Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **474,** 2635–2678 (2018)
Advance Access publication 2017 October 14

doi:10.1093/mnras/stx2666

Cluster mass calibration at high redshift: *HST* weak lensing analysis of 13 distant galaxy clusters from the South Pole Telescope Sunyaev–Zel'dovich Survey

T. Schrabback,[1,2,3★] D. Applegate,[1,4] J. P. Dietrich,[5,6] H. Hoekstra,[7] S. Bocquet,[4,5,6,8] A. H. Gonzalez,[9] A. von der Linden,[2,3,10,11] M. McDonald,[12] C. B. Morrison,[1,13] S. F. Raihan,[1] S. W. Allen,[2,3,14] M. Bayliss,[15,16,17] B. A. Benson,[4,18,19] L. E. Bleem,[4,8,20] I. Chiu,[5,6,21] S. Desai,[5,6,22] R. J. Foley,[23] T. de Haan,[24,25] F. W. High,[4,19] S. Hilbert,[5,6] A. B. Mantz,[2,3] R. Massey,[26] J. Mohr,[5,6,27] C. L. Reichardt,[28] A. Saro,[5,6] P. Simon,[1] C. Stern,[5,6] C. W. Stubbs[15,16] and A. Zenteno[29]

Affiliations are listed at the end of the paper

Accepted 2017 October 10. Received 2017 October 10; in original form 2016 October 28

ABSTRACT

We present an *HST*/Advanced Camera for Surveys (ACS) weak gravitational lensing analysis of 13 massive high-redshift ($z_{\rm median} = 0.88$) galaxy clusters discovered in the South Pole Telescope (SPT) Sunyaev–Zel'dovich Survey. This study is part of a larger campaign that aims to robustly calibrate mass–observable scaling relations over a wide range in redshift to enable improved cosmological constraints from the SPT cluster sample. We introduce new strategies to ensure that systematics in the lensing analysis do not degrade constraints on cluster scaling relations significantly. First, we efficiently remove cluster members from the source sample by selecting very blue galaxies in $V-I$ colour. Our estimate of the source redshift distribution is based on Cosmic Assembly Near-infrared Deep Extragalactic Legacy Survey (CANDELS) data, where we carefully mimic the source selection criteria of the cluster fields. We apply a statistical correction for systematic photometric redshift errors as derived from *Hubble* Ultra Deep Field data and verified through spatial cross-correlations. We account for the impact of lensing magnification on the source redshift distribution, finding that this is particularly relevant for shallower surveys. Finally, we account for biases in the mass modelling caused by miscentring and uncertainties in the concentration–mass relation using simulations. In combination with temperature estimates from *Chandra* we constrain the normalization of the mass–temperature scaling relation $\ln(E(z)M_{500c}/10^{14}\,{\rm M}_\odot) = A + 1.5\ln(kT/7.2\,{\rm keV})$ to $A = 1.81^{+0.24}_{-0.14}({\rm stat.}) \pm 0.09({\rm sys.})$, consistent with self-similar redshift evolution when compared to lower redshift samples. Additionally, the lensing data constrain the average concentration of the clusters to $c_{200c} = 5.6^{+3.7}_{-1.8}$.

Key words: gravitational lensing: weak – galaxies: clusters: general – cosmology: observations.

1 INTRODUCTION

Constraints on the number density of clusters as a function of their mass and redshift probe the growth of structure in the Universe, therefore holding great promise to constrain cosmological models (e.g. Haiman, Mohr & Holder 2001; Allen, Evrard & Mantz 2011; Weinberg et al. 2013). Previous studies using samples of at most a few hundred clusters have delivered some of the tightest cosmological constraints currently available on dark energy properties, theories of modified gravity and the species-summed neutrino mass (e.g. Rapetti et al. 2009, 2013; Schmidt, Vikhlinin & Hu 2009; Vikhlinin et al. 2009b; Mantz et al. 2010, 2015; Bocquet et al. 2015; de Haan et al. 2016). Recently, cosmic microwave background (CMB) experiments have begun to substantially increase the number of massive, high-redshift clusters found with well-characterized

★E-mail: schrabba@astro.uni-bonn.de

selection functions, detected via their Sunyaev–Zel'dovich (SZ; Sunyaev & Zel'dovich 1970, 1972) signature from inverse Compton scattering off the electrons in the hot cluster plasma (Hasselfield et al. 2013; Bleem et al. 2015; Planck Collaboration XXVII 2016a). Upcoming experiments such as SPT-3G (Benson et al. 2014) and eROSITA (Merloni et al. 2012) are expected to soon provide samples of 10^4–10^5 massive clusters with well-characterized selection functions, yielding a statistical constraining power that may mark the transition between 'Stage III' and 'Stage IV' dark energy constraints (see Albrecht et al. 2006) from clusters if systematic uncertainties are well controlled.

Cluster observables such as X-ray luminosity, SZ signal or optical/NIR richness and luminosity have been shown to scale with mass (e.g. Reiprich & Böhringer 2002; Lin, Mohr & Stanford 2004; Andersson et al. 2011). In order to adequately exploit the statistical constraining power of large cluster surveys, an accurate and precise calibration of the scaling relations between such mass proxies and mass is needed. Already for current surveys cosmological constraints are primarily limited by uncertainties in the calibration of mass–observable scaling relations (e.g. Rozo et al. 2010; Sehgal et al. 2011; Benson et al. 2013; von der Linden et al. 2014b; Mantz et al. 2015; Planck Collaboration XXIV 2016c). It is therefore imperative to improve this calibration empirically. In this context our work focuses especially on calibrating mass–observable relations at high redshifts, which together with low-redshift measurements, provides constraints on their redshift evolution. Particularly for constraints on dark energy properties, which are primarily derived from the redshift evolution of the cluster mass function, it is critical to ensure that systematic errors in the evolution of mass–observable scaling relations do not mimic the signature of dark energy. Most previous cosmological cluster studies had to rely on priors for the redshift evolution derived from numerical cluster simulations (e.g. Vikhlinin et al. 2009b; Benson et al. 2013; de Haan et al. 2016). It is crucial to test the assumed models of cluster astrophysics in these simulations by comparing their predictions to observational constraints on the scaling relations (e.g. Le Brun et al. 2014), and to shrink the uncertainties on the scaling relation parameters.

Progress in the field critically requires improvements in the cluster mass calibration through large multiwavelength follow-up campaigns. For example, high-resolution X-ray observations provide mass proxies with low intrinsic scatter, which can be used to constrain the relative masses of clusters (e.g. Vikhlinin et al. 2009a; Andersson et al. 2011; Reichert et al. 2011). On the other hand, weak gravitational lensing has been recognized as the most direct technique for the absolute calibration of the normalization of cluster mass–observable relations (Allen et al. 2011; Hoekstra et al. 2013; Applegate et al. 2014; Mantz et al. 2015). The main observable is the weak lensing reduced shear, a tangential distortion caused by the projected tidal gravitational field of the foreground mass distribution. It is directly related to the differential projected cluster mass distribution, and can be estimated from the observed shapes of background galaxies (e.g. Bartelmann & Schneider 2001; Schneider 2006).

To date, the majority of cluster weak lensing mass estimates have been obtained for lower redshift clusters ($z \lesssim 0.6$–0.7) using ground-based observations (e.g. High et al. 2012; Israel et al. 2012; Oguri et al. 2012; Applegate et al. 2014; Gruen et al. 2014; Umetsu et al. 2014; Ford et al. 2015; Hoekstra et al. 2015; Kettula et al. 2015; Battaglia et al. 2016; Lieu et al. 2016; Okabe & Smith 2016; van Uitert et al. 2016; Melchior et al. 2017; Simet et al. 2017). To constrain the evolution of cluster mass–observable scaling relations, these measurements need to be complemented with constraints for higher redshift clusters. Here, ground-based measurements suffer from low densities of sufficiently resolved background galaxies with robust shape measurements. This can be overcome using high-resolution *Hubble Space Telescope* (*HST*) images, where so far Jee et al. (2011) present the only weak lensing constraints for the cluster mass calibration of a large sample of massive high-redshift ($0.83 \leq z \leq 1.46$) clusters, which were drawn from optically, NIR- and X-ray-selected samples. Interestingly, their results suggest a possible evolution in the M_{2500c}–T_X scaling relation in comparison to self-similar extrapolations from low redshifts, with lower masses at the 20−30 per cent level. *HST* weak lensing measurements have also been used to constrain mass–observable scaling relations for lower (Leauthaud et al. 2010) and intermediate mass clusters (Hoekstra et al. 2011a).

