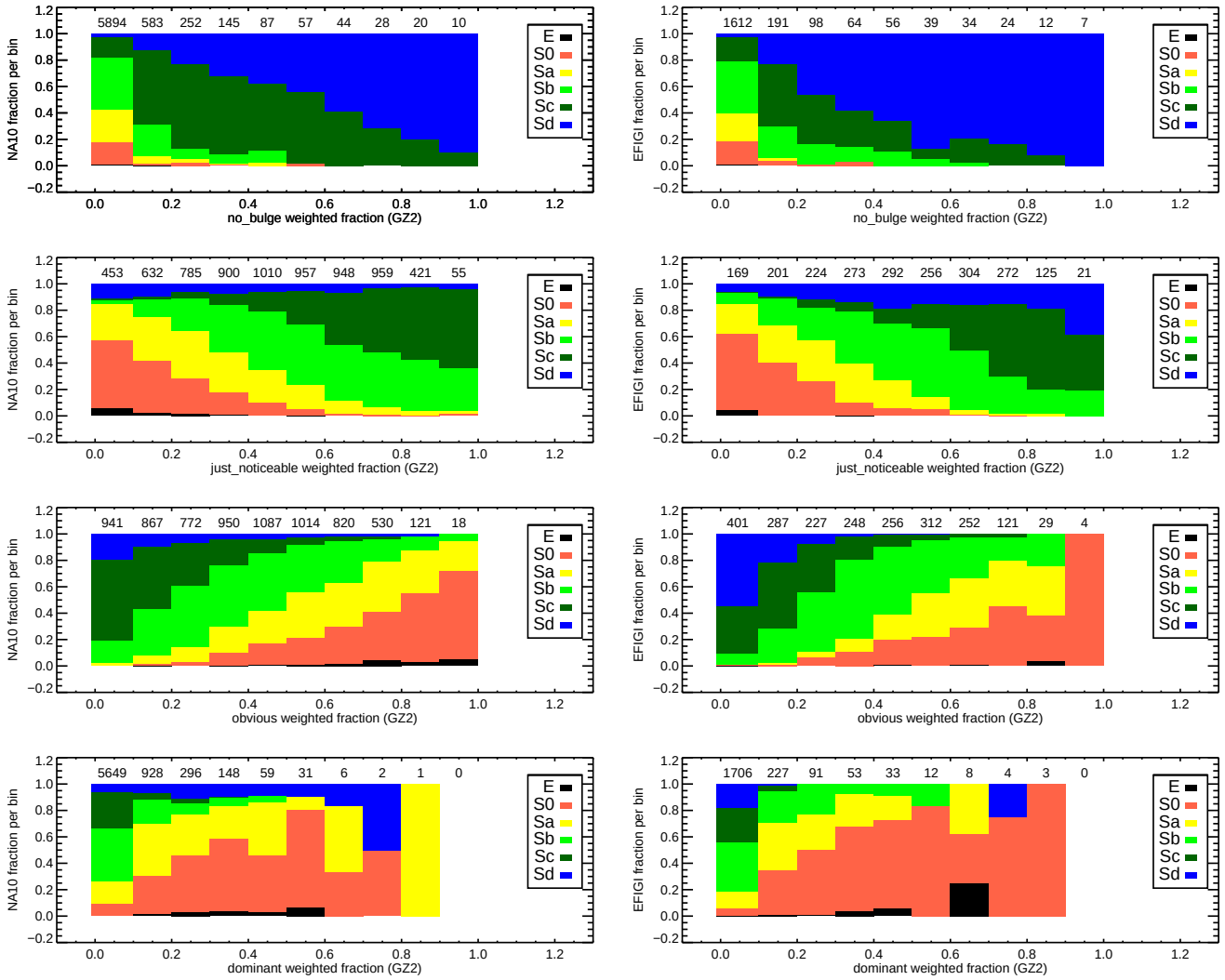

Figure 14. T-Type classifications compared to the GZ2 vote fractions for bulge prominence (Task 05). Left side is NA10 T-Types; right side is EFIGI T-Types. Data are for the 7,120 (NA10) and 2,321 (EFIGI) galaxies, respectively, with at least 10 GZ2 votes for Task 05. The number of galaxies per bin is indicated along the top of each panel.

5.2.7 T-Types to clicks

- How like a given T-Type is a galaxy?
- Take all GZ2 vote fractions, plus galaxy metadata (z, R_{50}, m_R, μ)
- Run a PCA on NA10 classifications to see which GZ2 parameters are most important
- Apply it and get the likelihood of a particular galaxy from the 20 axes
- This is the training set that is applied to the rest of Zoo2, accounting for evolution in the metadata properties on which the PCA can act.

5.3 EFIGI

Baillard et al. (2011) performed individual morphological classifications of 4,458 galaxies for EFIGI (Extractions de Formes Idéalisées de Galaxies en Imagerie). The sample is a subset of the RC3 catalog for which 5-color imaging in the SDSS DR4 was available. Images are supplemented by redshift information from several different sources. The galaxies have no strong redshift or volume limit on the sample, with almost all galaxies at $0.0001 < z < 0.08$. Classifications on composite *gri* images were performed by a group of 11 astronomers, each of whom classified a subset of 445 galaxies. A training set of 100 galaxies was classified by all astronomers in the group to adjust for individual bias.

EFIGI contains two types of morphological classification: T-Types and attributes. T-Types are assigned using a slightly modified version of the RC3 Hubble classifications. Peculiar galaxies are not considered a separate stage, and ellipticals are subdivided into various types: compact, elongated (standard elliptical), cD (giant elliptical), and dwarf spheroidals. They also classify late-type lenticulars ($S0^+$; T-Type=-1) that are not included in the classification of NA10. The remaining morphological information, called attributes, is divided into six groups:

- appearance: `inclination/elongation`
- environment: `multiplicity, contamination`

