Mon. Not. R. Astron. Soc. **000**, 000–000 (0000) Printed 7 October 2014 (MN LATEX style file v2.2)

The impact of cosmic variance on simulating weak lensing surveys

Arun Kannawadi[1⋆], Rachel Mandelbaum[1], Claire Lackner[2]
[1] *McWilliams Center for Cosmology, Carnegie Mellon University, Pittsburgh, PA 15217, USA*
[2] *Kavli Institute for the Physics and Mathematics of the Universe (WPI), Todai Institutes for Advanced Study, the University of Tokyo, Kashiwa, Japan.*

7 October 2014

ABSTRACT
Upcoming weak lensing surveys will survey large cosmological volumes in order to measure the growth of cosmological structure with time and thereby constrain dark energy. One major systematic uncertainty in this process is the calibration of the weak lensing shape distortions, or shears. Nearly all upcoming surveys plan to test their shear estimation algorithms using sophisticated image simulations that attempt to mimick real data as closely as possible, including realistic galaxy populations based on high-resolution data from the Hubble Space Telescope (*HST*). However, the existing datasets from the *HST* cover very small cosmological volumes, such that cosmic variance could cause the galaxy population represented in them to be atypical. This is particularly the case when selecting galaxy populations in limited redshift slices, which could be dominated by a single large overdensity or underdensity. In that case, the morphology-density relation could cause changes in the local galaxy populations that would result in incorrect conclusions to be drawn about the calibration of shear estimates as a function of redshift. We directly test this scenario using the COSMOS survey, which is the largest-area *HST* survey to date, and show how the statistical properties of shape (axis ratio or ellipticity) and morphological parameters such as Sérsic n or the bulge-to-total ratio are influenced by redshift-dependent cosmic variance in this survey. This study requires a careful distinction between environment effects from large-scale structure, which we do not wish to include in simulations, and general trends in the galaxy population with redshift. We conclude that this effect is large enough to exceed the systematic error budget in future surveys, but can be mitigated with careful choice of training dataset and sufficiently large redshift binning.

Key words: Gravitational lensing: weak — Cosmology: Large-scale structure of Universe — Galaxies: evolution.

1 INTRODUCTION

Weak gravitational lensing, the deflection of light by mass, is one of the cleanest ways to study the nature of dark energy by tracking the growth of structure in the Universe as a function of time (e.g., Bartelmann & Schneider 2001; Albrecht et al. 2006; Weinberg et al. 2013). As light from background sources passes by matter (including dark matter) on its way to us, the apparent shapes of the background galaxies get distorted, and the galaxies get slightly magnified as well. Because of its sensitivity to dark matter and dark energy, major surveys such as the Hyper Suprime-Cam (HSC; Miyazaki et al. 2006), Dark Energy Survey (DES; The Dark Energy Survey Collaboration 2005), the KIlo-Degree Survey (KIDS; de Jong et al. 2013), the Panoramic Survey Telescope and Rapid Response System (PanSTARRS; Kaiser et al. 2010), the Large Synoptic Survey Telescope (LSST; LSST Science Collaboration et al. 2009), Euclid[1] (Laureijs et al. 2011), and Wide-Field Infrared Survey Telescope (WFIRST; Green et al. 2012) are planned for the next two decades to gather enormous quantities of weak lensing data that will lead to precise constraints on the growth of structure with time, and therefore cosmological parameters.

For the upcoming surveys to achieve their promise, their systematic error budgets must be below their statistical error budgets. Systematic error budgets for weak lensing surveys typically include astrophysical effects, such as intrinsic

⋆ akannawa@andrew.cmu.edu

[1] http://sci.esa.int/euclid/, http://www.euclid-ec.org

alignments of galaxy shapes with large scale density fields (e.g., Troxel & Ishak 2014) and the effect of baryons on the matter power spectrum (e.g., van Daalen et al. 2011), as well as observational uncertainties such as the ability to robustly infer shears from galaxy observed shapes or photometric redshifts from their observed colors. Given the expected sub-per cent errors on upcoming surveys, systematic errors must be reduced from their typical level in the current state-of-the-art measurements that typically achieve ~ 5 per cent statistical errors at best (e.g., Schrabback et al. 2010; Heymans et al. 2013; Jee et al. 2013; Mandelbaum et al. 2013).

One method that is commonly used to test for the presence of systematic errors in the shear estimation process is image simulation, where we can cleanly test whether our methods of shear estimation recover the ground truth. This is a valuable test, considering the numerous sources of additive and multiplicative bias such as a mismatch between galaxy model assumptions and actual galaxy light profiles (e.g., Voigt & Bridle 2010; Melchior et al. 2010), biases due to the effects of pixel noise on the shear estimates (Kacprzak et al. 2012; Melchior & Viola 2012; Refregier et al. 2012), and ellipticity gradients (Bernstein 2010). These biases often differ for galaxies with different morphologies (e.g., disks vs. ellipticals), sizes, S/N, and shape (Bridle et al. 2010; Kitching et al. 2012). A general requirement for simulations used to test shear recovery is that they should be as realistic as possible.

Realistic simulations may use samples based on images from the Hubble Space Telescope (*HST*). Software packages like GalSim[2] (Rowe et al. 2014) can generate images of galaxies from the *HST* as they would appear with an additional lensing shear and viewed by some lower resolution telescope. Examples of training samples from the *HST* include the COSMOS survey (used by the GREAT3 challenge, Mandelbaum et al. 2014) or the Ultra Deep Field (UDF, used by Jee et al. 2013). These two examples serve as the extremes in the *HST* samples used as the basis for image simulation, with COSMOS being shallower but representing the current widest contiguous area surveyed by the *HST*, and the latter being extremely deep but narrow.

For a variety of physical reasons, some of which are still not fully understood, the shape and morphology of galaxies depends on their local environment (e.g., Carollo et al. 2014; De Propris et al. 2014). Hence, local overdensities or underdensities observed in these *HST* fields may (given the small size of the field) cause the properties of the galaxy population in redshift slices to be atypical depending on the environment in that slice. This would have the undesired consequence of including variation in galaxy properties due to the COSMOS (or other) survey cosmic variance in the simulated galaxy sample in that redshift slice, rather than only including ensemble effects that would appear in a large cosmological volume, such as true redshift evolution of galaxy properties. Our goal is to quantify the degree to which the morphology-density correlations in COSMOS cause noticeable changes in the galaxy populations in narrow redshift slices at a level that could result in difficulty using the sample to derive redshift-dependent shear calibrations. Upcoming surveys will study lensing as a function of redshift and therefore need to simulate galaxy samples at different redshifts in order to assess the shear calibration at each redshift.

The paper is structured as follows: in Sec. 2, we describe the data that we use for this study. In Sec. 3, we describe our methods for deriving the relevant galaxy properties like environment, morphology, and shape. Using these ingredients, we present our results in Sec. 4 and discuss their implications in Sec. 5.

2 DATA

The COSMOS survey (Scoville et al. 2007; Koekemoer et al. 2007; Leauthaud et al. 2007) is a flux-limited, narrow deep field survey covering a contiguous area of 1.64 deg^2 of sky, with images taken using the Advanced Camera for Surveys (ACS) Wide Field Channel (WFC) in the Hubble Space Telescope (HST). We use the COSMOS survey to define a parent sample of galaxy images to be used for making image simulations, following the approach taken to this problem in Mandelbaum et al. (2012, 2014).

We apply the following set of initial cuts to the COSMOS data, the first two of which are motivated and explained in more detail by Leauthaud et al. (2007):

(i) `MU_CLASS=1`: This criterion uses a comparison between the peak surface brightness and the background level to achieve a robust star/galaxy separation, with galaxies having `MU_CLASS=1`.

(ii) `CLEAN=1`: Objects near bright stars or those containing saturated pixels were removed; the rest pass this cut on `CLEAN`.

(iii) `GOOD_ZPHOT_SOURCE =1`: This cut requires that photometric redshifts be reliable and good enough to draw conclusions about the population (see Mandelbaum et al. 2012 for details).

High resolution images taken through the wide F814W filter (broad I) for all galaxies passing the above cuts were used to create a collection of postage stamp images for the GREAT3 challenge (Mandelbaum et al. 2014), using the procedure described in Mandelbaum et al. (2012). Each galaxy postage stamp image has a corresponding PSF image that can be used by GalSim or other software to remove the effects of the *HST* PSF before simulating the galaxy image as it would appear at lower resolution.

To better characterize the galaxy population, parametric models were fit to the light profiles of these galaxies. These were carried out using the method described in Lackner & Gunn (2012), and include Sérsic profile fits and 2 component bulge + disk fits described in detail in Mandelbaum et al. (2014) and briefly in Sec. 3.3 of this work.

In addition to the ACS/WFC (F814W) imaging, the COSMOS field has also been imaged by Subaru Suprime-Cam, the Canada-French Hawaii Telescope (CFHT) and KPNO/CTIO, yielding many bands of imaging data from which to determine high-fidelity photometric redshifts. Photometric redshifts were determined by Ilbert et al. (2009). The accuracy of photometric redshifts for $m_{\rm F814W} \leqslant 22.5$ is $\sigma_{\Delta z} = 0.007(1+z)$; for $m_{\rm F814W} \leqslant 24$, $\sigma_{\Delta z} = 0.012(1+z)$.

[2] `https://github.com/GalSim-developers/GalSim`

The photometric redshift values become noisier beyond $z \sim 1.2$, and the fits to the galaxy light profiles are also somewhat noisy once we go beyond $m_{\rm F814W} \sim 23.5$. For this reason, we will exclude all galaxies that have F814W magnitude fainter than 23.5. However, we will use the $m_{\rm F814W} \leqslant 25.2$ sample that was generated for the GREAT3 challenge to estimate the completeness, which is useful when generating a volume-limited sample (Sec. 3.2). We first use the $z < 1$ flux-limited sample to fit parametric redshift distribution models (Sec. 3.1), and then restrict ourselves to $z \leqslant 1$ sample for all further analysis.