This paper is part of a larger effort to obtain improved observational constraints on the calibration of cluster masses as a function of redshift. Here we analyse new *HST* observations of 13 massive high-z clusters detected by the South Pole Telescope (SPT; Carlstrom et al. 2011) via the SZ effect. This constitutes the first high-z sample of clusters with *HST* weak lensing observations which were drawn from a single, well-characterized survey selection function. As a major part of this paper, we carefully investigate and account for the relevant sources of systematic uncertainty in the weak lensing mass analysis, and discuss their relevance for future studies of larger samples.

The primary technical challenges for weak lensing studies are accurate measurements of galaxy shapes from noisy data in the presence of instrumental distortions, and the need for an accurate knowledge of the source redshift distribution which enters through the geometric lensing efficiency. Within the weak lensing community substantial progress has been made on the former issue through the development of improved shape measurement algorithms tested using image simulations (e.g. Miller et al. 2013; Hoekstra et al. 2015; Bernstein et al. 2016; Fenech Conti et al. 2017). For the latter issue, previous studies have typically estimated the redshift distribution from photometric redshifts (photo-zs) given the incompleteness of spectroscopic redshift samples (spec-zs) at the relevant magnitudes, requiring that the photo-z-based estimates are sufficiently accurate. If sufficient wavelength coverage is available, photo-zs can be estimated directly for the weak lensing survey fields of interest (used in the cluster context e.g. by Leauthaud et al. 2010; Applegate et al. 2014; Ford et al. 2015). Otherwise, photo-zs can be used from external reference deep fields, requiring that statistically consistent and sufficiently representative galaxy populations are selected in both the survey and reference fields. For cluster weak lensing studies both approaches are complicated by the fact that the presence of a cluster means that the corresponding line of sight is overdense at the cluster redshift, while both the default priors of photo-z codes and the reference deep fields ought to be representative for the cosmic mean distribution. Previous studies employing reference fields have typically dealt with this issue by applying colour selections ('colour cuts') that remove galaxies at the cluster redshift (e.g. High et al. 2012; Hoekstra et al. 2012; Okabe & Smith 2016). In case of incomplete removal the approach can be complemented by a statistical correction for the residual cluster member contamination if that can be estimated sufficiently well (e.g. Hoekstra et al. 2015). For cluster weak lensing studies a further complication arises when parametric models are fitted to the measured tangential reduced shear profiles, as issues such as miscentring (e.g. Johnston et al. 2007; George et al. 2012) or uncertainties regarding assumed cluster concentrations can lead to

non-negligible biases, introducing the need for calibrations using simulations (e.g. Becker & Kravtsov 2011).

This paper is organized as follows: Section 2 summarizes relevant aspects of weak lensing theory. This is followed by a description of our cluster sample in Section 3 and a description of the analysed data and image processing in Section 4. Section 5 details on the weak lensing shape measurements and a new test for signatures of potential residuals of charge-transfer inefficiency (CTI) in the weak lensing catalogues. In Section 6 we describe in detail our approach to remove cluster galaxies via colour cuts and reliably estimate the source redshift distribution using data from the Cosmic Assembly Near-infrared Deep Extragalactic Legacy Survey (CANDELS) fields. In Section 7 we present our weak lensing shear profile analysis, mass reconstructions and mass estimates, which we use in Section 8 to constrain the mass–temperature scaling relation. Finally, we discuss our findings in Section 9 and conclude in Section 10.

Throughout this paper we assume a standard flat Λ cold dark matter (CDM) cosmology characterized by $\Omega_\mathrm{m} = 0.3$, $\Omega_\Lambda = 0.7$ and $H_0 = 70h_{70}\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ with $h_{70} = 1$, as approximately consistent with recent CMB constraints (Hinshaw et al. 2013; Planck Collaboration XIII 2016b). For the computation of large-scale structure noise on the weak lensing estimates and the concentration–mass relation according to Diemer & Kravtsov (2015) we furthermore assume $\sigma_8 = 0.8$, $\Omega_\mathrm{b} = 0.046$ and $n_\mathrm{s} = 0.96$. All magnitudes are in the AB system and are corrected for extinction according to Schlegel, Finkbeiner & Davis (1998).

2 SUMMARY OF RELEVANT WEAK LENSING THEORY

The images of distant background galaxies are distorted by the tidal gravitational field of a foreground mass concentration, see e.g. the reviews by Bartelmann & Schneider (2001) and Schneider (2006), as well as Hoekstra et al. (2013) in the context of galaxy clusters. In the weak lensing regime the size of a source is much smaller than the characteristic scale on which variations in the tidal field occur. In this case the lens mapping as a function of observed position $\boldsymbol{\theta}$ can be described using the reduced shear $g(\boldsymbol{\theta})$ and the convergence $\kappa(\boldsymbol{\theta}) = \Sigma(\boldsymbol{\theta})/\Sigma_\mathrm{crit}$, which is the ratio of the surface mass density $\Sigma(\boldsymbol{\theta})$ and the critical surface mass density

$$\Sigma_\mathrm{crit} = \frac{c^2}{4\pi G}\frac{1}{D_\mathrm{l}\beta}, \tag{1}$$

with the speed of light c, the gravitational constant G and the geometric lensing efficiency

$$\beta = \max\left[0, \frac{D_\mathrm{ls}}{D_\mathrm{s}}\right], \tag{2}$$

where D_s, D_l and D_ls indicate the angular diameter distances to the source, to the lens, and between lens and source, respectively. The reduced shear

$$g(\boldsymbol{\theta}) = \frac{\gamma(\boldsymbol{\theta})}{1-\kappa(\boldsymbol{\theta})} \tag{3}$$

describes the observable anisotropic shape distortion due to weak lensing. It is a two-component quantity, conveniently written as a complex number

$$g = g_1 + \mathrm{i}g_2 = |g|\mathrm{e}^{2\mathrm{i}\varphi}, \tag{4}$$

where $|g|$ constitutes the strength of the distortion and φ its orientation with respect to the coordinate system. The reduced shear $g(\boldsymbol{\theta})$ is a rescaled version of the unobservable shear $\gamma(\boldsymbol{\theta})$, and can be estimated from the ensemble-averaged PSF-corrected ellipticities $\epsilon = \epsilon_1 + \mathrm{i}\epsilon_2$ of background galaxies (see Section 5), with the expectation value

$$\langle\epsilon\rangle = g. \tag{5}$$

Due to noise from the intrinsic galaxy shape distribution and measurement noise we need to average the ellipticities of a large ensemble of galaxies

$$\langle\epsilon_\alpha\rangle = \frac{\sum \epsilon_{\alpha,i} w_i}{\sum w_i} \tag{6}$$

to obtain useful constraints, where $\alpha \in \{1, 2\}$ indicates the two ellipticity components and i indicates galaxy i. The shape weights $w_i = 1/\sigma^2_{\epsilon,i}$ are included to improve the measurement signal-to-noise ratio, where $\sigma_{\epsilon,i}$ contains contributions both from the measurement noise and the intrinsic shape distribution (see Appendix A, where we constrain both contributions empirically using CANDELS data).

It is often useful to decompose the shear, reduced shear and the ellipticity into their tangential components, e.g. g_t, and cross components, e.g. $g_\times$, with respect to the centre of a mass distribution as

$$g_\mathrm{t} = -g_1\cos 2\phi - g_2 \sin 2\phi, \tag{7}$$

$$g_\times = +g_1\sin 2\phi - g_2 \cos 2\phi, \tag{8}$$

where ϕ is the azimuthal angle with respect to the centre. The azimuthal average of the tangential shear γ_t at a radius r around the centre of the mass distribution is linked to the mean convergence $\bar{\kappa}(<r)$ inside r and $\bar{\kappa}(r)$ at r via

$$\langle\gamma_\mathrm{t}\rangle(r) = \bar{\kappa}(<r) - \bar{\kappa}(r). \tag{9}$$

The weak lensing convergence and shear scale for an individual source galaxy at redshift z_i with the geometric lensing efficiency $\beta(z_i)$, which is often conveniently written as

$$\gamma = \beta_\mathrm{s}(z_i)\gamma_\infty,\ \kappa = \beta_\mathrm{s}(z_i)\kappa_\infty, \tag{10}$$

where κ_∞ and γ_∞ correspond to the values for a source at infinite redshift, and $\beta_\mathrm{s}(z_i) = \beta(z_i)/\beta_\infty$. In practice, we average the ellipticities of an ensemble of galaxies distributed in redshift, providing an estimate for

$$\langle g\rangle = \left\langle \frac{\beta_\mathrm{s}(z_i)\gamma_\infty}{1-\beta_\mathrm{s}(z_i)\kappa_\infty}\right\rangle. \tag{11}$$

While one could in principle compute the exact model prediction for this from the source redshift distribution weighted by the lensing weights, a sufficiently accurate approximation is provided in Hoekstra, Franx & Kuijken (2000):

$$g^\mathrm{model} \simeq \left[1 + \left(\frac{\langle\beta_\mathrm{s}^2\rangle}{\langle\beta_\mathrm{s}\rangle^2} - 1\right)\langle\beta_\mathrm{s}\rangle\kappa_\infty^\mathrm{model}\right]\frac{\langle\beta_\mathrm{s}\rangle\gamma_\infty^\mathrm{model}}{1-\langle\beta_\mathrm{s}\rangle\kappa_\infty^\mathrm{model}} \tag{12}$$

(see also Seitz & Schneider 1997; Applegate et al. 2014), where

$$\langle\beta_\mathrm{s}\rangle = \frac{\sum\beta_\mathrm{s}(z_i)w_i}{\sum w_i},\ \langle\beta_\mathrm{s}^2\rangle = \frac{\sum\beta_\mathrm{s}^2(z_i)w_i}{\sum w_i} \tag{13}$$

need to be computed from the estimated source redshift distribution, taking the shape weights into account.