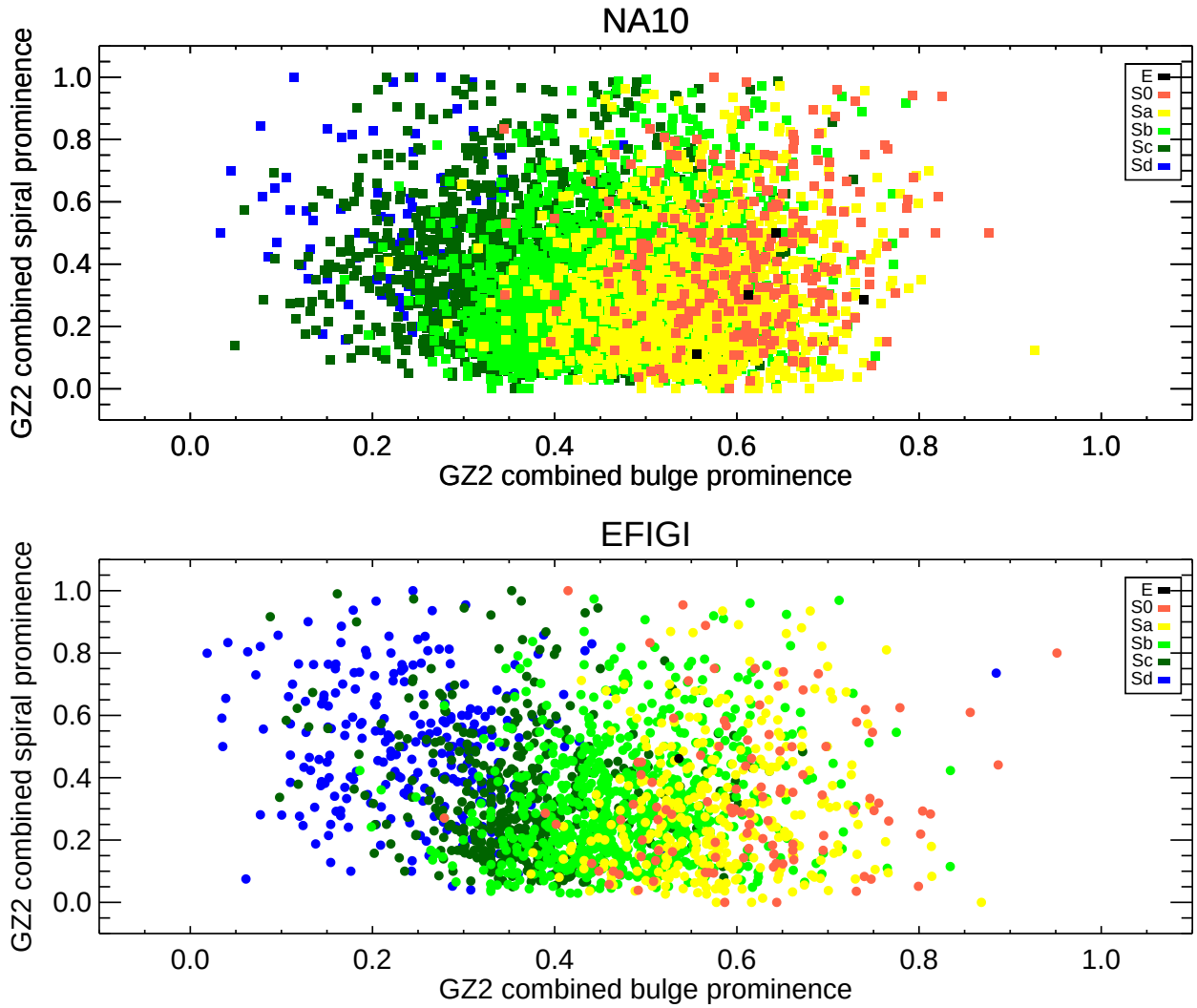

Figure 15. Weighted bulge classifications vs. weighted spiral classifications from GZ2. Galaxies are color-coded by their morphologies in EFIGI (*top*) and NA10 (*bottom*). Data are for galaxies with at least 10 classifications for both Task 05 (bulge dominance) and Task 10 (spiral structure) in GZ2.

- bulge: `B/T ratio`
- spiral arms: `arm strength, arm curvature, rotation`
- texture: `visible dust, dust dispersion, flocculence, hot spots`
- dynamics: `bar length, inner ring, outer ring, pseudo-ring, perturbation`

Attributes are defined on a five-step scale from 0 to 1 (0, 0.25, 0.50, 0.75, 1) that describe the strength of the feature in question. For some attributes (eg, arm strength, rings), the scale is set by the fraction of the flux contribution of the feature relative to that of the entire galaxy; this scale may not be linear. For others (eg, inclination or multiplicity), it ranges between the extrema of possible values. A 70% confidence interval (roughly 1σ) is estimated by setting lower and upper limits on the same five-point scale.

EFIGI is compared in detail to NA10 in Baillard et al. (2011). Only $\sim$ 10% of the NA10 catalog overlaps with EFIGI classifications; roughly one-third of the EFIGI sample lies at redshifts below the NA10 lower limit of $z = 0.01$, and also contain significant number of galaxies fainter than $g = 16$. T-Types agree well between the two samples; EFIGI lenticular and early spirals have slightly later average classifications in NA10, while later EFIGI galaxies have slightly earlier NA10 T-Types. EFIGI has a major fraction of galaxies with slight-to-moderate perturbations that have no interaction flags in the NA10 catalog, indicating that NA10 is less sensitive toward more benign features (eg, spiral arm asymmetry). The bar length scale is consistent between the two samples; good agreement is also found for ring classifications.

3,411 galaxies appear in both EFIGI and GZ2. This constitutes 77% of the EFIGI galaxies and 1.2% of the GZ2 sample.

5.3.1 Bars

GZ2 asks users to identify whether a bar is present in the galaxy. EFIGI's scale is based on bar length (not necessarily

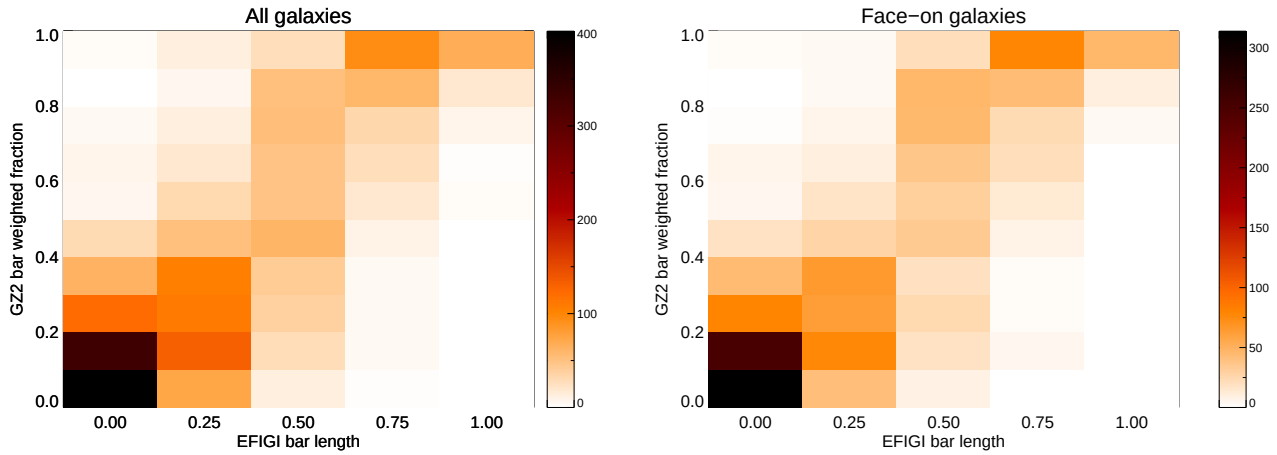

Figure 16. EFIGI bar length classifications compared to their GZ2 weighted fractions for the presence of a bar. Data on the left are for the 3,354 galaxies in both samples; the subset of 2,099 face-on galaxies is on the right. The dashed line is not a fit to the data, but is the one-to-one correlation.