Stellar mass estimates were obtained (Leauthaud et al. 2010) using the Bayesian code described in Bundy et al. (2006). This process involves constructing a grid of models that vary in age, star formation history, dust content and metallicity (always assuming a Chabrier IMF; Chabrier 2003), to which the observed galaxy spectral energy distributions (SEDs) and photometric redshift are compared. At each grid point, the probability that the SED fits the model is calculated, and by marginalizing over the nuisance parameters in the grid, the stellar mass probability distribution is obtained. The median of this distribution is taken as the stellar mass estimate.

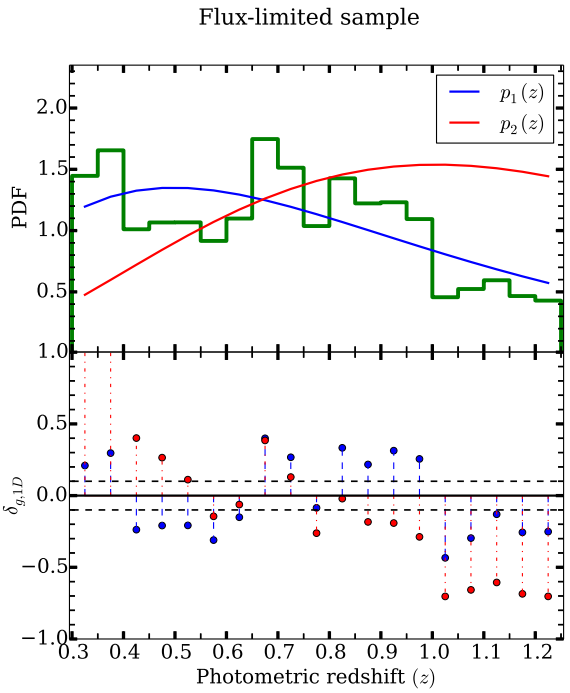

Figure 1. Upper panel: Redshift distribution of flux-limited ($m_{\rm F814W} \leqslant 23.5$) sample with photometric redshift bins that are 0.05 wide. Two analytical functions, $p_1(z)$ and $p_2(z)$, defined in Eqns. 1 and 2 with best fit parameters are plotted over it. Lower panel: Plot of $\delta_{g,1D} = N/N_{\rm mod} - 1$ with each functional form as the model for each redshift bin. Note to self: Give parameter values. How does this figure look if you show the plot range with the histogram going to $z = 0$ (still only fitting to $0.3 < z < 1.25$)? It might be nicer to do that and put a vertical line at $z = 0.3$ to remind the reader that you only fit above that redshift.

3 METHODS

In order to study the variation in the intrinsic ellipticity distribution and various morphological indicators with the galaxy environment, there are three main steps to be carried out:

(i) Identify overdense and underdense environments in our survey from the redshift distribution of galaxies (Sec. 3.1);
(ii) volume-limit the sample such that Malmquist bias is minimized before comparing galaxies in different redshift slices (Sec. 3.2); and
(iii) estimate the galaxy axis ratios and other morphological indicators such as Sérsic index and bulge-to-total ratios (Sec. 3.3).

In this section we will describe how these steps were carried out.

3.1 Finding overdensities

It is important to keep in mind when considering the environment estimation that our goal is not to create a full 3D mapping of the density field within the COSMOS region (a task that was already addressed by Kovač et al. 2010 using the zCOSMOS spectroscopic sample). Instead, we make a coarse, 1D division of the COSMOS survey into redshift slices, just as would be done when making galaxy redshift slices as input to a weak lensing survey simulation. For each redshift slice, we can then check whether the environment is overdense or underdense on average. Our approach will tend to wash out some real trends from a 3D study, but is appropriate given our scientific goal of testing effects of the environment on weak lensing simulations based on the COSMOS survey.

For our (flux-limited) sample of galaxies, up to $z =$ 1.0 (1.25) (fix this to the correct one once we make final decisions), we fit parametric models to the histograms of photometric redshifts in order to assign values of overdensity. We choose our bins to be 0.05 wide starting from $z = 0.3$, where the bin width is selected to be somewhat larger than the photometric redshift error but narrow enough that we can still identify rather than averaging over real cosmological structures. We neglect the lowest redshifts which have negligible cosmological volume and where the galaxy population tends to be intrinsically bright and large enough that a non-negligible fraction is lost due to the cuts we impose (Sec. 2).

The parametric redshift distributions that we use are

$$p_1(z) \propto z^{a-1} \exp\left[-bz\right] \qquad (1)$$

and

$$p_2(z) \propto z^2 \exp\left[-\left(\frac{z}{z_0}\right)^{1.5}\right], \qquad (2)$$

the latter of which was first presented by Baugh & Efstathiou (1993). Here a, b, and z_0 are free parameters that are to be determined. The normalization constants depend not only on the parameters but also on the lower and the upper limit of the redshifts considered, where we fix the normalization to ensure that the predicted number of galaxies in the range used ($0.3 < z < 1$) is

equal to the actual number. Fig. 1 shows the photometric redshift histogram together with the best-fitting parametric distributions. New suggestion re: 2nd redshift distribution: it just looks so bad, I worry that this functional form is simply wrong and we should use something newer that has been shown to work with modern redshift surveys. Let me suggest something you can do as an immediate sanity check. Look at Coil et al. (2004) table 4, which gives parametrized fits to the redshift distribution from DEEP2 for samples binned by $R_{\rm AB}$ and $I_{\rm AB}$. Use the results for the $18 < I_{\rm AB} < 23$ bin and the $18 < I_{\rm AB} < 24$ bin, interpolating linearly between them since your sample is limited at 23.5. Their results suggest that either $z^2 \exp\left[-z/0.262\right]$ or $z^2 \exp\left[-(z/0.361)^{1.2}\right]$ should be decent descriptions of your sample. Just as a quick sanity check can you please plot those distributions (appropriately normalized) on top of what you have in the top panel of figure 1, and see how they look? If they look OK, then I would be inclined to completely eliminate the 2nd distribution, and just use the 1st one that you're fitting and one of the ones from Coil as your way of checking overdensities and underdensities.

The estimated overdensity in a redshift bin is defined by comparing the observed galaxy counts in the bin with the counts that are predicted in that bin by one of the models in Eqs. (1) and (2):

$$\delta_{g,\rm 1D} = \frac{(N - N_{\rm mod})}{N_{\rm mod}}, \tag{3}$$

where

$$N_{\rm mod} = \int_{z_{\rm min}}^{z_{\rm max}} p(z)\,{\rm d}z \tag{4}$$

is determined by integrating the redshift distribution within the limits of that redshift slice. Note that $\delta_{g,\rm 1D}$ is dependent on our choice of model redshift distribution, and should have a mean value of 0 over the entire redshift range.

Our decision criterion for identifying overdense and underdense redshift slices involves leaving a 10 per cent margin around an overdensity of zero; i.e., if $|\delta_{g,\rm 1D}| < 0.1$, that is considered "neutral" (neither overdense nor underdense on average). We can then label each redshift slice as either overdense, underdense, or neutral as follows: We label a redshift bin as overdense if at least one model gives a value of $\delta_{g,\rm 1D} > 0.1$ while the other gives $\delta_{g,\rm 1D} > -0.1$ (neutral or overdense), and vice versa for the underdense regions. We label a redshift bin as neutral if both models give $\delta_{g,\rm 1D}$ within the neutral region, *or* if use of one model redshift distribution results in the conclusion that the bin is overdense while the over leads to the conclusion that it is underdense. We thus identify the regions $z = 0.30 - 0.40$, $0.65 - 0.75$, and $0.80 - 0.85$ as overdense; $z = 0.55 - 0.65$ and $0.75 - 0.80$ as underdense; and $z = 0.40 - 0.55$ as unclassified.

We have adopted this purely 1D environment classification for reasons explained at the beginning of this section. However, as a sanity check we can compare it with a more rigorous study that includes information about structure in the plane of the sky. Kovač et al. (2010) used a sample of ∼10 000 zCOSMOS spectroscopic galaxies with $I_{AB} < 22.5$ to reconstruct the three dimensional overdensity field up to $z \sim 1$. We find that our classification of overdensities and underdensities agrees with this work, except for our two highest redshift bins. We believe that this disagreement is due to the errors in our photometric redshifts, with the overdensity reported by Kovač et al. (2010) in the $z = 0.875 - 1$ range leaking into our $z = 0.80 - 0.85$ slice.

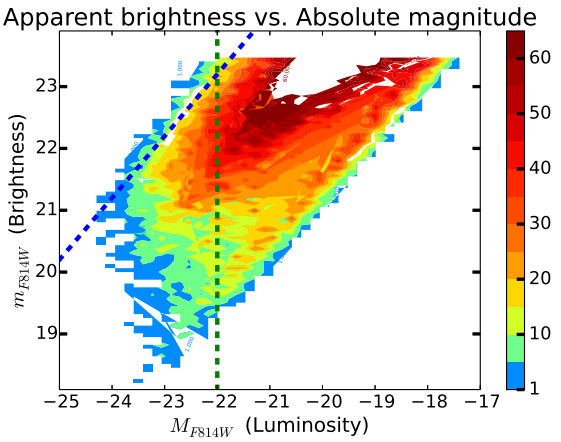

Figure 2. 2D histogram of galaxies in apparent magnitude ($m_{\rm F814W}$) and absolute magnitude (M_I) in the redshift range $[0.3, 1.0]$. The blue dotted line with a unit slope is a boundary at $z = 1$, fitted by eye. The green vertical dotted line represents our luminosity cut.