Table 1. The cluster sample.

Cluster name	z_l	ξ	Coordinates centres (deg J2000)						$M_{500c,SZ}$	Sample
			SZ α	SZ δ	X-ray α	X-ray δ	BCG α	BCG δ	$(10^{14}\,M_\odot\,h_{70}^{-1})$	
SPT-CL J0000−5748	0.702	8.49	0.2499	−57.8064	0.2518	−57.8094	0.2502	−57.8093	4.56 ± 0.80	V10
SPT-CL J0102−4915	0.870	39.91	15.7294	−49.2611	15.7350	−49.2667	15.7407	−49.2720	14.43 ± 2.10	W11
SPT-CL J0533−5005	0.881	7.08	83.4009	−50.0901	83.4018	−50.0969	83.4144	−50.0845	3.79 ± 0.73	V10
SPT-CL J0546−5345	1.066	10.76	86.6525	−53.7625	86.6532	−53.7604	86.6569	−53.7586	5.05 ± 0.82	V10
SPT-CL J0559−5249	0.609	10.64	89.9251	−52.8260	89.9357	−52.8253	89.9301	−52.8241	5.78 ± 0.95	V10
SPT-CL J0615−5746	0.972	26.42	93.9650	−57.7763	93.9652	−57.7788	93.9656	−57.7802	10.53 ± 1.55	W11
SPT-CL J2040−5725	0.930	6.24	310.0573	−57.4295	310.0631[a]	−57.4287	310.0552	−57.4209	3.36 ± 0.70	R13
SPT-CL J2106−5844	1.132	22.22	316.5206	−58.7451	316.5174	−58.7426	316.5192	−58.7411	8.35 ± 1.24	W11
SPT-CL J2331−5051	0.576	10.47	352.9608	−50.8639	352.9610	−50.8631	352.9631	−50.8650	5.60 ± 0.92	V10
SPT-CL J2337−5942	0.775	20.35	354.3523	−59.7049	354.3516	−59.7061	354.3650	−59.7013	8.43 ± 1.27	V10, W11
SPT-CL J2341−5119	1.003	12.49	355.2991	−51.3281	355.3009	−51.3285	355.3014	−51.3291	5.59 ± 0.89	V10
SPT-CL J2342−5411	1.075	8.18	355.6892	−54.1856	355.6904	−54.1838	355.6913	−54.1848	3.93 ± 0.70	V10
SPT-CL J2359−5009	0.775	6.68	359.9230	−50.1649	359.9321	−50.1697	359.9324	−50.1722	3.60 ± 0.71	V10

Note. Basic data from Bleem et al. (2015) and Chiu et al. (2016a) for the 13 clusters targeted in this weak lensing analysis. *Column 1:* Cluster designation. *Column 2:* Spectroscopic cluster redshift. *Column 3:* Peak signal-to-noise ratio of the SZ detection. *Columns 4–9:* Right ascension α and declination δ of the cluster centres used in the weak lensing analysis from the SZ peak, X-ray centroid and BCG position. *Column 10:* Mass derived from the SZ-Signal. *Column 11:* SPT parent sample for *HST* follow-up selection.
[a]X-ray centroid from *XMM–Newton* data, otherwise *Chandra* (see Section 8).

When the signal of lenses at different redshifts is compared or stacked, it can be useful to conduct the analysis in terms of the differential surface mass density

$$\Delta\Sigma(r) = \frac{\sum_i w_i \, (\epsilon_t \Sigma_{\rm crit})_i}{\sum_i w_i} \tag{14}$$

to compensate for the redshift dependence of the signal, where the summation is conducted over sources in a separation interval around r.

Gravitational lensing leaves the surface brightness invariant. Accordingly, a relative change in the observed flux of a source due to lensing is solely given by the relative magnification of the source

$$\mu = \frac{1}{(1-\kappa)^2 - |\gamma|^2}. \tag{15}$$

Together with the change in solid angle this also changes the observed density of background sources and their redshift distribution, as investigated in Section 6.7.

3 THE CLUSTER SAMPLE

We study a total of 13 distant galaxy clusters detected by the SPT in the redshift range $0.57 \leq z \leq 1.13$ via the SZ effect; see Table 1 for details and Fig. 1 for a comparison of the cluster redshift distribution to recent large weak lensing cluster samples from the Canadian Cluster Comparison Project (Hoekstra et al. 2015), Weighing the Giants (von der Linden et al. 2014a), the Cluster Lensing And Supernova survey with *Hubble* (Umetsu et al. 2014), the Local Cluster Substructure Survey (Okabe & Smith 2016), and the analysis of *HST* observations of X-ray, optically and NIR-selected high-redshift clusters by Jee et al. (2011).

The SPT clusters were observed in *HST* Cycles 18 and 19. At the time of the target selection, the SPT cluster follow-up campaign was still incomplete. From the clusters with measured spectroscopic redshifts prior to the corresponding cycle, we selected the most massive SPT-SZ clusters at $0.6 \lesssim z \lesssim 1.0$ for the Cycles 18 programme, and the most massive clusters at $z \gtrsim 0.9$ for the Cycle 19 programme. Nine clusters in our overall sample originate from the first 178 deg^2 of the sky surveyed by SPT (Vanderlinde et al. 2010, hereafter V10). Using updated estimates

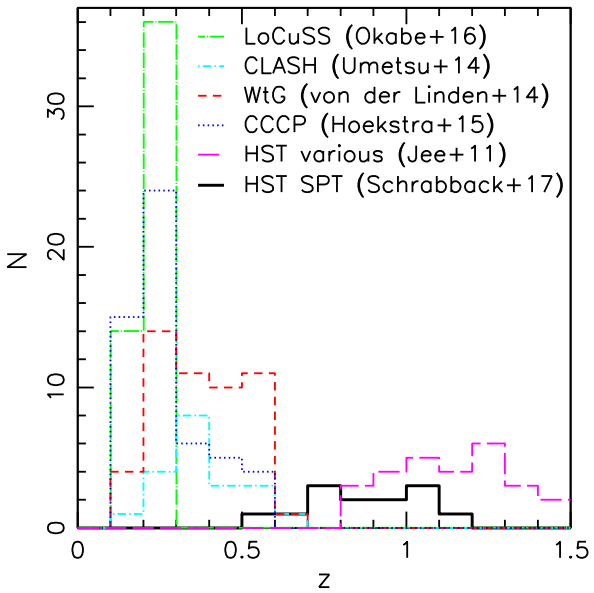

Figure 1. Comparison of the cluster redshift distribution of our sample with several recent independent studies, plus the larger high-redshift sample from Jee et al. (2011), which includes a combination of optically, NIR- and X-ray-selected clusters.

of the SZ detection significance ξ from the cluster catalogue for the full 2500 deg^2 SPT-SZ survey (Bleem et al. 2015, hereafter B15), our selection of clusters from the V10 sample includes all clusters from the first 178 deg^2 at $z \geq 0.57$ with $\xi \geq 8$ plus all clusters at $z \geq 0.70$ with $\xi \geq 6.6$ (see Table 1), except for SPT-CL J0540 −5744 ($\xi = 6.74$). Additionally, our sample includes all clusters at $z \geq 0.70$ from Williamson et al. (2011, henceforth W11), who present a catalogue of the 26 most significant SZ cluster detections in the full 2500 deg^2 SPT survey region. This adds three clusters in addition to SPT-CL J2337−5942, which is part of both samples. Finally, with SPT-CL J2040−5725 a single further cluster is included from Reichardt et al. (2013, hereafter R13), who present the cluster sample constructed from the first 720 deg^2 of the SPT cluster survey. In addition to the aforementioned sample papers, more detailed

studies of individual clusters were published for SPT-CL J0546−5345 (Brodwin et al. 2010) and SPT-CL J2106−5844 (Foley et al. 2011). Spectroscopic cluster redshift measurements are described in Ruel et al. (2014) and Bayliss et al. (2016). In Table 1 we also list X-ray centroids as estimated from the available *Chandra* or *XMM–Newton* data (detailed in Andersson et al. 2011; Benson et al. 2013; McDonald et al. 2013; Chiu et al. 2016a, see also Section 8), and BCG positions from Chiu et al. (2016a).