corresponding to strength) with respect to D_{25}, the decimal logarithm of the mean isophote diameter at a surface brightness of $\mu_B = 25$ mag arcsec^{-2}. A value of 1.0 (the strongest bar) extends more than half the length of D_{25}, while the median value of 0.5 would be about one-third the length of D_{25}.

There is a strong correlation between the GZ2 weighted fractions for bars and the bar attribute strength from EFIGI (Figure 16). 65% of GZ2 galaxies in the overlap sample have no strong evidence for a bar (less than 0.3); of those, 77% had EFIGI bar attributes of 0.0 and 94% had 0.25 or less. For higher values of the GZ2 bar weighted fraction, the EFIGI attribute is slightly lower; the largest number of galaxies with GZ2 weighted fraction above 0.8 have EFIGI values of 0.75. The correlation coefficient between the variables is 0.51; if only face-on galaxies are considered, this increases to 0.75.

If the Masters et al. (2011) GZ2 weighted fraction of $\geqslant 0.5$, at least 10 bar votes, and face-on galaxies is applied, then 98% (646/660) galaxies overlapping with the EFIGI catalog have a bar attribute above 0. The mean EFIGI attribute (weighted by the confidence intervals) for galaxies barred by the Masters et al. (2011) criteria is 0.62, indicating a selection preference toward medium-length bars, between one-third and one-half of D_{25}.

The overall fraction of barred galaxies in EFIGI is 42% (1439/3354); this is essentially unchanged if only face-on galaxies are considered (915/2099 = 44%). This is significantly higher than the mean bar fraction of Masters et al. (2011), at 29.5%, but consistent with results using automated ellipse-fitting techniques (Barazza et al. 2008; Aguerri et al. 2009). The higher fraction in EFIGI is due to the contributions of galaxies with bar length attributes of 0.25, the majority of which have GZ2 weighted fractions below 0.5. If only EFIGI galaxies at 0.5 and above are considered to be barred, then the bar fraction falls to 17%. Only some of the galaxies in the 0.25 EFIGI bin are being classified by the GZ2 users as barred; since Baillard et al. (2011) defines this as a "barely visible" bar, it may be expected that less than half of GZ2 users would have detected it.

Masters et al. (2011) suggest in their paper that trends of bar fraction should ideally be discussed in the context of large or strong bars; an EFIGI cut of 0.5 or above would likely suit this philosophy best.

5.3.2 Arm curvature

EFIGI measures the arm curvature of each galaxy, with classifications very similar to the "tightness of spiral arms" question (Task 10) in GZ2. If both tasks and classifiers agree, one would expect galaxies with high GZ2 weighted fractions/votes for tight spirals to have EFIGI classifications at 0.75–1.0; GZ2 galaxies classified as medium spirals to be centered around 0.5; and loose spirals to have arm curvatures of 0.0–0.25.

The EFIGI arm curvature classifications broadly follow the trends expected from matching targets with GZ2. The tight spiral weighted fraction follows the trend of the EFIGI arm curvature (Figure 17). The Spearman's correlation coefficient for tight spirals is $\rho = 0.62$. The medium spiral weighted fraction is clustered in the middle of the EFIGI values, where galaxies with the highest GZ2 weighted fraction have EFIGI values of 0.25–0.50, with $\rho = -0.26$. Loose spirals shows an anti-correlation ($\rho = -0.54$); very few galaxies have GZ2 weighted fractions above 0.5, but those which do have low EFIGI arm curvature values (0.0–0.25).

Trends described above are quantitatively similar for both the full matching sample and for face-on galaxies only (right side of Figure 17). Somewhat surprisingly, the distribution also appears similar if only considering galaxies with a mininum number of GZ2 votes on spiral winding arms. Lower limits of 5, 10, 20, and 30 votes produce the same patterns for all three categories. The correlation coefficient for both tight and medium spirals does decrease to $|\rho < 0.2|$ if a 10-vote lower limit is applied.

Most useful: look at high weighted fractions and high numbers of vote counts for the spiral winding category of choice. That might correlate most strongly with EFIGI values.