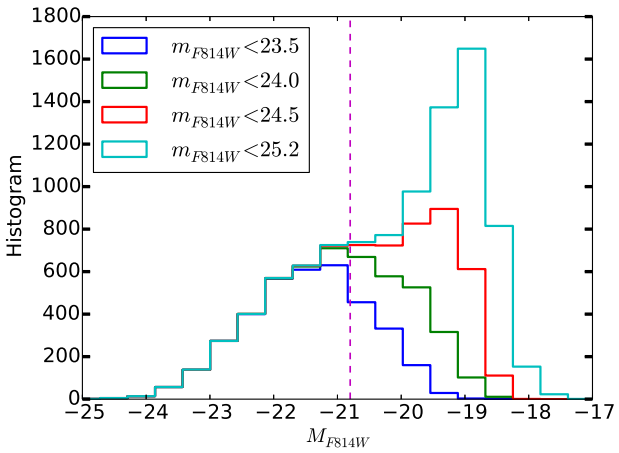

Figure 3. Distribution of absolute magnitude M_I for various flux-limited samples in the redshift range $[0.80, 0.85]$ are plotted together. The vertical line corresponds to our luminosity cut of -20.8, brighter than which the $m_{\rm F814W} < 23.5$ sample includes > 95.3 per cent of the galaxies in the $m_{\rm F814W} < 25.2$ sample.

3.2 Volume limiting

COSMOS is a flux-limited survey and is therefore affected by Malmquist bias, with the galaxy samples at higher redshift being intrinsically brighter on average. Our analysis involves comparing galaxies in different redshift slices to identity significant differences in morphology that arise due to morphology-density correlations. Such an analysis would be very difficult with a flux-limited sample because there would

be some variation in morphology with redshift just due to the intrinsic change in the sample properties. For a fair comparison, we must restrict ourselves to galaxies that are bright enough that they would be observed at all redshifts that we consider, which is achieved by volume-limiting the sample. We consider three different ways of carrying out this process, which results in three different galaxy sample selections, all of which we will use in the remainder of the analysis.

Our first approach is to generate a volume-limited sample that is complete up to $z = 1$, by applying a cut on luminosity such that only galaxies instrinsically brighter than a certain threshold (determined in detail below) is considered. This threshold is set on the k-corrected I-band absolute magnitudes (M_I) from the COSMOS PSF-matched photometry catalog. The joint distribution of $m_{\rm F814W}$ and M_I is shown in Fig. 2. Since the parent sample contains fainter galaxies and is quite complete to $m_{\rm F814W} = 25.2$, we compare the M_I distribution of the $m_{\rm F814W} = 23.5$ sample with flux-limited samples that have fainter flux limits, to see where the $m_{\rm F814W} < 23.5$ sample that we want to use for our tests is no longer complete. At $M_I \sim -22.0$, the $m_{\rm F814W} < 23.5$ sample is beginning to lose galaxies in the $0.9 < z < 1.0$ redshift bin due to the flux limit. However, the $0.85 < z < 1$ redshift bin was found to be only moderately overdense, so we choose to disregard this region for the rest of the analysis, and instead restrict to $z < 0.85$, which is advantageous because it allows us to choose a somewhat fainter intrinsic luminosity limit for the analysis. We relax our luminosity cut so that the sample is volume-limited *not* until $z = 1$ but until $z = 0.85$. We impose the cut at $M_I = -20.8$, which gives 95.3 per cent completeness in the $0.8 < z \leqslant 0.85$ bin (see Fig. 3). The resulting sample, which has 11 169 galaxies, will be called sample $\mathcal{S}1$ in the remainder of this work.

However, previous studies (e.g., Wolf et al. 2003; Giallongo et al. 2005; Willmer et al. 2006; Faber et al. 2007) have shown that galaxy intrinsic luminosities evolve with redshift. Thus, we should also let the luminosity cut that we apply to volume-limit the sample evolve with redshift. Unfortunately, the majority of the published work on evolution of the luminosity function uses B and V band data, and it is not apparent that the results should be the same in a redder passband like I. We use the results from Faber et al. (2007) for the evolution of B-band magnitudes from the DEEP2 and COMBO-17 surveys, which is $\Delta M_B^* \sim -1.23$ mag per unit redshift (with the sign indicating that galaxies were intrinsically brighter in the past), for a combined sample of blue and red galaxies. Typically, estimates of evolution in the much redder K band are less than the estimates of evolution in B and V bands. (Please include a citation for this statement.) Assuming that the evolution is a smooth function of the wavelength, the evolution in I-band should be in between B and K band. Therefore, by considering no evolution (a lower limit) as in our $\mathcal{S}1$, and a second sample $\mathcal{S}2$ constructed using the B-band evolution (as an upper bound on the I-band evolution), we can assume that these two samples bracket reality, which should be somewhere in between.

Thus, $\mathcal{S}2$ is constructed by letting the luminosity cut evolve, starting from $M_I = -20.8$ (same as in $\mathcal{S}1$) for the $0.8 < z \leqslant 0.85$ bin. The cut values for the other bins are defined by allowing 1.23 magnitudes of evolution to fainter magnitudes as a function of redshift (evaluated using the bin centers). Because of the sign of redshift evolution, $\mathcal{S}2$ includes more galaxies.

Redshift	Environment	$\mathcal{S}1$	$\mathcal{S}2$	$\mathcal{S}3$
0.3-0.4	Overdense	1726	2505	1260
0.4-0.475	Neutral	988	1317	710
0.475-0.55	Neutral	1410	1788	902
0.55-0.65	Underdense	1797	2193	1183
0.65-0.75	Overdense	4059	4476	2593
0.75-0.8	Underdense	1159	1196	675
0.8-0.85	Overdense	2428	2428	1630

Table 1. List of different redshift bins, their environmental classification and the number of galaxies per redshift bin for volume-limited samples constructed in three different ways: using a hard luminosity cut ($\mathcal{S}1$), using a redshift-dependent luminosity cut ($\mathcal{S}2$) and using stellar-mass cuts ($\mathcal{S}3$).

One might wonder why we cannot use the luminosity function in F814W based on the COSMOS observations to directly determine the luminosity function for our sample, thus simplifying this exercise. However, this turns out to be highly non-trivial for two reasons. First, the F814W observations are relatively shallow compared to the deep ground-based observations used in many other works for determination of luminosity evolution. As a result, it is difficult to get a handle on the faint end of the luminosity function, and the unknown faint-end slope turns out to be degenerate with the evolution of the typical luminosity. Second, the photometric redshift error is a complicating factor that requires sophisticated techniques to remove. A derivation of the I-band luminosity evolution is therefore beyond the scope of this work.

Finally, we can circumvent the problem of redshift evolution of the luminosity by imposing cuts on stellar mass instead. In Fig. 4, we show the stellar mass function (SMF) of our sample for various F814W flux limits. Tomczak et al. (2014) report the SMFs for the ZFOURGE survey, which includes COSMOS. They considered for this work a single stellar population (What do you mean by a single stellar population?) following a Chabrier IMF (Chabrier 2003). We plot their SMF for *all* galaxies in Fig. 4 for reference. Their SMF is higher than ours since they reach a K_s-band 5σ depth of 24.9. As for M_I, we compare the stellar mass function in the $m_{\rm F814W} \leqslant 23.5$ sample with that in the $m_{\rm F814W} \leqslant 25.2$ sample. The sample with $\log(M/M_\odot) > 10.15$ is ~ 95 per cent complete in the redshift bin $0.75 \leqslant z < 0.85$ and has 10 341 galaxies in total across all redshifts. Thus, we construct a third volume-limited sample $\mathcal{S}3$ by imposing the stellar mass cut mentioned above. The number of galaxies in redshift slices are tabulated in Table 1 for all three ways discussed in this section for obtaining a volume-limited sample. The stellar-mass limited sample is the smallest, most likely because when converting from flux to stellar mass, the stellar mass-to-light ratios vary strongly with galaxy type, so red galaxies with high M_*/L simply have too low a flux compared to the blue galaxies at the same M_*, and are not observed.

There is one subtlety in our method used for estimating completeness. We have used the full $m_{\rm F814W} \leqslant 23.5$ sample for identifying overdensities and for the completeness cal-

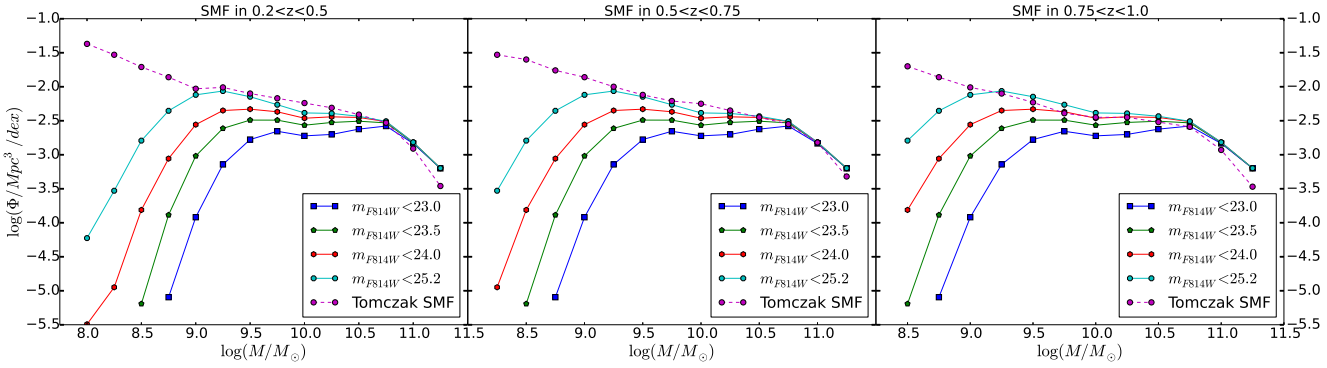

Figure 4. Stellar mass distribution for various flux-limited samples are shown in three redshift ranges as separate panels. The redshift bins have been chosen to facilitate the comparison with a study of the SMF in Tomczak et al. (2014). At high mass, the distributions are the same for various flux limits, indicating that the samples are complete in that mass range. The curves begin to deviate at low masses due to incompleteness coming from the flux limit. The point at which the deviation exceeds our threshold determines where the mass cutoff should be to volume-limit the sample.