4 DATA AND DATA REDUCTION

In this section we provide details on the data analysed in this study and their reduction. For the SPT clusters we make use of *HST* observations (Section 4.1.1) for shape and colour measurements, as well as Very Large Telescope (VLT) observations (Section 4.2) for colour measurements in the outer cluster regions. To optimize our weak lensing pipeline, and to be able to apply consistent selection criteria to photo-z catalogues from Skelton et al. (2014), we also process *HST* observations of the CANDELS fields (Section 4.1.3).

4.1 *HST* /ACS data

4.1.1 SPT cluster observations

We measure weak lensing galaxy shapes from high-resolution *HST* imaging obtained during Cycles 18 and 19 as part of programmes 12246 (PI: C. Stubbs) and 12477[1] (PI: F. W. High), and observed between 2011 September 29 and 2012 October 24 under low sky background conditions. Each cluster was observed with a 2 × 2 Advanced Camera for Surveys (ACS)/WFC mosaic in the *F*606*W* filter, where each tile consists of four dithered exposures of 480 s, adding to a total exposure time of 1.92 ks per tile. These mosaic observations allow us to probe the cluster weak lensing signal out to approximately the virial radius. Additionally, a single tile was observed with ACS in the *F*814*W* filter on the cluster centre (1.92 ks). These data are included in our photometric analysis (Section 6). For the weak lensing shape measurements we chose observations in the *F*606*W* filter as it is the most efficient ACS filter in terms of weak lensing galaxy source density (see e.g. Schrabback et al. 2007). However note that our analysis in Appendix A4 suggests that future programmes could benefit from mosaic observations in both *F*606*W* and *F*814*W* to simultaneously obtain robust shape measurements and colour estimates. In fact, a 2 × 2 *F*814*W* ACS mosaic was obtained for one of the clusters in our sample, SPT-CL J0615−5746, through the independent *HST* programme 12757 (PI: Mazzotta), with observations conducted 2012 January 19–22. For the current analysis we include these additional data in the colour measurements but not the shape analysis.

We denote magnitudes measured from the ACS *F*606*W* and *F*814*W* images as V_{606} and I_{814}, respectively. By default these correspond to magnitudes measured in circular apertures with a diameter of 0.7 arcsec unless explicitly stated differently.

4.1.2 HST data reduction

For basic image reductions we largely employ the standard ACS calibration pipeline CALACS. The main exception is our use of the Massey et al. (2014, M14 henceforth) algorithm for the correction of CTI. CTI constitutes an important systematic effect for *HST* weak lensing shape analyses if left uncorrected (e.g. Rhodes et al. 2007; Schrabback et al. 2010, S10 henceforth). It is caused by radiation damage in space. The resulting CCD defects act as charge traps during the read-out process, introducing non-linear charge-trails behind objects in the parallel-transfer read-out direction. M14 updated their time-dependent model of the charge trap densities by fitting charge trails behind hot pixels in CANDELS ACS/*F*606*W* imaging exposures of the COSMOS field (Grogin et al. 2011), which were obtained at a similar epoch as our cluster data (between 2011 December 6 and 2012 April 15). Given that we conduct the CTI correction using the M14 code, we also have to CTI-correct the master dark frames using this pipeline. As further differences to standard CALACS processing we compute accurately normalized r.m.s. noise maps as detailed in S10 and optimize the bad pixel mask, where we flag satellite trails and cosmic ray clusters, and unflag the removed CTI trails of hot pixels.

The further data reduction for the individual ACS tiles closely follows S10, to which we refer the reader for details. As the first step, we carefully refine relative shifts and rotations between the exposures by matching the positions of compact objects. We then use MULTIDRIZZLE (Koekemoer et al. 2003) for the cosmic ray removal and stacking, where we employ the `lanczos3` kernel at the native pixel scale 0.05 arcsec to minimize noise correlations while only introducing a low level of aliasing for ellipticity measurements (Jee et al. 2007). The pipeline also generates correctly scaled r.m.s. noise maps for stacks that are used for the object detection. We conduct weak lensing shape measurements on these individual stacked ACS tiles (see Section 5).

For the joint photometric analysis with available VLT data (Section 6.4 with details given in Appendix D) we additionally generate stacks for the 2 × 2 ACS mosaics. Here we iteratively align neighbouring tiles by first resampling them separately on to a common pixel grid, only stacking the exposures of the corresponding tile. We then use the differences between the positions of matched objects in the overlapping regions to compute shifts and rotations, in order to update the astrometry.

4.1.3 CANDELS HST data

When estimating the redshift distribution of our source sample (see Section 6) we need to apply the same selection function (consisting of photometric, shape and size cuts) to the galaxies in the CANDELS fields, which act as our reference sample. To be able to employ consistent weak lensing cuts, we reduce and analyse ACS imaging in the CANDELS fields with the same pipeline as the *HST* observations of the SPT clusters. This includes data from the CANDELS (Grogin et al. 2011, Proposal IDs 12440, 12064), GOODS (Giavalisco et al. 2004, Proposal IDs 9425, 9583), GEMS (Rix et al. 2004, Proposal ID 9500) and AEGIS (Davis et al. 2007, Proposal ID 10134) programmes. Here we perform a tile-wise analysis, always stacking exposures with good spatial overlap which add to approximately 1-orbit depth, roughly matching the depth of our cluster field data (see Appendix A2 for additional information).

We use these blank field data also as a calibration sample to derive an empirical weak lensing weighting scheme that is based on the measured ellipticity dispersion as a function of logarithmic signal-to-noise ratio and employed in our cluster lensing analysis (see Appendix A5). This analysis also provides updated constraints on the dispersion of the intrinsic galaxy ellipticities and allows

[1] This programme also includes observations of SPT-CL J0205−5829 ($z = 1.322$). However, we do not include it in the current analysis given its high redshift, which would require deeper z-band observations for the background selection (see Section 6) than currently available.

Table 2. The VLT/FORS2 $I_{\rm FORS2}$ imaging data.

Cluster name	$t_{\rm exp}$ (ks)	$I_{\rm lim}$	IQ (arcsec)	Used V_{606} range Bright cut	Faint cut
SPT-CL J0000−5748	2.1	26.0	0.65	24.0–25.5	25.5–26.0
SPT-CL J0102−4915	2.1	25.8	0.75	24.0–25.0	25.0–25.5
SPT-CL J0533−5005	2.1	25.8	0.73	24.0–25.5	–
SPT-CL J0546−5345	2.1	25.7	0.75	24.0–25.0	25.0–25.5
SPT-CL J0559−5249	1.9	25.6	0.65	24.0–25.0	25.0–25.5
SPT-CL J0615−5746	2.5	25.6	0.93	24.0–24.5	24.5–25.5
SPT-CL J2040−5725	2.9	25.7	0.70	24.0–25.0	25.0–25.5
SPT-CL J2106−5844	4.8	25.8	0.80	24.0–25.0	25.0–25.5
SPT-CL J2331−5051	2.4	25.9	0.83	24.0–25.5	25.5–26.0
SPT-CL J2337−5942	2.1	25.7	0.80	24.0–25.5	25.5–26.0
SPT-CL J2341−5119	2.1	25.8	0.80	24.0–25.5	25.5–26.0
SPT-CL J2342−5411	2.1	25.7	0.93	24.0–25.0	25.0–25.5
SPT-CL J2359−5009	2.1	25.9	0.68	24.0–25.5	25.5–26.0

Note. Details of the analysed VLT/FORS2 imaging data. *Column 1:* Cluster designation. *Column 2:* Total co-added exposure time. *Column 3:* 5σ-limiting magnitude computed for 1.5 arcsec apertures in the stack from the single pixel noise r.m.s. values of the contributing exposures. *Column 4:* Image Quality defined as 2 × FLUX_RADIUS from SOURCE EXTRACTOR. *Column 5:* V_{606} magnitude range with low photometric colour scatter $\sigma_{\Delta(V-I)} < 0.2$, for which the 'bright' colour cut is applied (see Table D1 in Appendix D). *Column 6:* V_{606} magnitude range with increased photometric colour scatter $0.2 < \sigma_{\Delta(V-I)} < 0.3$, for which the 'faint' colour cut is applied (see Table D1 in Appendix D).

us to compare the weak lensing performance of the ACS $F606W$ and $F814W$ filters, aiding the preparation of future weak lensing programmes (see Appendix A4).