EFIGI values are matched to specific pitch angles of the

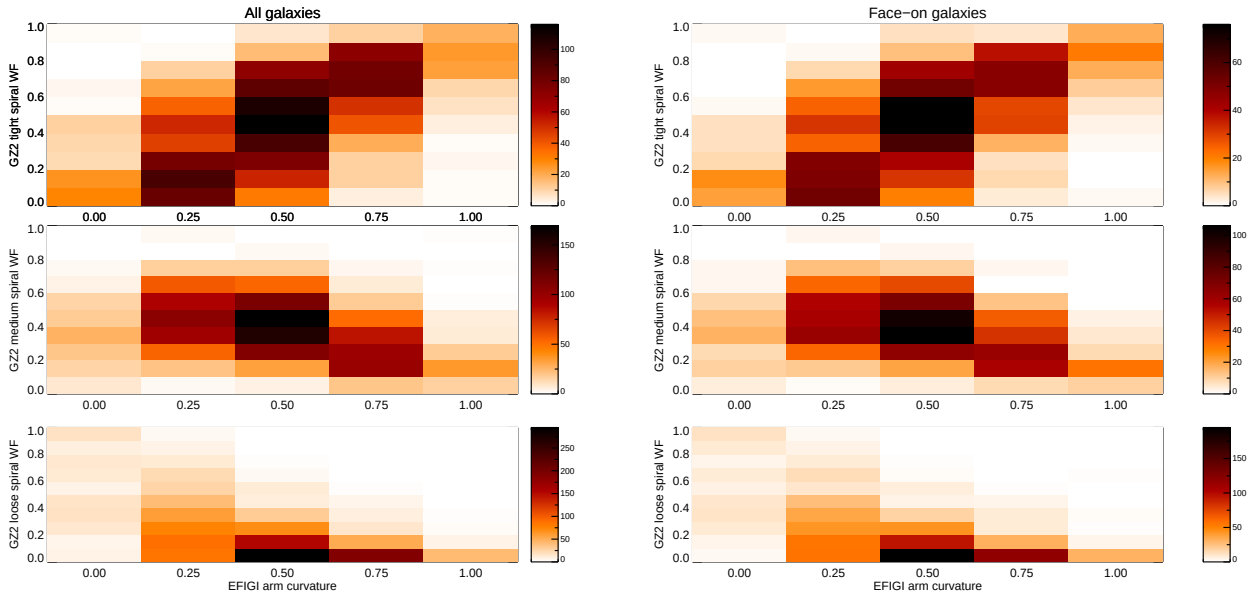

Figure 17. EFIGI arm curvature classifications compared to their GZ2 weighted fractions for the presence of a bar. Data on the left are for the 3,411 galaxies in both samples; the subset of 2,099 face-on galaxies is on the right. Dashed lines on the top and bottom pairs of plots show the one-to-one correlation, a Gaussian with $\mu = 0.5$ and $\sigma = 0.25$, and the one-to-one anti-correlation, respectively.

galaxy. For galaxies in which the GZ2 WF does not match the EFIGI, what is happening? Are the vote categories laterally skewed, are we not sensitive to weakest spirals, or is something else going on?

Not convinced that the face-on criterion is working.

Change bar plot to single-panel version. Change right side of arm curvature to cuts on count and weighted vote fraction, show stronger correlations. Discuss the percentage of the total number of galaxies that this constitutes for the expanded GZ2 sample, and how many "clean" galaxies from the EFIGI criteria we might be able to extract (subject to bias at lower surface brightnesses).

5.4 Huertas-Company – automated classifications

Huertas-Company et al. (2011) published a study in which they used training sets of galaxies to create artificial classifications, and then compared the results to GZ1. There have been no comparisons to GZ2; however, the broad nature of their probabilities (four broad morphological categories) are not suited for the fine structure questions such as bar and spiral number.

The classified sample of HC is the SDSS DR7 spectroscopic sample, limited to galaxies with $z < 0.25$ with good photometric data and clean spectra. Their total of 698,420 galaxies is approximately twice the size of the GZ2 sample, which has a similar redshift range. The HC sample goes to fainter magnitudes, with more than 400,000 galaxies below the GZ2 limit of $r > 17$. Their algorithm for classification is implemented using support vector machine software that tries to find boundaries between points in N-dimensional space. The training set for galaxy morphology is the 2,253 galaxies in Fukugita et al. (2007), classified by T-Type. Each galaxy is assigned a probability of being in one of four subclasses: E, S0, Sab, and Scd.

Huertas-Company et al. (2011) directly compare their results to the GZ1 sample from Lintott et al. (2011). They find that robust classifications in GZ1 (flagged in our clean sample as being either confirmed ellipticals or spirals) have median probabilities of 0.92 according to their algorithm, indicating that sure GZ1 classifications are also sure in their catalog. They also find a (but not strictly linear) relationship between the GZ1 debiased vote fraction and the HC probabilities. This is one of the first independent confirmations that the vote fractions may be related to the actual *probability* of a galaxy possessing a particular morphology.

HC also compare their results to the NA10 data, most of which are not included in the HC training sample. They find a good correlation between the NA10 T-Types and the HC probabilities, especially for ellipticals and Scd spirals. S0 galaxies are more difficult to separate; for the NA10 lenticulars, HC11 give only a 0.4 probability of S0, with 0.32 of elliptical and 0.2 of being Sab. Sab galaxies have an average probability of 0.55 being an Sab in HC, but also 0.15 of being S0 or Scd. The HC algorithm is thus very good at identifying galaxies at the extreme ends of the Hubble tuning fork, but have larger amounts of overlapping probabilities for intermediate states.

Since the GZ2 galaxies are a subset of the GZ1 sample, the results for Task 01 are expected to be similar to those described in Huertas-Company et al. (2011). Figure 18 shows the distributions of the HC early- and late-type probabilities for GZ2 galaxies robustly identified ($p > 0.8$) as either smooth or having features/disks. The median HC early-type probability for GZ2 ellipticals is 0.85, and the late-type probability for GZ2 spirals is 0.95. This confirms the result that

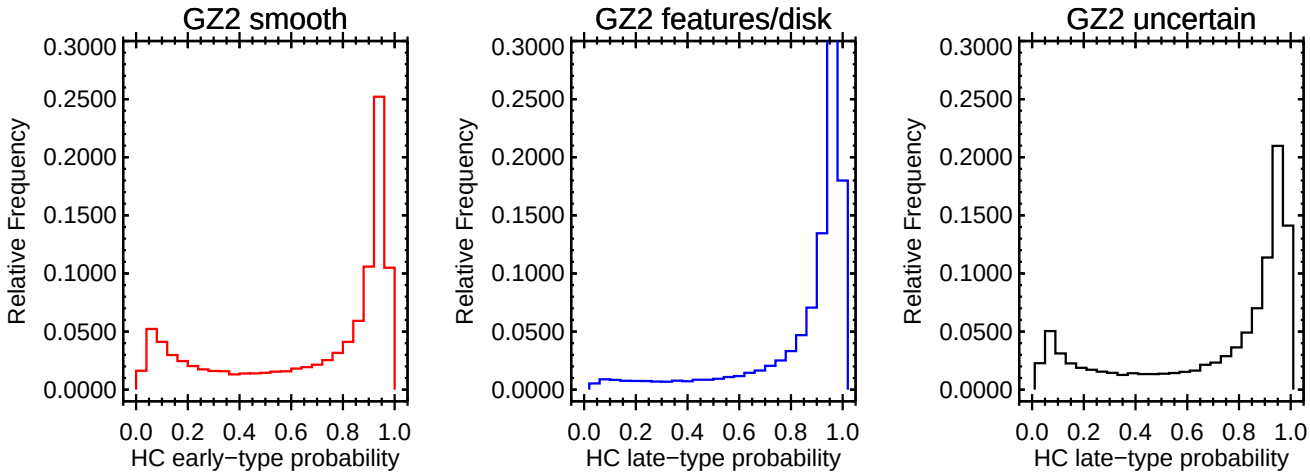

Figure 18. Left: distribution of HC early-type probabilities for galaxies with Task 01 "smooth" weighted fraction above 0.8. Middle: distribution of HC late-type probabilities for galaxies with Task 01 "features or disk" weighted fraction above 0.8. Right: HC late-type probabilities for uncertain galaxies ($p < 0.8$ for all responses to Task 01).

robust classifications in Galaxy Zoo agree with the automated algorithm.