culations that motivated our definitions of volume-limited samples. However, everywhere else in the paper, we consider only those galaxies for which there are postage stamp images used to create weak lensing simulations, in part because this is the sample for which fits to Sérsic profiles were carried out, which is a requirement for our morphology analysis. 12 per cent of the galaxies that pass our cuts do not have an associated postage stamp image. Postage stamps may not exist because, given the size of the galaxy, the size of the postage stamp we want to draw around it (including some blank space) intersects the edge of the CCD. If all galaxies were the same size, this would be a purely random effect, but in fact bigger galaxies are more likely to get excluded by this cut. It is commonly the case that galaxies that are nearby and intrinsically very bright do not have postage stamps associated with them, an effect that is dominant at lower redshifts (and is part of our reason for excluding $z < 0.3$). Our completeness calculation is done at high redshifts, and thus we believe that our conclusions are not affected by this bias.

The functional forms for the (flux-limited) redshift distribution that we used in Sec. 3.1 are not physically motivated. We fit them again to the redshift distribution of a volume-limited sample ($\mathcal{S}1$). Fig. 5 shows that when we do this, the values of $\delta_{g,1\mathrm{D}}$ for the $z = 0.40 - 0.55$ bin increase and are within the $[-0.1, 0.1]$ range that we have defined as neutral. This is the reason that in Sec. 3.1 we classified them as neutral as opposed to underdense. We will see in Sec. 4 that they are more similar to overdense regions as opposed to underdense regions. The other redshift slices seem to exhibit a consistent behavior in Fig. 5 as in Fig. 1.

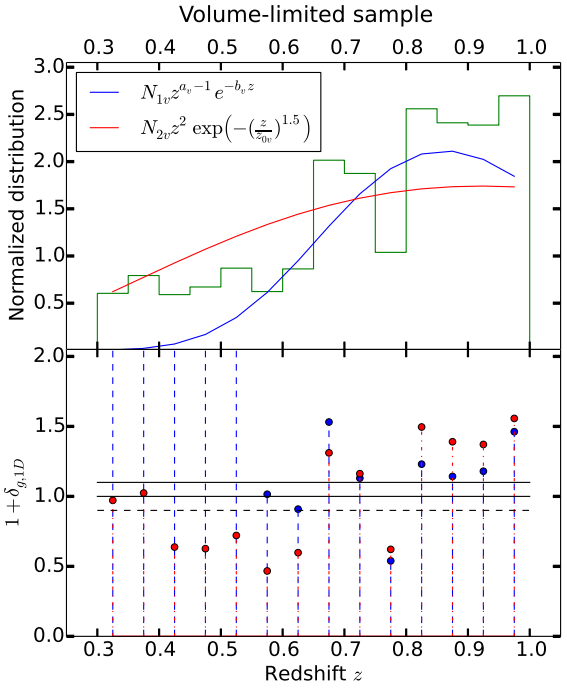

Figure 5. Upper panel: Redshift distribution of volume-limited ($M_I < -20.8$) $\mathcal{S}1$ sample with photometric redshift bins that are 0.05 wide. Two analytical functions with best fit parameters are plotted over it. Lower panel: Plot of $(1 + \delta_{g,1\mathrm{D}}) = N/N_{\mathrm{mod}}$ with each functional form as the model for each redshift bin. I am quite concerned that something has gone wrong with this fit (in the sense that there may be a bug). Why is the redshift distribution shown in blue forced to zero at $z = 0.3$? It should go to zero at $z = 0$, not 0.3. It is clearly not normalized to the same total number as the redshift histogram, either. In terms of formatting, please implement the same changes that I suggested for the flux-limited sample redshift distribution plot.

3.3 Describing galaxy morphology and shape

We choose simple and well-motivated ways to parametrize galaxy shapes and morphology based on existing methods in the literature. These methods have the advantage of being stable and well-defined in nearly all cases. However, for highly irregular galaxies the meaning of the structural parameters that we derive is not entirely clear. In all cases, our methods account for the effect of the *HST* PSF.

One method to estimate the galaxy ellipticities and other morphological parameters is to fit parametric models convolved with the PSF to the observed galaxy light profile. We use the fits from Mandelbaum et al. (2014), which used

the methods and software from Lackner & Gunn (2012) to fit the images to the following profiles:

(i) A Sérsic profile given by the expression

$$I_S(x,y) = I_{1/2} \exp\left[-k(R(x,y)/R_{\rm eff})^{1/n} - 1\right], \quad (5)$$

where

$$R^2(x,y) = ((x-x_0)\cos\Phi + (y-y_0)\sin\Phi)^2 + ((y-y_0)\cos\Phi - (x-x_0)\sin\Phi)^2/q^2,$$

$R_{\rm eff}$ is the half-light radius of the profile defined along hte major axis, I_0 is the surface brightness at $R = R_{\rm eff}$, (x_0, y_0) is the centroid of the image, Φ is the position angle, n is the Sérsic index, k is a n-dependent normalization factor required to ensure that half the light is enclosed within the half-light radius, and q is the axis ratio of the elliptical isophotes. Thus, the Sérsic profile has 7 free parameters.

(ii) A sum of two Sérsic component fits: a de Vaucouleurs bulge ($n = 4$) plus an exponential disc profile ($n = 1$). In this case, there are 10 free parameters, because the Sérsic indices are fixed, and the two components are constrained to have the same centroid.

More details about the fitting algorithm can be found in Lackner & Gunn (2012).

To quantify galaxy morphology and shape, we will use several quantities from the above fits. First, from the single Sérsic profile fits, we use the Sérsic index and the axis ratio. The axis ratio can also be used to derive a distortion,

$$e = \frac{1-q^2}{1+q^2} \quad (6)$$

or an ellipticity,

$$\varepsilon = \frac{1-q}{1+q}. \quad (7)$$

As an alternative morphological indicator (instead of Sérsic index) we use a bulge-to-total ratio derived from the double Sérsic profile fits. This ratio is defined in terms of bulge and disk fluxes as

$$\frac{B}{T} = \frac{f_{\rm bulge}}{f_{\rm bulge} + f_{\rm disk}}. \quad (8)$$

(Note, the equation that was previously in here was wrong. It's not defined in terms of surface brightnesses, but rather in terms of flux. Were you using numbers from my files for B/T? If so, then it was calculated properly. But if your analysis uses the equation that was in here before, then you will have to redo it. Sorry for not noticing this earlier.)

I think it would be useful to take the entire sample that we use for science, and show the overall distributions of axis ratio, distortion (one curve for Claire's and one curve for re-Gaussianization), Sérsic index, and Bulge-to-Total . This would be a nearly full-page four panel figure, but I think it's worthwhile to illustrate the nature of the sample. For example, it will clearly show why we cannot use the distributions of Sérsic index and Bulge-to-Total , because of the hard cutoffs.

We also consider an alternative method for estimating the galaxy ellipticity. This method is based on using the observed weighted moments of the galaxy and PSF, and correcting those of the galaxy for those of the PSF. This PSF correction scheme is the re-Gaussianization method described in section 2.4 of Hirata & Seljak (2003) (hereafter HS03) as implemented in the GalSim software package (with implementation details described in Rowe et al. 2014). This method models the true PSF $g(\mathbf{x})$ as a Gaussian $G(\mathbf{x})$ and the residual $\epsilon(\mathbf{x}) = g(\mathbf{x}) - G(\mathbf{x})$ is assumed to be small. Thus, the Gaussian-convolved intrinsic image, f, can be modeled as $I' = G \otimes f = I - \epsilon \otimes f$, where I is the observed image. The crucial idea here is that, when ϵ is small, we get a reasonably accurate estimate of I' even if we use an approximate form for f. The form assumed for f is that of a Gaussian with covariance $M_f^{(0)} = M_I - M_g$, where M_I and M_g are the elliptical Gaussian-weighted adaptive covariances of the measured object and PSF respectively, described in section 2.1 of HS03 and Bernstein & Jarvis (2002). We refer to the re-Gaussianization estimates of the PSF-corrected distortion as "moments-based shape estimates". The value in including them in this analysis is that they have quite different radial weighting from the Sérsic profile fits, with the outer regions being quite downweighted when calculating adaptive moments. Thus, if ellipticity gradients are important, we could get different results using these two shape estimators.

4 RESULTS

Having identified the overdense and underdense regions in a volume-limited sample (Secs. 3.1 and 3.2), we will now see whether the morphological parameters of the galaxies described in Sec. 3.3 depend noticably on the environment of the redshift slice in which they reside. Note that for true 3D overdensities there is already substantial evidence in the literature that we should see variation of properties with the environment. Our test is necessary to see whether such morphology-density correlations are evident in the kind of 1D redshift slices that would be used for constructing weak lensing simulations, or whether our use of an area the size of COSMOS will wash out these trends (which would be good news for weak lensing simulations based on that dataset).

As described in Sec. 3.2, we have three different ways of volume-limiting our sample:

(i) no redshift evolution of luminosity cut ($\mathcal{S}1$),
(ii) using B-band luminosity evolution applied to the I-band luminosities ($\mathcal{S}2$), and
(iii) impose stellar mass cuts instead of luminosity ($\mathcal{S}3$).