4.2 VLT/FORS2 data

For our analysis we make use of VLT/FORS2 imaging of all of our targets taken as part of programmes 086.A-0741 (PI: Bazin), 088.A-0796 (PI: Bazin), 088.A-0889 (PI: Mohr) and 089.A-0824 (PI: Mohr) in the $I_{\rm BESS}$ pass-band, which we call $I_{\rm FORS2}$. The FORS2 focal plane is covered with two $2k \times 4k$ MIT CCDs. The data were taken with the standard resolution collimator in 2×2 binning, providing imaging over a 6.8 arcmin × 6.8 arcmin field of view with a pixel scale of 0.25 arcsec, matching the size of our ACS mosaics well.

We reduced the data using THELI (Erben et al. 2005; Schirmer 2013), applying bias and flat-field correction, relative photometric calibration and sky background subtraction using SOURCE EXTRACTOR (Bertin & Arnouts 1996). We use the object positions in the *HST* $F606W$ image as an astrometric reference for the distortion correction. For an initial absolute photometric calibration using the stars located in the central *HST* I_{814} tile we employ the relation

$$I_{\rm FORS2} - I_{814} = -0.052 + 0.0095(V_{606} - I_{814}), \quad (16)$$

which was derived employing the Pickles (1998) stellar library. This relation is valid for $V_{606} - I_{814} < 1.7$ and assumes total magnitudes for the computation of $I_{\rm FORS2} - I_{814}$. We list total exposure times, limiting magnitudes and delivered image quality for the co-added images in Table 2. For further details on the data reduction see Chiu et al. (2016a), who also analyse observations obtained with FORS2 in the $B_{\rm HIGH}$ and $z_{\rm GUNN}$ pass-bands. In our analysis we do not include these additional bands. Our initial testing indicates that their inclusion would only yield a minor increase in the usable background galaxy source density given the depth of the different observations and typical colours of the dominant background source population.

5 WEAK LENSING GALAXY SHAPES

5.1 Shape measurements

For the generation of weak lensing shape catalogues we employ the pipeline from S10, which was successfully used for cosmological weak lensing measurements that typically have more stringent requirements on the control of systematics than cluster weak lensing studies. We refer the reader to this publication for a more detailed pipeline description. Here we summarize the main steps and provide details on recent changes to our pipeline only. One of the main changes is the application of the pixel-based CTI correction from M14 (Section 4.1.2), which is more accurate than the catalogue-level correction employed in S10. This change has become necessary as we analyse more recent ACS data with stronger CTI degradation.

As the first step in the catalogue generation we use SOURCE EXTRACTOR (Bertin & Arnouts 1996) to detect objects in the $F606W$ stacks and measure basic object properties. For the ellipticity measurement and correction for the point-spread function (PSF) we employ the KSB+ formalism (Kaiser, Squires & Broadhurst 1995; Luppino & Kaiser 1997; Hoekstra et al. 1998) as implemented by Erben et al. (2001) with modifications from Schrabback et al. (2007) and S10. We interpolate the spatially and temporally varying ACS PSF using a model derived from a principal component analysis of PSF variations in dense stellar fields. S10 showed that the dominant contribution to ACS PSF ellipticity variations can be described with a single principal component (related to the *HST* focus position). This one-parameter PSF model is sufficiently well constrained by the ∼10–20 high-S/N stars available for PSF measurements in extragalactic ACS pointings. We obtain a PSF model for each contributing exposure based on stellar ellipticity and size measurements in the image prior to resampling (to minimize noise), from which we compute the combined model for the stack. For the current work we recalibrated this algorithm using archival ACS $F606W$ stellar field observations taken after Servicing Mission 4. We processed these data with the same CTI correction method as our cluster field data.

Following S10 we select galaxies in terms of their half-light radius $r_{\rm h} > 1.2 r_{\rm h}^{*,\rm max}$, where $r_{\rm h}^{*,\rm max}$ is the upper limit of the 0.25 pixel wide stellar locus, and 'pre-seeing' shear polarizability tensor P^g with $\mathrm{Tr}[P^g]/2 > 0.1$. Deviating from S10 we exclude very extended galaxies with $r_{\rm h} > 7$ pixels, as they are poorly covered by the employed postage stamps. As done in S10 we mask galaxies close to the image boundaries, large galaxies or bright stars.

S10 introduced an empirical correction for noise bias in the ellipticity measurement as a function of the KSB signal-to-noise ratio from Erben et al. (2001). S10 calibrated this correction using simulated images of ground-based weak lensing observations from STEP2 (Massey et al. 2007), and verified that the same correction robustly corrects simulated high-resolution ACS-like weak lensing data with less than 2 per cent residual multiplicative ellipticity bias (0.8 per cent on average). However, as recently shown by Hoekstra et al. (2015), the STEP2 image simulations lack sources at the faint end, affecting the derived bias calibration (see also Hoekstra, Viola & Herbonnet 2017). Also, deviations in the assumed intrinsic galaxy shape distribution influence the noise-bias correction (e.g. Viola, Kitching & Joachimi 2014). To minimize the impact of such uncertainties we apply a more conservative galaxy selection

requiring $S/N = (Flux/Fluxerr)_{auto} > 10$ from SOURCE EXTRACTOR.[2] To be conservative, we additionally double the systematic uncertainty for the shear calibration in the error budget of our current cluster study (4 per cent), which is comparable to the mean shear calibration correction of the galaxies passing our cuts (average factor 1.05). In the context of cluster weak lensing studies a relevant question is also if the image simulations probe the relevant range of shears sufficiently well. We expect that this is not a major concern for our study given that $\langle g_t \rangle \lesssim 0.1$–$0.15$ for all of our clusters within the radial range used for the mass constraints (see Section 7). For comparison, the basic KSB+ implementation used in our analysis was tested in Heymans et al. (2006) using shears up to $g = 0.1$, where no indications were found for significant quadratic shear bias terms that would result in an inaccurate correction using our linear correction scheme.

We apply the same shape measurement pipeline to the CANDELS data discussed in Section 4.1.3. When mimicking our cluster field selection in these catalogues and assigning weights, we rescale the S/N values prior to the S/N cut to account for slight differences in depth. Hence, if a CANDELS tile is slightly shallower (deeper) compared to the cluster tile considered, we will apply a correspondingly slightly lower (higher) S/N cut in the CANDELS tile to select consistent galaxy samples. On average the depth of our CANDELS stacks agrees well with the depth of the cluster field stacks (to 0.065 mag). Together with the fact that $\langle \beta \rangle$ depends only weakly on V_{606} for our colour-selected sample at the faint end (see Section 6.5), we therefore ignore second-order effects such as incompleteness differences between the CANDELS and cluster field catalogues.

5.2 Test for residual CTI signatures in the ACS cluster data

CTI generates charge-trails behind objects dominantly in the parallel-transfer read-out direction. For raw ACS images this corresponds to the *y*-direction, and this is approximately also the case for distortion-corrected images if MULTIDRIZZLE is run using the native detector orientation. M14 test the performance of their pixel-based CTI correction by averaging the PSF-corrected ellipticity estimates of galaxies in blank field CANDELS data. Images without CTI correction show a prominent alignment with the *y*-axis ($\langle \epsilon_1 \rangle < 0$), where the magnitude of the effect increases with the *y*-separation relative to the readout amplifiers. In contrast, this alignment is undetected if the correction is applied.

We cannot apply the same test to our ACS data of the cluster fields given the presence of massive clusters, which are always located at the same position within the mosaics, and whose weak gravitational lensing shear would add to the saw-tooth CTI signature. However, we can make use of the fact that CTI primarily affects the ϵ_1 ellipticity component (measured along the image axes) but not the ϵ_2 ellipticity component (measured along the field diagonals). The tangential and cross components of the ellipticity with respect to the cluster centre

$$\epsilon_t = \epsilon_{t,1} + \epsilon_{t,2} \tag{17}$$

$$\epsilon_\times = \epsilon_{\times,1} + \epsilon_{\times,2} \tag{18}$$

[2] This cut is more conservative than the cut $S/N_{KSB} > 2$ from S10, which is based on the Erben et al. (2001) signal-to-noise ratio definition that includes a radial weak lensing weight function. $S/N_{KSB} > 2$ approximately corresponds to $S/N = (Flux/Fluxerr)_{auto} \gtrsim 6.5$ for our typical source galaxies, but note that there is a significant scatter between both estimates due to the different radial weighting.