An exception to this is a population of galaxies classified as "smooth" by GZ2 users, but which have very low early-type probabilities from HC (the bump on the left side of the first panel in Figure 18). The average GZ2 vote fraction for these galaxies is consistent with those with high early-type probabilities – these galaxies are not marginally classified as ellipticals in GZ2. The roundness of the galaxy (Task 07 in GZ2) seems to play some role, as the low-HC smooth galaxies have fewer round galaxies and many more "cigar-shaped" galaxies in this sample. A high axial ratio might have trained the HC algorithm to infer the existence of a disk; the absence of any obvious spiral features or bulge/disk separation (verified by eye in a small subsample of the imaages) lead GZ2 users to categorize them as "smooth". There is a clear dependence on apparent magnitude; the lower peak disappears if only galaxies with $r < 16$ are plotted. The lower peak is also significantly bluer than the higher peak, with respective colors of $(g - r) = 0.67$ and $(g - r) = 0.97$. Since the SVM method does include SDSS colors as a parameter, we conjecture that the low HC early-type probability is in part due to the fact that they are blue, in addition to morphological features such as shape and concentration. It would be interesting to see what fraction of the blue ellipticals (Schawinski et al. 2009) fall in this peak.

The right panel of Figure 18 shows the distribution of "unclassified" galaxies, for which none of the reponses for Task 01 had a weighted fraction above 0.8. The HC probability for these galaxies is bimodal, with the larger fraction classified as HC late-type and a smaller fraction as HC early-type.

Figure 19 plots the HC probability values against the GZ2 vote fraction for both smooth and feature/disk galaxies, similar to Figure 8 in Huertas-Company et al. (2011). Significant excesses are seen at the lower left and upper right corners for both samples, confirming the earlier result that robust classifications using both methods tend to agree. The relationship of the HC probability to GZ2 vote fraction, however, seems distinctly non-linear and significantly different from the relationship shown in Huertas-Company et al. (2011). Galaxies with low HC early-type probabilities show no strong correlation with the GZ2 weighted fraction, resulting in the vertical stripe in Figure 19. Intermediate values of the HC probabilities are typically classified by GZ2 users as "smooth", with a concentration at the highest HC early-type probabilities. It should be noted that without GZ2 debiasing, the two panels in Figure 19 are essentially mirror images of each other (since $p_{HC,early} + p_{HC,late} \equiv 1$ and the GZ2 weighted fractions only have marginal contributions from either the star/artifact response or downweighting of inconsistent users).

Splitting the HC morphology types into the E, S0, Sab, and Scd subclasses, Figure 20 shows the correlations to the GZ2 weighted fractions. While the direction of the correlation is the same as seen in Figures 9 and 10 in Huertas-Company et al. (2011), the behavior at the extrema is quite different. The debiased GZ1 data shows a strong cluster of low-HC and low-GZ probabilities for all four subclasses; no such cluster is seen in any of the plots in Figure 20. Furthermore, high HC probabilities have the lowest fraction of galaxies in the GZ1 published data, where our plots show a concentration of galaxies along the high end for both P(E) and P(S0). It should be investigated whether the scaling on these sets of figures is truly plotting the same value.

- To do: check how the presence of a bar affects the HC classifications.
- Plot the HC T-Type vs. the GZ2 bulge prominence (Figure 21); analyze.

5.5 Other classification schemes

Mention the volume classification of de Vaucouleurs (1959); recent developments by Kormendy & Bender (2012); Laurikainen et al. (2011); Cappellari et al. (2011); Krajnović et al. (2011) that have refined these ideas. Buta (2011) is a useful review of modern morphology.

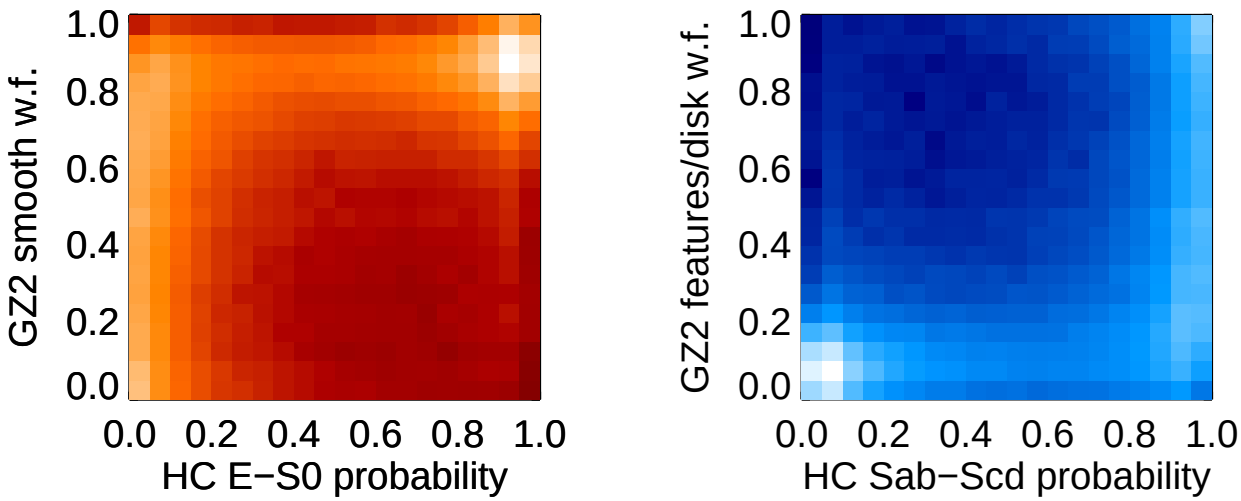

Figure 19. Left: GZ2 smooth weighted fraction as a function of Huertas-Company et al. (2011) early-type probability. Right: GZ2 features/disk weighted fraction as a function of HC late-type probability. Whiter values in both indicate a larger fraction of galaxies in that bin (logarithmic scale).