We will present our results in all three cases to check for their robustness to how the sample is selected.

4.1 Axis ratios

We can test the influence of environment on the galaxy shapes by comparing the distributions of the axis ratios for the overdense and underdense redshift slices, or by encapsulating that distribution as a single number, the RMS (root mean squared) ellipticity. By volume-limiting the sample, we have avoided issues wherein the flux limit leads to artificial changes in the sample as a function of redshift. We will also carry out tests to differentiate between environmental effects versus evolution of the population with redshift (at fixed mass).

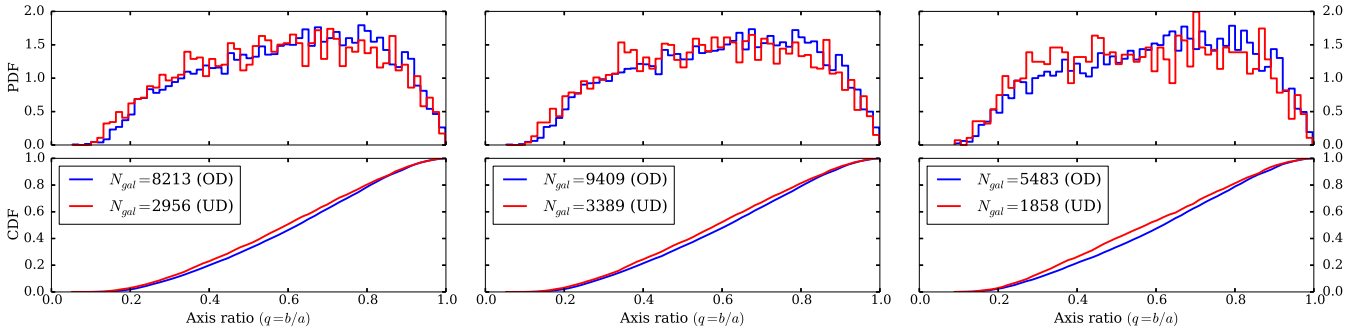

Figure 6. The distributions of axis ratios of galaxies in *all* overdense (OD) and *all* underdense (UD) regions in the case of the luminosity-selected sample $\mathcal{S}1$ (left), luminosity-selected sample with B-band evolution taken into account $\mathcal{S}2$ (center), and thestellar-mass-selected sample $\mathcal{S}3$ (right). The upper panels show the histograms, and the bottom panels show the cumulative distribution functions (CDF). The p-values computed using these CDFs are shown in Table. 2.

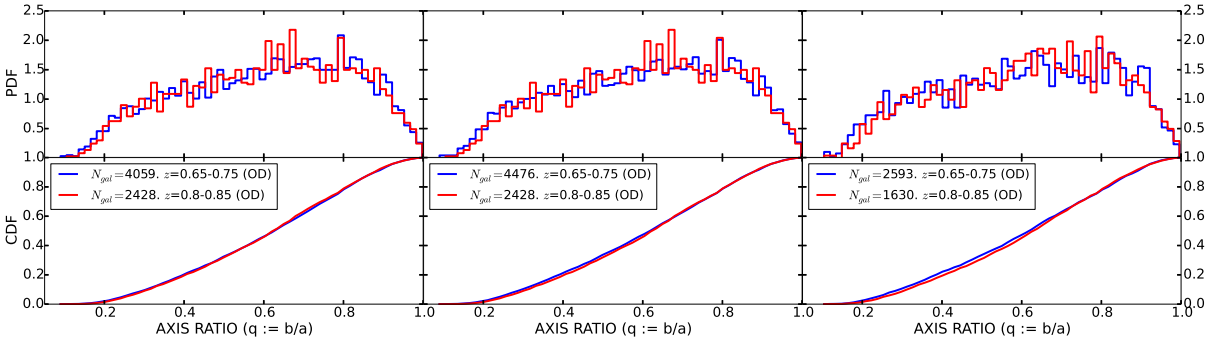

Figure 7. Galaxy axis ratio distributions in two overdense redshift bins, $z = 0.65 - 0.75$ and $z = 0.80 - 0.85$, to check for consistency in the case that the environment is the same even if the redshift differs. The p-values from the KS and AD tests are given in Table. 2. Figure does not appear to be centered.

Redshift bins	$\mathcal{S}1$	$\mathcal{S}2$	$\mathcal{S}3$
All overdense vs. All underdense	1.1×10^{-4} 1×10^{-5}	2.6×10^{-5} $< 1 \times 10^{-5}$	1.9×10^{-6} $< 1 \times 10^{-5}$
$[0.65, 0.75]$ (OD) vs. $[0.80, 0.85]$ (OD)	0.613 0.494	0.431 0.237	0.231 0.130
$[0.65, 0.75]$ (OD) vs. $[0.55, 0.65]$ (UD)	5.8×10^{-4} 9.8×10^{-4}	1.5×10^{-5} $< 1 \times 10^{-5}$	3.5×10^{-6} $< 1 \times 10^{-5}$

Table 2. p-values from the Kolmogorov-Smirnov (top) and Anderson-Darling (bottom) tests obtained by comparing the distributions of axis ratios for three cases: *all* overdense (OD) vs. *all* underdense (UD), two overdense bins that are not very separated in redshift, and a pair of adjacent overdense and underdense bins. $\mathcal{S}1$, $\mathcal{S}2$, $\mathcal{S}3$ refer to the three different types of volume-limited samples. The Anderson-Darling p-values are correct only up to 5 decimal places. What does this last sentence mean? Why? I think you need clarification of this point in this table and the next one.

Redshift bins	$\mathcal{S}1$	$\mathcal{S}2$	$\mathcal{S}3$
All overdense vs. All underdense	5.6×10^{-4} 3×10^{-5}	1.0×10^{-4} 1×10^{-5}	3.3×10^{-6} $< 1 \times 10^{-5}$
$[0.65, 0.75]$ (OD) vs. $[0.80, 0.85]$ (OD)	0.9563 0.5162	0.7476 0.3352	0.5359 0.2290
$[0.65, 0.75]$ (OD) vs. $[0.55, 0.65]$ (UD)	6.0×10^{-3} 1.2×10^{-2}	2.5×10^{-4} 2.5×10^{-4}	2.4×10^{4} 5×10^{-5}

Table 3. p-values from the Kolmogorov-Smirnov (top) and Anderson-Darling (bottom) obtained by comparing the second moments-based ellipticities for the three same three cases as in Table 2. Should I write only two significant digits for higher p-values? Yes. The same comment applies to the previous table, too. The Anderson-Darling p-values are correct only up to 5 decimal places.

4.1.1 Comparing distributions

We begin by comparing the entire axis ratio distributions $p(q)$ between pairs of redshift slices. Unless otherwise mentioned, the axis ratios refer to the values obtained using the method of Lackner & Gunn (2012) to fit single Sérsic profiles to each galaxy image. To compare the distributions and make statistical statements about their consistency, we use two statistical tests, the Kolmogorov-Smirnov (KS) test and Anderson-Darling (AD) test, the latter of which is carried out using the `adk` package in `R`.

We first compare the distribution of galaxy axis ratios in *all* overdense bins against that for *all* underdense bins in Fig. 6, with different panels showing the comparison for $\mathcal{S}1$, $\mathcal{S}2$, and $\mathcal{S}3$. The cumulative distinction functions are also

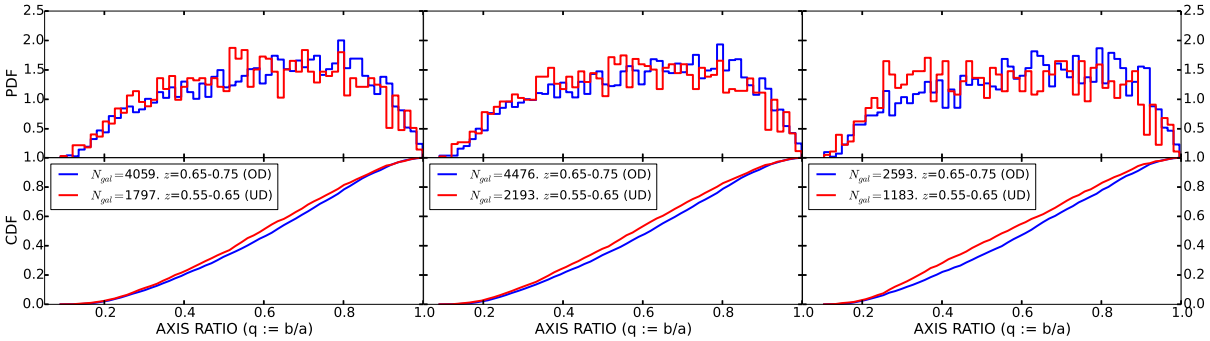

Figure 8. Galaxy axis ratio distributions in a single underdense redshift slice, $z = 0.55 - 0.65$, and a single overdense redshift slice, $z = 0.65 - 0.75$. The p-values from the KS and AD tests are given in Table. 2. Figure does not appear to be centered.

shown, since they form the basis for our statistical statements about consistency using the KS and AD tests. The results of these tests are shown in the first two rows in Table 2. For all three ways of volume-limiting the sample, the p-values from both the KS and AD tests are well below 0.05 (a maximum of 1.1×10^{-4}, but often smaller than that). We can therefore reject the null hypothesis that the overdense and underdense regions have the same underlying axis ratio distributions at high significance.