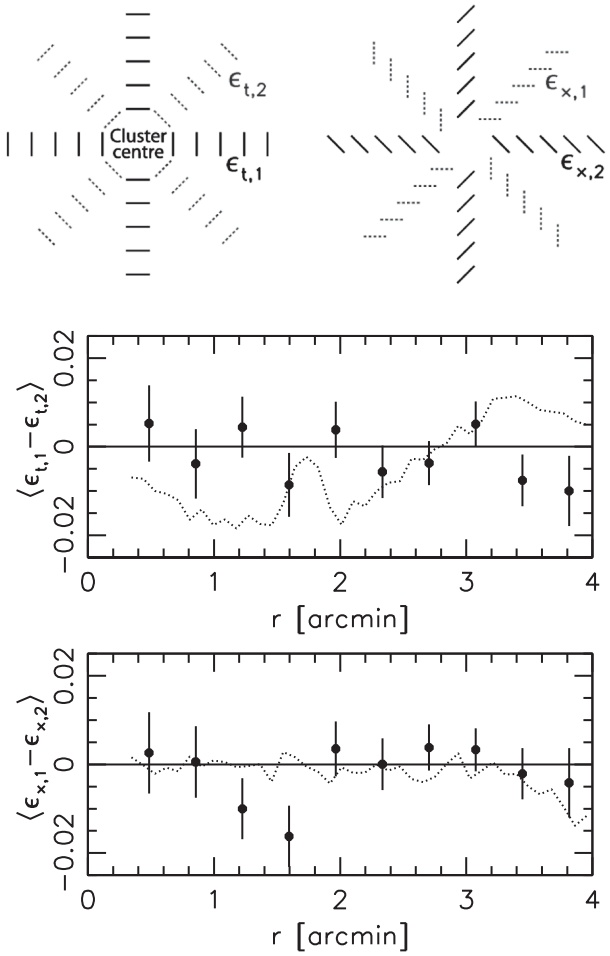

Figure 2. Testing for residual CTI systematics in the cluster fields. *Top:* Illustration for the separation of the tangential and cross components of the ellipticity into components affected by CTI ($\epsilon_{t,1}, \epsilon_{\times,1}$), and those unaffected by CTI ($\epsilon_{t,2}, \epsilon_{\times,2}$). The *middle* (*bottom*) panel shows the difference in the tangential (cross) ellipticity component with respect to the cluster centre as estimated from the CTI-affected and the CTI-unaffected components. Here we combine the signal from all galaxies passing the shape cuts with $24 < V_{606,auto} < 26.7$ in all cluster fields. The points are consistent with zero (χ^2/d.o.f. = 0.96) suggesting that the CTI has been fully corrected within the statistical precision of the data. For comparison, the dotted curve shows the signal which would be measured from an uncorrected CTI saw-tooth ellipticity pattern with $\langle e_1 \rangle = -0.05$, where small wiggles are caused by the sampling at the galaxy positions and the masks applied.

(compare equations 7 and 8) receive contributions from both ellipticity components with

$$\epsilon_{t,1} = -\epsilon_1 \cos 2\phi \tag{19}$$

$$\epsilon_{t,2} = -\epsilon_2 \sin 2\phi \tag{20}$$

$$\epsilon_{\times,1} = +\epsilon_1 \sin 2\phi \tag{21}$$

$$\epsilon_{\times,2} = -\epsilon_2 \cos 2\phi, \tag{22}$$

see the sketch in the top panel of Fig. 2 for an illustration of these components. In our test we stack the signal from all clusters. Here we expect that any anisotropy in the reduced shear pattern due to cluster halo ellipticity will average out leading to an approximately circularly symmetric shear field. Accordingly, in the absence of residual systematics we expect that $\langle \epsilon_{t,1} - \epsilon_{t,2} \rangle$ and $\langle \epsilon_{\times,1} - \epsilon_{\times,2} \rangle$ are

consistent with zero when averaged azimuthally. Fig. 2 shows that this is indeed the case for our data (χ^2/d.o.f. = 0.96), confirming the success of the CTI correction within the statistical precision of the data. For comparison, the dotted line in Fig. 2 shows the signal that would be caused by a typical uncorrected CTI ellipticity saw-tooth pattern with $\langle \epsilon_1 \rangle = -0.05$.[3]

6 CLUSTER MEMBER REMOVAL AND ESTIMATION OF THE SOURCE REDSHIFT DISTRIBUTION

Robust weak lensing mass measurements require accurate knowledge of the mean geometric lensing efficiency $\langle \beta \rangle$ of the source sample and its variance $\langle \beta^2 \rangle$ (see Section 2). For a given cosmological model these depend only on the source redshift distribution and cluster redshift. Surveys with sufficiently deep imaging in sufficiently many bands can attempt to estimate the probability distribution of source redshifts directly via photo-zs (e.g. Applegate et al. 2014). However, such data are not available for our cluster fields. Hence, we have to rely on an estimate of the redshift distribution from external reference fields. Here we use photometric redshift estimates for the CANDELS fields from the 3D-HST team (Skelton et al. 2014) as primary data set (see Section 6.1). Additionally, we use spectroscopic and grism redshift estimates for galaxies in the CANDELS fields, as well as much deeper data from the *Hubble* Ultra Deep field (HUDF) to investigate and statistically correct for systematic features in the CANDELS photo-zs (Section 6.3).

Given that our cluster fields are overdense at the cluster redshift we have to apply a colour selection that robustly removes galaxies at the cluster redshift both in the reference catalogue and our actual cluster field catalogues. Here we use colour estimates from the *HST*/ACS *F*606*W* and *F*814*W* images in the inner regions ('ACS-only' selection, Section 6.2), and we use VLT/FORS2 *I*-band imaging for the cluster outskirts ('ACS+FORS2' selection, Section 6.4 with details given in Appendix D). As discussed in Appendix E we also explored a different analysis scheme which substitutes the colour selection with a statistical correction for cluster member contamination, but we found that we could not control the systematics of the correction to the needed level due to the limited radial range probed by the *F*606*W* images. We optimize the analysis by splitting the colour-selected sources into magnitude bins (Section 6.5), investigate the influence of line-of-sight variations (Section 6.6), and account for weak lensing magnification (Section 6.7). Section 6.8 presents consistency checks for our analysis based on the source number density measured as a function of magnitude and cluster-centric distance.

6.1 CANDELS photometric redshift reference catalogues from 3D-HST

We make use of photometric redshift catalogues computed by the 3D-HST team (Brammer et al. 2012; Skelton et al. 2014, hereafter S14) for the CANDELS fields (Grogin et al. 2011), which consist of five independent lines of sight (AEGIS, COSMOS, GOODS-North, GOODS-South, UDS). Hence, their combination efficiently suppresses the impact of sampling variance. All CANDELS field were observed by *HST* with ACS and WFC3, including ACS *F*606*W* and

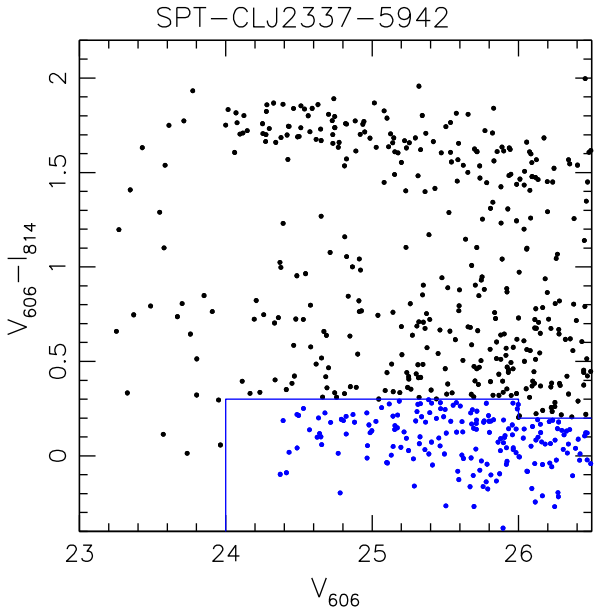

Figure 3. Measured $V_{606} - I_{814}$ colours as a function of V_{606} for galaxies in the field of SPT-CL *J*2337 −5942 that pass our weak lensing shape cuts, and that are located within the central I_{814} ACS tile. The blue lines indicate the region of blue galaxies that pass our colour selection. The cluster red sequence is clearly visible at $V_{606} - I_{814} \sim 1.7$.

*F*814*W*[4] imaging mosaics that have at least the depth of our cluster field observations (see Koekemoer et al. 2011). This includes observations from the CANDELS program (Grogin et al. 2011) and earlier projects (Giavalisco et al. 2004; Rix et al. 2004; Davis et al. 2007; Scoville et al. 2007). The S14 catalogues are based on detections from combined *HST*/WFC3 NIR *F*125*W*+*F*140*W*+*F*160*W* images, and include photometric measurements from a total of 147 distinct imaging data sets from *HST*, *Spitzer* and ground-based facilities with a broad wavelength coverage from 0.3−8 μm (18–44 data sets per field). S14 compute photometric redshifts using EAZY (Brammer, van Dokkum & Coppi 2008), which fits the observed spectral energy distribution (SED) constraints of each object with a linear combination of galaxy templates.