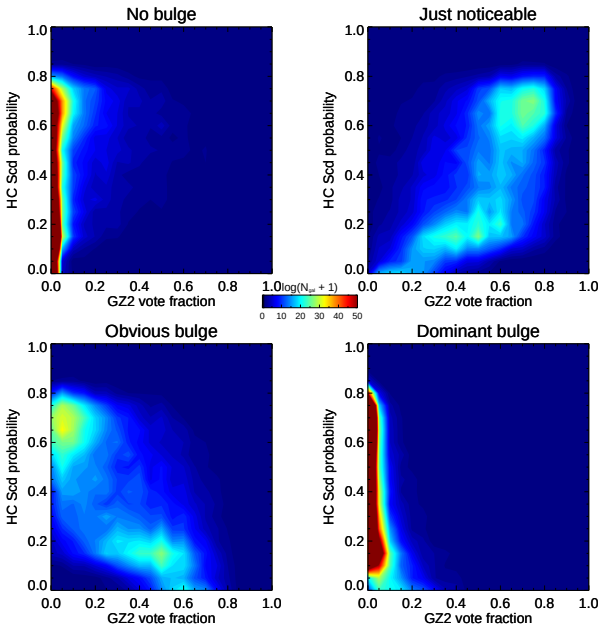

Figure 21. Huertas-Company et al. (2011) early-type probability as a function of the GZ2 weighted vote fraction for bulge dominance. The color of the contours is log $(N_{gal} + 1)$, where N_{gal} ranges from 0 to 1.5×10^3. Only galaxies with at least 10 classifications for Task 05 are shown.

6 GALAXY WARS

7 ASTRONOMY

8 CONCLUSIONS

REFERENCES

Abazajian K. N., Adelman-McCarthy J. K., Agüeros M. A., Allam S. S., Allende Prieto C., An D., Anderson K. S. J., Anderson S. F., Annis J., Bahcall N. A., et al. 2009, ApJS, 182, 543

Aguerri J. A. L., Méndez-Abreu J., Corsini E. M., 2009, A&A, 495, 491

Annis J., Soares-Santos M., Strauss M. A., Becker A. C., Dodelson S., Fan X., Gunn J. E., Hao J., Ivezic Z., Jester S., Jiang L., Johnston D. E., Kubo J. M., Lampeitl H., Lin H., Lupton R. H., Miknaitis G., Seo H.-J., Simet M., Yanny B., 2011, ArXiv e-prints

Baillard A., Bertin E., de Lapparent V., Fouqué P., Arnouts S., Mellier Y., Pelló R., Leborgne J.-F., Prugniel P., Makarov D., Makarova L., McCracken H. J., Bijaoui A., Tasca L., 2011, A&A, 532, A74

Bamford S. P., Nichol R. C., Baldry I. K., Land K., Lintott C. J., Schawinski K., Slosar A., Szalay A. S., Thomas D., Torki M., Andreescu D., Edmondson E. M., Miller C. J., Murray P., Raddick M. J., Vandenberg J., 2009, MNRAS, 393, 1324

Banerji M., Lahav O., Lintott C. J., Abdalla F. B., Schawinski K., Bamford S. P., Andreescu D., Murray P., Raddick M. J., Slosar A., Szalay A., Thomas D., Vandenberg J., 2010, MNRAS, 406, 342

Barazza F. D., Jogee S., Marinova I., 2008, ApJ, 675, 1194

Buta R., Combes F., 1996, FCP, 17, 95

Buta R. J., 2011, ArXiv e-prints

Cappellari M., Emsellem E., Krajnović D., McDermid R. M., Serra P., Alatalo K., Blitz L., Bois M., Bournaud F., et al. 2011, MNRAS, 416, 1680

Darg D. W., Kaviraj S., Lintott C. J., Schawinski K., Sarzi M., Bamford S., Silk J., Proctor R., Andreescu D., Murray P., Nichol R. C., Raddick M. J., Slosar A., Szalay A. S., Thomas D., Vandenberg J., 2010, MNRAS, 401, 1043

Davis D., Hayes W., 2013, ArXiv e-prints

de Vaucouleurs G., 1959, Handbuch der Physik, 53, 275

de Vaucouleurs G., de Vaucouleurs A., Corwin Jr. H. G., Buta R. J., Paturel G., Fouque P., 1991, Third Reference Catalogue of Bright Galaxies. Springer-Verlag