One might imagine that the disagreement between the distributions is at least partly due to the fact that the overdense and underdense sample have different redshift distributions and there could be some evolution of ellipticity distributions with redshift. To show that this redshift evolution effect is subdominant to environmental effects, we will compare distributions between pairs of two overdense (or pairs of underdense) redshift slices, where we expect to find similarity even if the redshifts are different if the environmental effects dominate. We will also compare between overdense and underdense regions that are selected to be nearby in redshift, so that any redshift evolution effects should be minimal. Figures 7 shows that the axis ratio distributions are indeed consistent when the environments are similar but the redshifts are different. Likewise Fig. 8 shows that for adjacent redshift slices with different environments, the axis ratio distributions are inconsistent. The results of statistical tests for the distributions in these figures are given in Table 2, and support our statement that the morphology-density correlation is the dominant effect when comparing overdense and underdense redshift slices, with redshift evolution of the population being negligible. Comparing other pairs of redshift bins leads to similar conclusions. (This argument would be even stronger if Fig. 7 showed two redshift slices that were further separated in redshift, for example $0.3 - 0.4$ against $0.75 - 0.8$. However the former might not have enough galaxies?)

Finally, we can check whether these findings are particular to the axis ratios from the Sérsic fits, or whether we reproduce this finding when we use the shapes from the centrally-weighted moments-based re-Gaussianization method, which estimates a distortion (Eq. 6) for each galaxy. After neglecting a small fraction (< 0.01 per cent) of galaxies for which the method does not converge, we carry out the same statistical tests from Table 2, but using the moments-based shape estimates. The results of the KS and AD tests are tabulated in Table 3. We see that all of our findings with the Sérsic fit-based shapes carry over to shapes from a centrally-weighted moments-based shape estimate.

4.1.2 RMS ellipticites

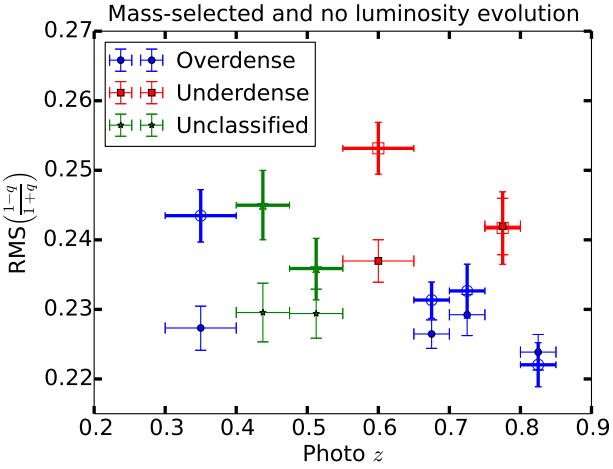

Figure 9. RMS distortions as a function of redshift. The horizontal errorbars indicate the width of the redshift bin, while the vertical ones are 1σ errorbars obtained by bootstrapping. The colors and shapes of the points indicate their environmental classification, as shown in the legend. The title and figure clearly show two samples. The caption needs to explain which are solid points and which are open. Also, the horizontal axis label should be either z or "Photometric redshift (z)", not "Photo z" (which is not a phrase we have used throughout the paper). The change in horizontal axis label is needed for the other RMS ellipticity plots and anywhere else that it shows up in the paper.

Overall, I find it quite confusing that the underdense regions are red and the overdense regions are blue on all of the plots in this section and the section on Sérsic n and B/T. The underdense regions are dominated by blue galaxies and the overdense are dominated by red galaxies, so having the plot colors reversed creates some cognitive dissonance. Can you reverse them to more naturally match up point color with galaxy type?

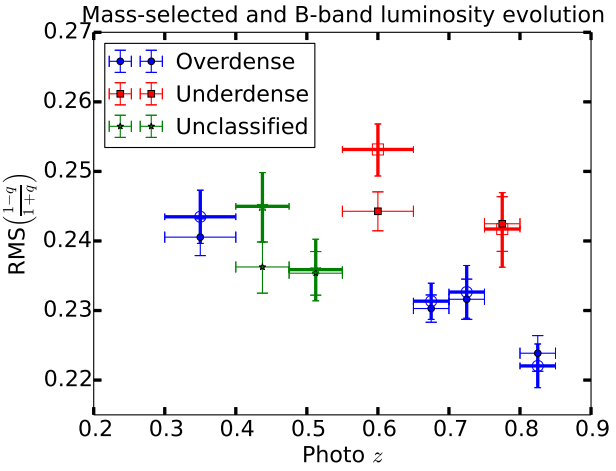

Figure 10. RMS distortions as a function of redshift. The solid points correspond to the sample where the luminosity cut evolves by -1.23 magnitudes per unit redshift ($\mathcal{S}2$) and the open points correspond to the sample obtained from stellar mass cuts ($\mathcal{S}3$). If I compare this with Fig. 9 I would have concluded that the samples were the opposite of what is stated here. The solid points appear the same in both plots, so I thought that was the stellar mass-selected sample. Please check and confirm which way is correct, and include the information in both captions.

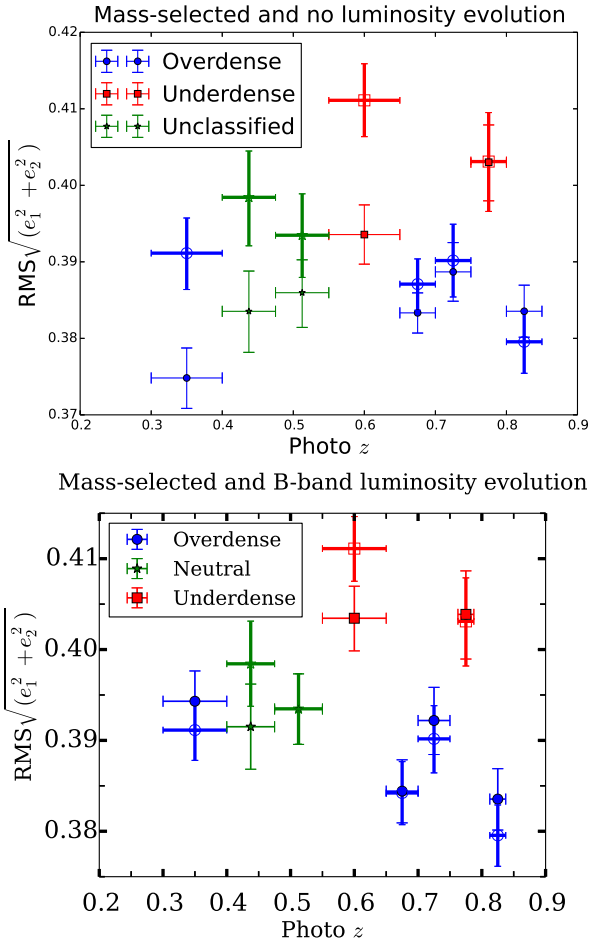

Figure 11. RMS distortions as a function of redshift, with points defined in a similar way as in Fig. 9. In this case, the distortions are the moments-based shape estimates, rather than from fits to Sérsic profiles.

We can also carry out tests on a single statistic of the galaxy shape distribution in each redshift slice, like the RMS ellipticity. While tests of a single quantity may seem less powerful than tests that use the entire shape distributions, the advantage is that instead of picking out pairs of redshift slices for our tests, we can easily compute our statistic of interest for every single redshift slice, and look for trends with both redshift and environment.

For the luminosity-selected sample without any evolution ($\mathcal{S}1$), the RMS distortions (Eq. 6) of galaxies in each redshift bin are shown in Fig. 9. (The figure seems to show two samples, not just $\mathcal{S}1$. Update the text to reflect this?) For this and all similar figures, the colors of the points were selected to easily differentiate between galaxies in overdense, neutral, and underdense redshift slices. As shown, the underdense regions have higher values for RMS ellipticities when compared to the overdense regions. The difference between the underdense and overdense regions for $z > 0.5$ is significantly larger than any redshift evolution across the $z > 0.5$ range. Our conclusions are very similar if we use the RMS ellipticity from Eq. (7) instead of the distortion. (Our statements at the low redshift end regarding the amount of redshift evolution will depend on whether we're using the solid or open points, and the text and caption do not say which samples they correspond to, so I was not able to address this in the text. Please do that once it is your turn to edit.)

The sign of the dependence on the local environment is reasonable when compared with previous work on the morphology-density relation. Overdense regions typically contain many old, elliptical galaxies which are close to round (large axis ratio and low RMS ellipticity). In contrast, the underdense regions typically contain a larger population of younger, disk galaxies, which have lower axis ratios and higher RMS ellipticity. Important note: this argument was given completely backwards. I fixed it, but please watch out for this in future since it's a key point behind this work.

From Figs. 1 and 5, the $0.4 \leqslant z < 0.55$ redshift range shows signs of being marginally underdense, but has somewhat low RMS ellipticities that agree with the rest of the overdense regions.

When the B-band luminosity evolution is taken into account in selecting the sample, a systematic increase in the ellipticity at lower redshifts can be observed. We plot these along with stellar-mass selected samples, where a similar trend is found, in Fig. 10.

Finally, we show an analogous plot of RMS ellipticities for all three volume-limited sample using the moments-based shape estimates in Fig. 11. The conclusions are quite similar to those using the shapes from the Sérsic profile fits, with the underdense regions standing out in having larger RMS distortions than the other redshift slices, an effect that is substantially larger than any average redshift evolution of the RMS distortions.

However, it is generally the case that the statistical significance of trends in this section using a single statistic of the shape distribution (the RMS) is less than the significance

of the trends seen using the entire axis ratio distributions in Sec. 4.1.1.

4.2 Other morphological parameters

For the other morphological parameters that we described in Sec. 3.3, the Sérsic index and bulge-to-total ratio, we do not compare the distributions directly. Doing so is relatively tricky because both distributions have hard cutoffs that are enforced in the fitting process (Sérsic n in the range $[0.25, 6]$ and $0.05 < B/T < 0.95$). The KS and AD statistical tests are both sensitive to exactly what happens at these hard boundaries in the distributions. So, instead of using the full distributions, we will study the dependence of these quantities on environment by computing the median values in different redshift slices. Median values are preferred over the full distributions or even the sample means since the medians are more robust to what happens at the edges of the distributions.