We have matched the S14 catalogues with our *F*606*W*-detected shape catalogues of the CANDELS fields (see Section 5). After applying weak lensing cuts, accounting for masks, and restricting the analysis to the overlap region of the ACS and WFC3 mosaics, we find that ∼97.6 per cent of the galaxies in the shape catalogues with $24 < V_{606} < 26.5$ have a direct match within 0.5 arcsec in the S14 catalogues, showing that they are nearly complete within our employed magnitude range (see Appendix B for an investigation of the ∼2.4 per cent of non-matching galaxies which shows that they have a negligible impact).

6.2 Source selection using ACS-only colours

In the inner cluster regions we apply a colour selection (indicated in Fig. 3) using our ACS *F*606*W* and *F*814*W* images, selecting only

[3] M14 measure an average uncorrected CTI-induced galaxy ellipticity at $V \sim 26.5$ of $\langle \epsilon_1 \rangle \simeq -0.04$ from CANDELS/COSMOS *F*606*W* images, which were observed at a similar epoch but have higher background levels than our data, and thus weaker CTI signals.

[4] For the GOODS-North field we estimate the I_{814} magnitudes from the S14 flux measurements in the *F*775*W* and *F*850LP filters. When conducting selections or binning in V_{606} based on the S14 photometry we undo their correction for total magnitudes in order to employ aperture magnitudes that are consistent with our cluster field measurements.

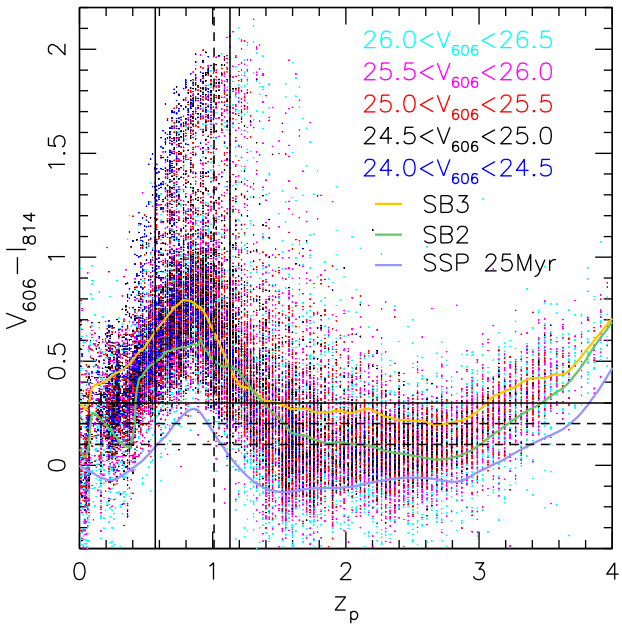

Figure 4. $V_{606} - I_{814}$ colours of galaxies in the CANDELS fields as a function of the peak photometric redshift z_p from S14. The colour coding splits the galaxies into our different magnitude bins. The horizontal lines mark our different colour cuts (dependent on cluster redshift and galaxy magnitude, see Section 6.2), while the vertical lines indicate the cluster redshift range $0.57 \leq z \leq 1.13$ (solid), as well as $z = 1.01$ (dashed), at which cluster redshift the colour cuts change. The curves indicate synthetic $V_{606} - I_{814}$ colours of galaxy SED templates from Coe et al. (2006).

galaxies that are bluer than nearly all galaxies at the cluster redshift. This is illustrated in Fig. 4, where we plot the EAZY peak photometric redshift z_p for the CANDELS galaxies as a function of $V_{606} - I_{814}$ colour from S14 (measured with the same 0.7 arcsec aperture diameter as employed for our ACS colour measurements). Figs 4 and 5 illustrate that the selection of blue galaxies in $V_{606} - I_{814}$ colour in CANDELS is very effective in removing galaxies at our cluster redshifts, while it selects the majority of the $z_p \gtrsim 1.4$ background galaxies. The latter are high-redshift star-forming galaxies observed at rest-frame UV wavelength with very blue spectral slopes. In contrast, nearly all galaxies at the cluster redshifts show a redder $V_{606} - I_{814}$ colour, as they contain either the 4000 Å break (early type galaxies, see the cluster red sequence in Fig. 3) or the Balmer break (late type galaxies) within the filter pair.

We note that our approach rejects both red and blue cluster members. It is therefore more conservative and robust than redder colour cuts that some studies have used to remove red sequence cluster members only (e.g. Jee et al. 2011). Note that, in contrast, Okabe et al. (2013) select only galaxies that are redder than the red sequence. This is a useful approach for the low-redshift clusters targeted in their study, but less effective for the high-redshift clusters studied here, as most of the $z_p \gtrsim 1.4$ background galaxies are blue at optical wavelengths (see Fig. 5). Likewise, some studies of lower redshift clusters have used combinations of blue and red regions in colour space to minimize cluster member contamination (e.g. Medezinski et al. 2010; High et al. 2012; Umetsu et al. 2014). It is evident from Fig. 4 that a selection of blue galaxies in $V - I$ colour is inefficient for clusters at low redshifts $z \lesssim 0.4$, as it would either require extremely blue cuts that drastically shrink the source sample, or lead to a larger residual contamination by galaxies at the cluster redshift. Similar results were found by Ziparo et al. (2016), who conclude that optical observations alone are not sufficient to reduce the cluster member contamination below the per cent level for blue source samples and clusters at $z \sim 0.2$.

For clusters at $z < 1.01$ we select source galaxies with $V_{606} - I_{814} < 0.3$. This maximizes the background galaxy density while at the same time removing 98.5 per cent of the CANDELS galaxies at $0.6 < z_p < 1$ that pass the other weak lensing cuts, see the top left panel of Fig. 5. For the higher redshift clusters we apply a more stringent cut $V_{606} - I_{814} < 0.2$ which still yields a 97.6 per cent suppression of galaxies at $1 < z_p < 1.13$, at the expense of a slightly lower source density (top right panel of Fig. 5). When conducting the analysis for our cluster fields we apply slightly more conservative colour cuts that are bluer by 0.1 mag for the faintest sources in our analysis, as they show the largest photometric scatter. As a result, we obtain a similar fraction of removed galaxies at the cluster redshifts when taking photometric scatter into account (see Section 6.4 and Appendix D3).

In Fig. 4 we also overplot synthetic $V_{606} - I_{814}$ colours of redshifted SED templates for star-forming galaxies employed in the Bayesian Photometric Redshift (BPZ) algorithm (Benítez 2000). This includes the SB3 and SB2 starburst templates from Kinney et al. (1996) as recalibrated by Benítez et al. (2004). We additionally include a young starburst model [simple stellar population (SSP) 25 Myr], which is one of the templates introduced by Coe et al. (2006) into BPZ to improve photometric redshift estimates for very blue galaxies in the HUDF. The shown SED corresponds to an SSP model with an age of 25 Myr and metallicity $Z = 0.08$ (Bruzual & Charlot 2003). At the cluster redshifts, the colours of the SB3 and SB2 templates approximately describe the range of colours of typical blue cloud galaxies, which are well removed by our colour selection. In contrast, while the colour of the SSP 25 Myr model appears to be representative for a considerable fraction of the $z \gtrsim 1.4$ background galaxies, it approximately marks the location of the most extreme blue outliers at the cluster redshifts, which are not fully removed by our colour selection scheme. If the clusters contain a substantial fraction of such extremely blue galaxies, this might introduce some residual cluster member contamination in our lensing catalogue. We investigate this issue in Appendix F, concluding that such galaxies have a negligible impact for our analysis despite the physical overdensity of galaxies in clusters. We also present empirical tests for residual contamination by cluster galaxies in Section 6.8.