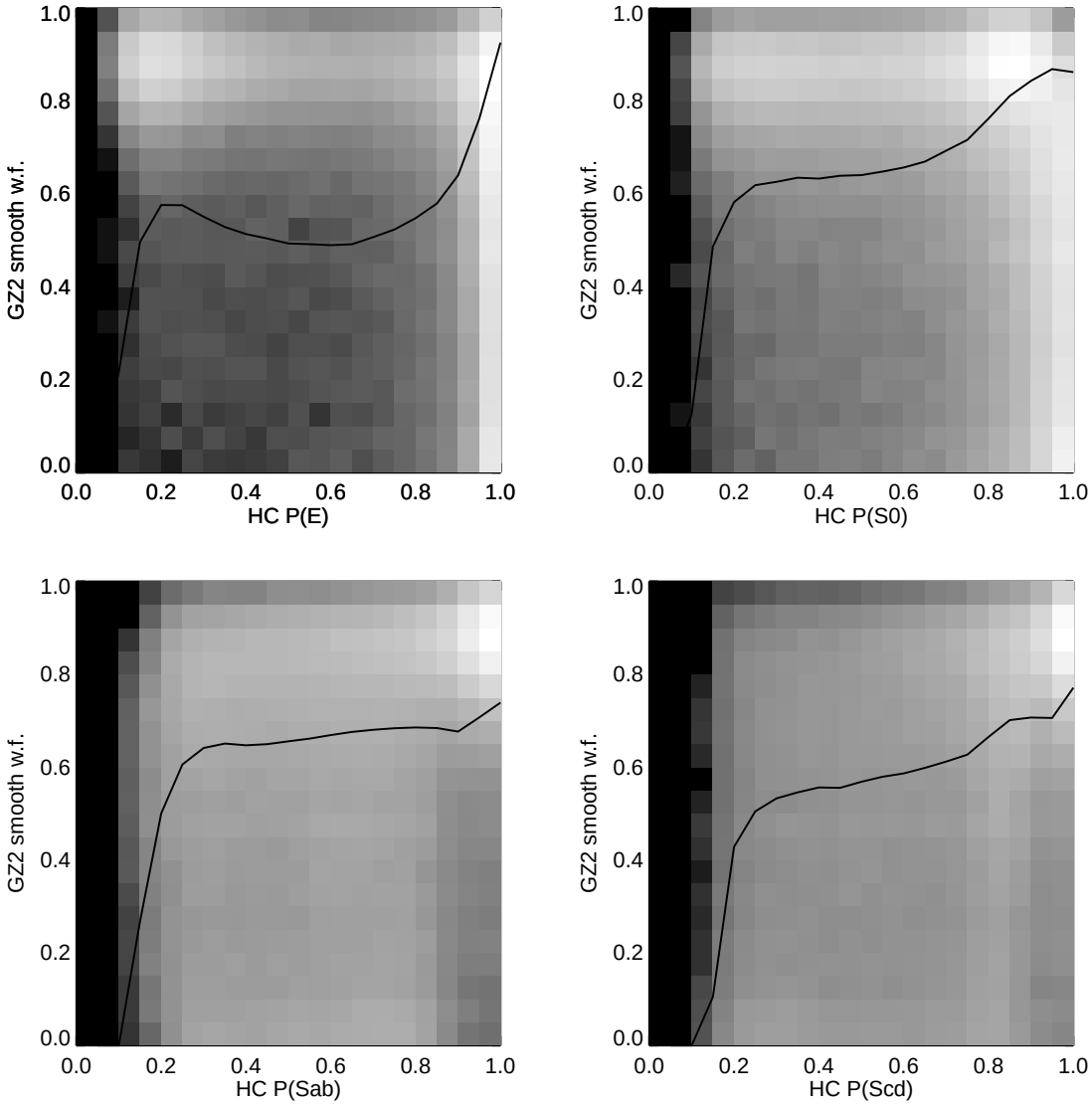

Figure 20. Left: GZ2 smooth weighted fraction as a function of Huertas-Company et al. (2011) early-type probability. Right: GZ2 features/disk weighted fraction as a function of HC late-type probability. Whiter values in both indicate a larger fraction of galaxies in that bin (logarithmic scale).

Fischer D., Schwamb M., Schawinski K., Lintott C., Brewer J., Giguere M., Lynn S., Parrish M., Sartori T., Simpson R., et al. 2011, ArXiv e-prints

Fukugita M., Nakamura O., Okamura S., Yasuda N., Barentine J. C., Brinkmann J., Gunn J. E., Harvanek M., Ichikawa T., Lupton R. H., Schneider D. P., Strauss M. A., York D. G., 2007, AJ, 134, 579

Hubble E. P., 1936, Realm of the Nebulae. Yale University Press

Huertas-Company M., Aguerri J. A. L., Bernardi M., Mei S., Sánchez Almeida J., 2011, A&A, 525, A157+

Kormendy J., Bender R., 2012, ApJS, 198, 2

Kormendy J., Kennicutt Jr. R. C., 2004, ARA&A, 42, 603

Krajnović D., Emsellem E., Cappellari M., Alatalo K., Blitz L., Bois M., Bournaud F., et al. 2011, MNRAS, 414, 2923

Laurikainen E., Salo H., Buta R., Knapen J. H., 2011, MNRAS, 418, 1452

Lintott C., Schawinski K., Bamford S., Slosar A., Land K., Thomas D., Edmondson E., Masters K., Nichol R. C., Raddick M. J., Szalay A., Andreescu D., Murray P., Vandenberg J., 2011, MNRAS, 410, 166

Lintott C. J., Schawinski K., Slosar A., Land K., Bamford S., Thomas D., Raddick M. J., Nichol R. C., Szalay A., Andreescu D., Murray P., Vandenberg J., 2008, MNRAS, 389, 1179

Martig M., Bournaud F., Croton D. J., Dekel A., Teyssier R., 2012, ApJ, 756, 26

Masters K. L., Nichol R. C., Hoyle B., Lintott C., Bamford S. P., Edmondson E. M., Fortson L., Keel W. C., Schawinski K., Smith A. M., Thomas D., 2011, MNRAS, 411, 2026

Nair P. B., Abraham R. G., 2010a, ApJS, 186, 427

Nair P. B., Abraham R. G., 2010b, ApJL, 714, L260

Nieto-Santisteban M. A., Szalay A. S., Gray J., 2004, in Ochsenbein F., Allen M. G., Egret D., eds, Astronomical Data Analysis Software and Systems (ADASS) XIII Vol. 314 of Astronomical Society of the Pacific Conference Series, ImgCutout, an Engine of Instantaneous Astronomical Discovery. p. 666

Sandage A., 1961, The Hubble atlas of galaxies. Carnegie Institute of Washington

Schawinski K., Lintott C., Thomas D., Sarzi M., Andreescu D., Bamford S. P., Kaviraj S., Khochfar S., Land K., Murray P., Nichol R. C., Raddick M. J., Slosar A., Szalay A., Vandenberg J., Yi S. K., 2009, MNRAS, 396, 818

Schwamb M. E., Lintott C. J., Fischer D. A., Giguere M. J., Lynn S., Smith A. M., Brewer J. M., Parrish M., Schawinski K., Simpson R. J., 2012, ApJ, 754, 129

Simpson R. J., Povich M. S., Kendrew S., Lintott C. J., Bressert E., Arvidsson K., Cyganowski C., Maddison S., Schawinski K., Sherman R., Smith A. M., Wolf-Chase G., 2012, MNRAS, 424, 2442

York D. G., Adelman J., Anderson Jr. J. E., Anderson S. F., Annis J., Bahcall N. A., Bakken J. A., Barkhouser R., Bastian S., Berman E., et al. 2000, AJ, 120, 1579

This paper has been typeset from a TEX/ LATEX file prepared by the author.

Table 3. Morphological classifications of GZ2 main sample galaxies with spectra

				t01_smooth_or_features_a01_smooth_						t01_smooth_or_features_a02_features_or_disk_						...
SDSS DR7 objID	sample	N_{class}	N_{votes}	count	wt_count	fraction	wt_fraction	debiased	flag	count	weight	fraction	wt_fraction	debiased	flag	
588017703996096547	original	44	349	1	0.1	0.023	0.002	0.002	0	42	42.0	0.955	0.975	0.975	1	
587738569780428805	original	45	185	5	5.0	0.111	0.115	0.115	0	38	38.0	0.844	0.873	0.873	1	
587735695913320507	original	46	372	0	0.0	0.000	0.000	0.000	0	44	44.0	0.957	0.966	0.966	1	
587742775634624545	original	45	289	8	8.0	0.178	0.178	0.178	0	37	37.0	0.822	0.822	0.822	1	
587732769983889439	extra	49	210	12	12.0	0.245	0.249	0.249	0	36	36.0	0.735	0.748	0.748	0	
588017725475782665	extra	42	149	27	27.0	0.643	0.686	0.686	0	12	12.0	0.286	0.305	0.305	0	
588017702391578633	original	45	356	0	0.0	0.000	0.000	0.000	0	45	45.0	1.000	1.000	1.000	1	
588297864730181658	original	45	206	4	4.0	0.089	0.091	0.091	0	39	38.5	0.867	0.871	0.871	1	
588017704545812500	original	43	360	0	0.0	0.000	0.000	0.000	0	43	43.0	1.000	1.000	1.000	1	
588017566564155399	extra	43	244	6	6.0	0.140	0.143	0.143	0	35	35.0	0.814	0.833	0.833	1	

Note. — The full, machine-readable version of this table is available at http://data.galaxyzoo.org. A portion is shown here for guidance on form and content. The full table contains 252,750 rows (one for every galaxy in the sample), and 226 columns, with six variables for each of the 37 GZ2 morphology classifications.