Fig. 12 shows the median value of the Sérsic index in each redshift slice, with and without taking into account of luminosity evolution when volume-limiting the samples ($\mathcal{S}1$ and $\mathcal{S}2$). Both panels also show the results with stellar mass-selected samples ($\mathcal{S}3$) for reference. When using the stellar mass-selected sample, we observe that the overdense regions tend to have higher Sérsic index than the underdense ones, with the redshift evolution being mild and the results for the underdense regions particularly standing out. This trend is consistent with our previous explanation for trends in RMS ellipticities; the underdense regions have more spiral galaxies and therefore a lower median value of Sérsic n. However, this trend is less evident for the luminosity-selected samples, where there seems to be some evolution with redshift that dominates over the environmental effects.

We also note that in Fig. 12, the median Sérsic indices of the stellar mass-selected samples ($\mathcal{S}3$) are systematically greater than those of the luminosity selected samples ($\mathcal{S}1$, $\mathcal{S}2$). We saw that $\mathcal{S}3$ is the smallest sample, and thus is likely limited to higher mass galaxies overall. It is therefore not surprising that it has a higher median Sérsic n. Finally, even for the stellar mass-selected sample there is some sign of redshift evolution. The sign of this evolution is as expected, with lower Sérsic n and B/T for higher redshift samples, which should have a higher fraction of disk and irregular galaxies and fewer galaxies with bulge-like morphology.

Finally, Fig. 13 shows the variation of the median bulge-to-total ratio with redshift. The results are quite consistent with those of Fig. 12. Thus, our results in this section suggest that the environment can significantly affect the median morphological parameters of galaxies selected in thin redshift slices, assuming that the galaxies represent a stellar mass-selected sample. The trend is less evident when using luminosity to select the galaxies.

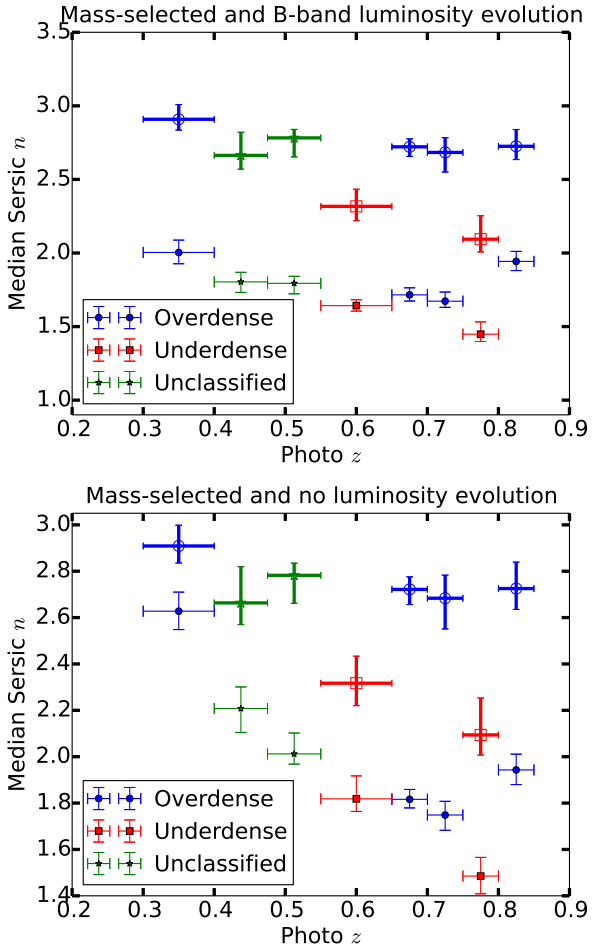

Figure 12. Median values of the Sérsic indices for volume-limited samples $\mathcal{S}1$ and $\mathcal{S}3$ are plotted (filled centers and thin errorbars) in top and bottom panels, respectively, for each redshift bin. Median values for the $\mathcal{S}2$ sample are plotted in both the panels (open centers and thick errorbars) in both the panels. Now I'm really confused. The text implies that the top and bottom are $\mathcal{S}1$ and $\mathcal{S}2$ as open points, with $\mathcal{S}3$ in both panels as solid points. The caption says something different regarding which sample is in both panels and which is open vs. closed. Please clarify and make everything consistent! Also, you only need to have a legend on one panel, since the legends on both panels are the same. Same comment applies to next figure.

5 CONCLUSIONS AND IMPLICATIONS

Here is one thing that is still missing from the results section. We had discussed a number of times lumping the galaxies together into slices with larger Δz and showing that the trends in the shape distributions disappear or are reduced. You had shown me plots for this at some point, but they don't seem to have gone into the paper. Please include them, because it would be useful to be able to give some guidance on how wide the redshift slices should be to avoid this issue.

In this study, we have shown that the shape distributions of galaxies (to a statistically significant degree) and morphological parameters like Sérsic n and bulge-to-total ratios (more marginally) depend on the local environments when dividing up the COSMOS sample into redshift slices along the line of sight. The redshift slices used had a width of $\Delta z = 0.05$. Our findings are robust to the choice of shape estimator from Sérsic profile fits vs. using centrally weighted moments-based shear estimates.

These findings are relevant to attempts to use *HST*-based galaxy samples to calibrate shear estimates in weak lensing surveys. In general, the approach would be to define galaxy samples using all galaxies in redshift slices, and de-

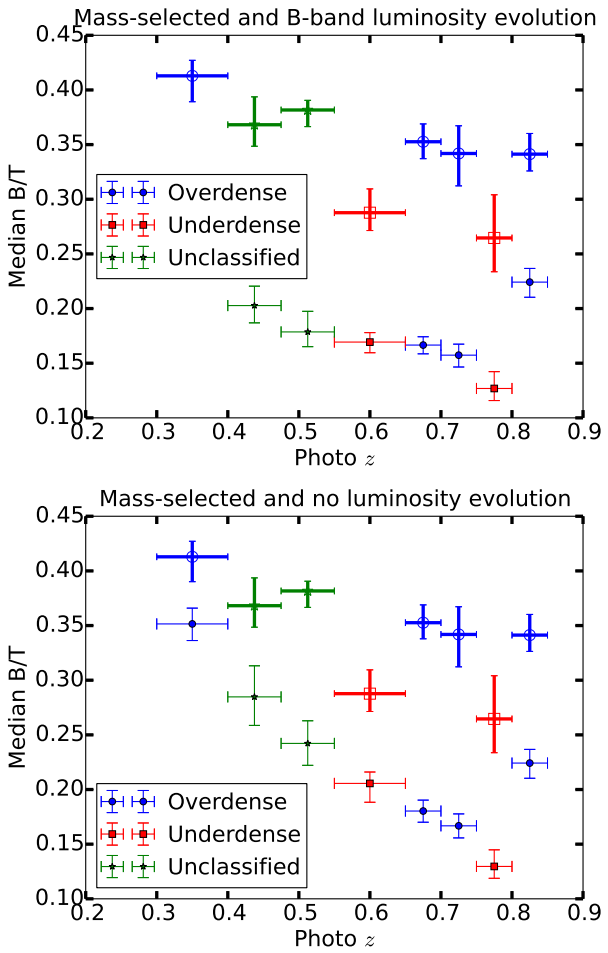

Figure 13. Same confusion here as in previous figure caption; please clarify! Median values of the bulge-to-total ratios for volume-limited samples $\mathcal{S}1$ and $\mathcal{S}3$ are plotted (filled centers and thin errorbars) in left and right panels respectively for each redshift bin. Median values for the $\mathcal{S}2$ sample are plotted in both the panels (open centers and thick errorbars) in both the panels. The horizontal errorbars simply correspond to the binwidth while the vertical ones are 1σ errorbars obtained by bootstrapping. These look so similar that I am questioning the need to have both. Perhaps remove one and mention that they are nearly identical in the text.

termine a redshift-dependent shear calibration. Our findings highlight the danger in such an approach: while we would like our simulations to include true evolution in galaxy properties with redshift, this approach also includes spurious variations in galaxy properties due to the large-scale structure within the COSMOS field. Since the fidelity of weak lensing shear estimates depends sensitively on the intrinsic shape distribution and galaxy morphological parameters, the conclusions for the redshift-dependent shear calibration would be incorrect. These errors would be reduced as the redshift slices that are used become wider, so that the impact of local overdensities becomes washed out.

It is important to keep in mind the nature of COSMOS with respect to other possible *HST* training samples. COSMOS represents the largest contiguous field surveyed by the *HST*, with the sizes of other *HST* fields such as GOODS and AEGIS lagging by at least an order of magnitude. Hence, if cosmic variance due to structures along the line-of-sight in COSMOS are problematic for its use as a training sample for weak lensing simulations, studies that use even smaller area training samples are even more prone to errors, with the UDF serving as an extreme case. If the size of the *HST* survey is small enough, there is no reason *a priori* to suppose that the galaxy population is typical even when using all galaxies along the line of sight, without any division into redshift bins. Of course, future surveys are unlikely to pick just a single survey to serve as the basis for their image simulation training sample, but rather to combine as many as possible. Combining multiple surveys will reduce the cosmic variance and therefore the significance of the effects discussed in this work. However, with COSMOS being so dominant in its size, the combination with other much smaller fields is unlikely to fully ameliorate this effect. Thus, it will be important to carefully choose the size of redshift slices used to derive properties of the galaxy population so as to be minimally affected by this issue.