Table 4. Morphological classifications of GZ2 main sample galaxies with photo-z

				t01_smooth_or_features_a01_smooth_						t01_smooth_or_features_a02_features_or_disk_						...
SDSS DR7 objid	sample	N_{class}	N_{votes}	count	wt_count	fraction	wt_fraction	debiased	flag	count	weight	fraction	wt_fraction	debiased	flag	
587722981736579107	original	43	181	27	27.0	0.628	0.648	0.648	0	14	14.0	0.326	0.336	0.336	0	
587722981741691055	original	44	133	40	40.0	0.909	0.909	0.909	1	1	1.0	0.023	0.023	0.023	0	
587722981745819655	original	46	221	17	17.0	0.370	0.378	0.378	0	22	22.0	0.478	0.489	0.489	0	
587722981746082020	original	44	172	31	31.0	0.705	0.771	0.771	0	6	6.0	0.136	0.149	0.149	0	
587722981746344092	original	43	358	0	0.0	0.000	0.000	0.000	0	43	43.0	1.000	1.000	1.000	1	
587722981747982511	original	45	156	37	37.0	0.822	0.850	0.633	0	3	3.0	0.067	0.069	0.289	0	
587722981748375814	original	52	198	44	44.0	0.846	0.846	0.846	1	8	8.0	0.154	0.154	0.154	0	
587722981748768914	original	46	350	3	3.0	0.065	0.065	0.065	0	43	43.0	0.935	0.935	0.935	1	
587722981748768984	original	42	140	37	36.2	0.881	0.900	0.872	1	4	4.0	0.095	0.100	0.128	0	
587722981749031027	original	50	158	46	45.8	0.920	0.932	0.932	1	2	1.4	0.040	0.028	0.028	0	

Note. — The full, machine-readable version of this table is available at http://data.galaxyzoo.org. A portion is shown here for guidance on form and content, which are identical to those in Table 3.

Table 5. GZ2 weighted vote fractions for the main and Stripe 82 normal-depth samples

User task and answer	Mean weighted vote fraction		Number of galaxies		Sample comparison	
	Main	Stripe 82	Main	Stripe 82	$f_m - f_s$	f_m / f_s
T01_SMOOTH_OR_FEATURES_A01_SMOOTH	0.642	0.613	273783	8437	0.029	1.048
T01_SMOOTH_OR_FEATURES_A02_FEATURES_OR_DISK	0.325	0.352	273783	8437	-0.027	0.924
T01_SMOOTH_OR_FEATURES_A03_STAR_OR_ARTIFACT	0.033	0.036	273783	8437	-0.003	0.924
T02_EDGEON_A04_YES	0.234	0.229	118147	4110	0.005	1.020
T02_EDGEON_A05_NO	0.766	0.771	118147	4110	-0.005	0.994
T03_BAR_A06_BAR	0.287	0.292	87647	3077	-0.004	0.986
T03_BAR_A07_NO_BAR	0.713	0.708	87647	3077	0.004	1.006
T04_SPIRAL_A08_SPIRAL	0.660	0.674	87621	3077	-0.014	0.980
T04_SPIRAL_A09_NO_SPIRAL	0.340	0.326	87621	3077	0.014	1.042
T05_BULGE_PROMINENCE_A10_NO_BULGE	0.129	0.114	87608	3076	0.015	1.129
T05_BULGE_PROMINENCE_A11_JUST_NOTICEABLE	0.479	0.491	87608	3076	-0.012	0.976
T05_BULGE_PROMINENCE_A12_OBVIOUS	0.321	0.333	87608	3076	-0.011	0.966
T05_BULGE_PROMINENCE_A13_DOMINANT	0.070	0.062	87608	3076	0.008	1.134
T06_ODD_A14_YES	0.199	0.185	273636	8391	0.014	1.073
T06_ODD_A15_NO	0.801	0.815	273636	8391	-0.014	0.983
T07_ROUNDED_A16_COMPLETELY_ROUND	0.330	0.326	229974	6956	0.005	1.014
T07_ROUNDED_A17_IN_BETWEEN	0.496	0.501	229974	6956	-0.004	0.991
T07_ROUNDED_A18_CIGAR_SHAPED	0.173	0.174	229974	6956	-0.000	0.999
T08_ODD_FEATURE_A19_RING	0.147	0.178	72439	2114	-0.031	0.827
T08_ODD_FEATURE_A20_LENS_OR_ARC	0.045	0.040	72439	2114	0.005	1.122
T08_ODD_FEATURE_A21_DISTURBED	0.121	0.125	72439	2114	-0.003	0.973
T08_ODD_FEATURE_A22_IRREGULAR	0.177	0.170	72439	2114	0.007	1.039
T08_ODD_FEATURE_A23_OTHER	0.282	0.245	72439	2114	0.037	1.152
T08_ODD_FEATURE_A24_MERGER	0.212	0.209	72439	2114	0.003	1.015
T08_ODD_FEATURE_A38_DUST_LANE	0.016	0.034	72439	2114	-0.018	0.475
T09_BULGE_SHAPE_A25_ROUNDED	0.538	0.567	21496	746	-0.029	0.948
T09_BULGE_SHAPE_A26_BOXY	0.103	0.096	21496	746	0.007	1.077
T09_BULGE_SHAPE_A27_NO_BULGE	0.359	0.338	21496	746	0.022	1.065
T10_ARMS_WINDING_A28_TIGHT	0.370	0.392	55800	2052	-0.022	0.943
T10_ARMS_WINDING_A29_MEDIUM	0.417	0.405	55800	2052	0.011	1.028
T10_ARMS_WINDING_A30_LOOSE	0.214	0.203	55800	2052	0.011	1.054
T11_ARMS_NUMBER_A31_1	0.074	0.067	55805	2053	0.008	1.117
T11_ARMS_NUMBER_A32_2	0.501	0.478	55805	2053	0.023	1.048
T11_ARMS_NUMBER_A33_3	0.093	0.098	55805	2053	-0.005	0.950
T11_ARMS_NUMBER_A34_4	0.038	0.043	55805	2053	-0.005	0.884
T11_ARMS_NUMBER_A36_MORE_THAN_4	0.030	0.033	55805	2053	-0.002	0.934
T11_ARMS_NUMBER_A37_CANT_TELL	0.263	0.281	55805	2053	-0.019	0.934

Note. — Data for each task is only for galaxies with at least 10 total responses to the classification question.