It is natural to ask the question of how relevant are variations in the galaxy population of the size found here for shear calibration bias estimates. It is likely that the answer to this question varies quite significantly with the type of shear estimation method used. Some will be sensitive to the variations in morphology, others to the variation in the intrinsic ellipticity distribution, and many will be sensitive to both at some level. We consider each in turn.

The intrinsic ellipticity distribution plays a role in nearly all shear estimators. In some, the role is explicit: for example, LensFit (Miller et al. 2007; Kitching et al. 2008; Miller et al. 2013) and the methods presented by Bernstein & Armstrong (2014) require accurate intrinsic ellipticity distributions as inputs, and uncertainty in the distribution was one of the important sources of systematic uncertainty in the CFHTLenS weak lensing results (Heymans et al. 2013; Miller et al. 2013). In others, the role is implicit. For example, the re-Gaussianization method and several other moment-based methods require calculation of a shear responsivity (Bernstein & Jarvis 2002; Hirata & Seljak 2003) that describes how the galaxy population overall responds to a shear, based on its intrinsic ellipticity distribution. The responsivity can be calculated based on the observed shape distribution, assuming that the uncertainties in the shears are known enough that their contribution to that distribution can be removed. If a simulated sample in some redshift slice has a different intrinsic ellipticity distribution and therefore responsivity, it could lead to incorrect conclusions about shear calibration. The responsivity scales roughly like $1 - e^2_{\rm RMS}$, which means that deviations in RMS ellipticity at the level of 0.01 would become fractional shear errors of

$$\frac{\Delta\gamma}{\gamma} \approx \frac{2e_{\rm RMS}\Delta e_{\rm RMS}}{1 - e^2_{\rm RMS}} \approx 0.01. \quad (9)$$

In the context of upcoming lensing surveys that seek to constrain shears to the per cent level or better, a systematic error of this magnitude in shear calibration is quite serious.

Regarding possible biases in the morphological mixtures of galaxies due to overdensities or underdensities in the training sample, there are results in the literature for several methods that show how shear biases vary with morphology. For example, for the maximum likelihood fitting

code IM3SHAPE, figure 2 in Kacprzak et al. (2012) shows multiplicative biases for two-component Sérsic profile galaxies as a function of their bulge-to-total ratios (denoted there as $F_b/(F_b + F_d)$, which we will equate with our B/T). The shear calibration bias scales roughly like $0.04 - 0.05(B/T)$ as B/T goes from 0 to 1. Our results suggest that typical (median) B/T values may be influenced by cosmic variance in the COSMOS field leading to fluctuations of order 0.05. The resulting variation in the calibration bias would therefore be $\sim 2.5 \times 10^{-3}$, or 0.25 per cent shear calibration uncertainty. For existing datasets this is not very problematic, but for surveys like LSST, Euclid, and WFIRST-AFTA, this would be a dominant part of the systematic error budget. To give another example, for re-Gaussianization, figure 9 of Mandelbaum et al. (2012) shows that as Sérsic n goes from 1 to 6, the shear calibration bias varies by 2 per cent. In this case, since we have shown that the median value of Sérsic n can vary by ~ 0.4 due to morphology-density correlations, this suggests that the shear calibration for re-Gaussianization could be misestimated by $\sim 2 \times 10^{-3}$, or 0.2 per cent. This too is acceptable in existing datasets, but not those that will be used for shear estimation in the next decade.

A final consideration is the question of how applicable these results using volume-limited samples are to simulations of upcoming weak lensing surveys, which will exclusively use flux-limited samples. For our analysis, the volume-limiting sample was necessary to avoid complications due to varying galaxy populations in each redshift slice, allowing us to isolate purely environmental effects. In principle, if the morphology-density correlations that we have identified turn out to not exist for intrinsically fainter galaxy populations, then at low redshift (where a flux limited sample will include galaxies that are intrinsically much fainter than at high redshift), the effects will be less serious for upcoming lensing surveys. However, we do not have any particular reason to believe that these effects will vanish for fainter galaxies. Moreover, at higher redshift where only intrinsically bright galaxies can be seen, the effect should be present at a level similar to what we have found here. Since higher redshift galaxies tend to dominate cosmological shear estimates (due to their higher shears), our findings will be important to take into account.

In conclusion, our results have serious implications for the plans to create realistic image simulations that will be used to derive redshift-dependent shear calibrations for upcoming weak lensing surveys. If care is not taken to mitigate this effect, then the cosmic variance in the training sample may bias the conclusions regarding shear calibration for redshift slices that represent significant overdensities or underdensities compared to the typical galaxy population. This is particularly a problem when using the smaller *HST* surveys, where a single galaxy cluster or a void could completely dominate the galaxy population in a given redshift slice. To mitigate this problem, it will be imperative to (a) collect training data from widely separated patches on the sky, and (b) take care to use redshift slices that are brought enough that these effects are reduced, so as to wash out the effect of any signal overdensity or underdensity on the simulated galaxy population. By employing these mitigation schemes, there is every reason to believe that the effect we have identified can be reduced to a small component of the systematic error budget of major upcoming lensing surveys.

ACKNOWLEDGMENTS

AK and RM acknowledge the support of NASA ROSES 12-EUCLID12-0004, and program HST-AR-12857.01-A, provided by NASA through a grant from the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, Incorporated, under NASA contract NAS5-26555. RM acknowledges the support of an Alfred P. Sloan Research Fellowship. We thank Alexie Leauthaud for many useful discussions.

Need to fix arXiv references so that they show up properly with e-print number.

REFERENCES

Albrecht A. et al., 2006, ArXiv Astrophysics e-prints
Bartelmann M., Schneider P., 2001, Phys.Rep., 340, 291
Baugh C. M., Efstathiou G., 1993, MNRAS, 265, 145
Bernstein G. M., 2010, MNRAS, 406, 2793
Bernstein G. M., Armstrong R., 2014, MNRAS, 438, 1880
Bernstein G. M., Jarvis M., 2002, AJ, 123, 583
Bridle S. et al., 2010, MNRAS, 405, 2044
Bundy K. et al., 2006, ApJ, 651, 120
Carollo C. M. et al., 2014, ArXiv e-prints
Chabrier G., 2003, PASP, 115, 763
Coil A. L., Newman J. A., Kaiser N., Davis M., Ma C.-P., Kocevski D. D., Koo D. C., 2004, ApJ, 617, 765
de Jong J. T. A. et al., 2013, The Messenger, 154, 44
De Propris R. et al., 2014, MNRAS, 444, 2200
Faber S. M. et al., 2007, ApJ, 665, 265
Giallongo E., Salimbeni S., Menci N., Zamorani G., Fontana A., Dickinson M., Cristiani S., Pozzetti L., 2005, ApJ, 622, 116
Green J. et al., 2012, ArXiv e-prints
Heymans C. et al., 2013, MNRAS, 432, 2433
Hirata C., Seljak U., 2003, MNRAS, 343, 459
Ilbert O. et al., 2009, ApJ, 690, 1236
Jee M. J., Tyson J. A., Schneider M. D., Wittman D., Schmidt S., Hilbert S., 2013, ApJ, 765, 74
Kacprzak T., Zuntz J., Rowe B., Bridle S., Refregier A., Amara A., Voigt L., Hirsch M., 2012, MNRAS, 427, 2711
Kaiser N. et al., 2010, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 7733, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series
Kitching T. D. et al., 2012, MNRAS, 423, 3163
Kitching T. D., Miller L., Heymans C. E., van Waerbeke L., Heavens A. F., 2008, MNRAS, 390, 149
Koekemoer A. M. et al., 2007, ApJS, 172, 196
Kovač K. et al., 2010, ApJ, 708, 505
Lackner C. N., Gunn J. E., 2012, MNRAS, 421, 2277
Laureijs R. et al., 2011, ArXiv e-prints
Leauthaud A. et al., 2010, ApJ, 709, 97
Leauthaud A. et al., 2007, ApJS, 172, 219
LSST Science Collaboration et al., 2009, ArXiv e-prints
Mandelbaum R., Hirata C. M., Leauthaud A., Massey R. J., Rhodes J., 2012, MNRAS, 420, 1518

Mandelbaum R. et al., 2014, ApJS, 212, 5
Mandelbaum R., Slosar A., Baldauf T., Seljak U., Hirata C. M., Nakajima R., Reyes R., Smith R. E., 2013, MNRAS, 432, 1544
Melchior P., Böhnert A., Lombardi M., Bartelmann M., 2010, A&A, 510, A75
Melchior P., Viola M., 2012, MNRAS, 424, 2757
Miller L. et al., 2013, MNRAS, 429, 2858
Miller L., Kitching T. D., Heymans C., Heavens A. F., van Waerbeke L., 2007, MNRAS, 382, 315
Miyazaki S. et al., 2006, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 6269, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series
Refregier A., Kacprzak T., Amara A., Bridle S., Rowe B., 2012, MNRAS, 425, 1951
Rowe B. et al., 2014, ArXiv e-prints
Schrabback T. et al., 2010, A&A, 516, A63
Scoville N. et al., 2007, ApJS, 172, 1
The Dark Energy Survey Collaboration, 2005, ArXiv Astrophysics e-prints
Tomczak A. R. et al., 2014, ApJ, 783, 85
Troxel M. A., Ishak M., 2014, ArXiv e-prints
van Daalen M. P., Schaye J., Booth C. M., Dalla Vecchia C., 2011, MNRAS, 415, 3649
Voigt L. M., Bridle S. L., 2010, MNRAS, 404, 458
Weinberg D. H., Mortonson M. J., Eisenstein D. J., Hirata C., Riess A. G., Rozo E., 2013, Phys.Rep., 530, 87
Willmer C. N. A. et al., 2006, ApJ, 647, 853
Wolf C., Meisenheimer K., Rix H.-W., Borch A., Dye S., Kleinheinrich M., 2003, A&A, 401, 73