6.3 Statistical correction for systematic features in the photometric redshift distribution

We base our estimate of the source redshift distribution on the CANDELS photo-z catalogues because of their high completeness at the depth of our SPT ACS observations (Section 6.1), allowing us to select galaxies that are representative for the galaxies used in our lensing analysis. However, it is important to realize that such photo-z estimates may contain systematic features (e.g. catastrophic outliers) that can bias the inferred redshift distribution and accordingly the lensing results. As an example, the cosmological weak lensing analysis of COSMOS data by S10 suggests that the majority of faint galaxies in the COSMOS-30 photometric redshift catalogue (Ilbert et al. 2009) that have a primary peak in their posterior redshift probability distribution $p(z)$ at low redshifts but also a secondary peak at high redshifts, are truly at high redshift. Likewise, the galaxy–galaxy lensing analysis of CFHTLenS data by Heymans et al. (2012) indicates that a significant fraction of galaxies with an assigned photometric redshift $z_{\rm photo} < 0.2$ are truly at high redshift. In the following subsections we exploit additional data sets to check

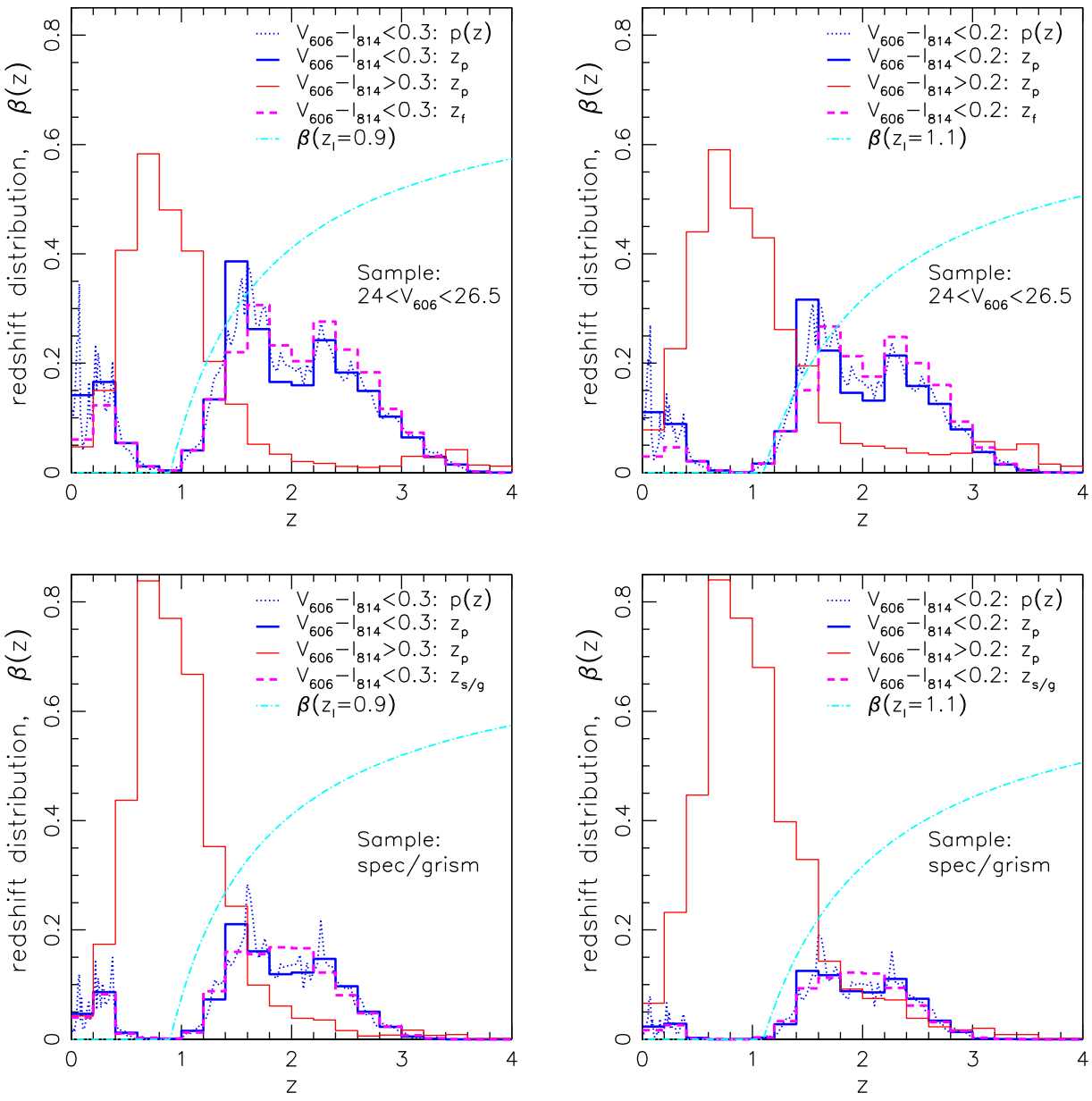

Figure 5. Redshift distribution of different galaxy samples in CANDELS: The *top* panels show the full photometric sample of galaxies which have $24.0 < V_{606} < 26.5$ and pass the shape cuts, whereas the sample is further reduced to contain only those galaxies with robust spec-zs or grism-zs in the *bottom* panels. In the *left*-hand (*right*-hand) panels, a colour cut $V_{606} - I_{814} < 0.3$ ($V_{606} - I_{814} < 0.2$) is used to separate the source sample (solid thick photo-z histogram and thin dotted averaged $p(z)$ in blue) from redder galaxies (thin solid red photo-z histogram) that contain most galaxies at the corresponding cluster redshifts. The magenta dashed histogram shows the distribution of spec-zs or grism-zs in the *bottom* panels, and the distribution of photo-zs after the statistical correction based on the HUDF analysis in the *top* panels. The histograms are normalized according to the total number of galaxies in the corresponding spectroscopic or photometric sample prior to the colour selection. The cyan dash–dotted curve shows the geometric lensing efficiency β for clusters at redshift $z_l = 0.9$ (*left*) and $z_l = 1.1$ (*right*). The presence of foreground galaxies in the source sample is not a concern as long as it is modelled accurately.

the accuracy of the CANDELS photo-zs and implement a statistical correction for relevant systematic features.

6.3.1 Tests and statistical correction based on HUDF data

The HUDF is located within one of the CANDELS fields (GOODS-South). The very deep multiwavelength observations conducted in the HUDF can therefore be used for cross-checks of the CANDELS photo-zs.

As first data set we use a combination of high-fidelity spectroscopic redshifts ('spec-zs', z_s) compiled by Rafelski et al. (2015),[5] and redshift estimates extracted by the 3D-HST team (Brammer et al. 2012, 2013) from the combination of deep *HST* WFC3/IR slitless grism spectroscopy and very deep *HST* optical/NIR

[5] Rafelski et al. (2015) note that the object 10157 in their catalogue is problematic as it consists of a blend of two galaxies at different redshifts. We therefore exclude it from the spec-z/grism-z sample used in our analysis.

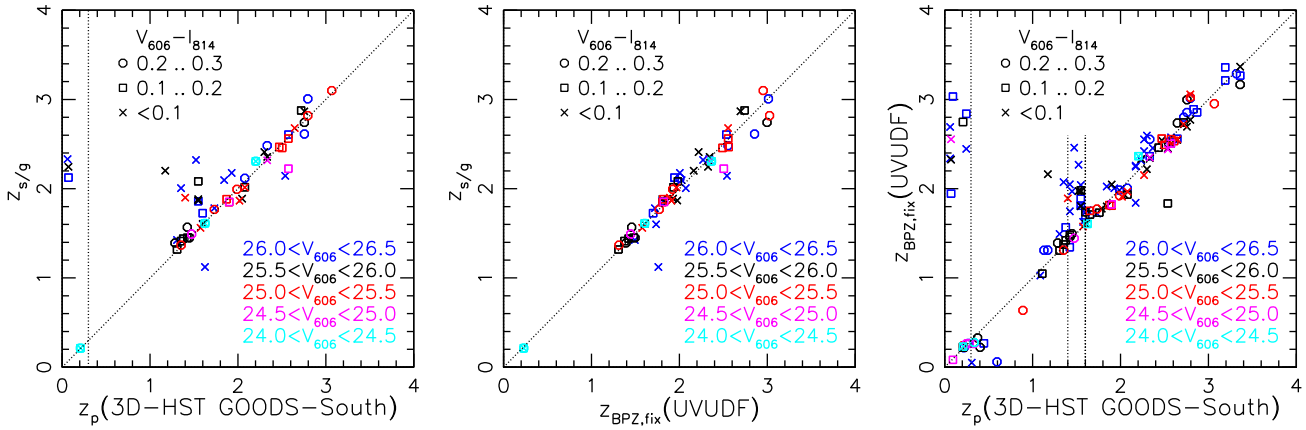

Figure 6. Comparison of redshift estimates in the HUDF including the peak photometric redshift from EAZY z_p estimated by the 3D-HST team in the GOODS-South field, the BPZ photometric redshift from the UVUDF project $z_{BPZ,fix}$ (with small bias corrections applied, see text), and a combined sample of spectroscopic and grism redshifts $z_{s/g}$. We regard the latter as a true but incomplete reference sample, which reveals the presence of significant outliers for the z_p but not the $z_{BPZ,fix}$ photo-zs. We therefore use the $z_{BPZ,fix}$ photo-zs, which do not suffer from incompleteness at the relevant depth, to derive a statistical correction for the z_p photo-zs. The symbols split the galaxies according to $V_{606} - I_{814}$ colour and the different colours indicate different magnitude bins (based on the 3D-HST photometry). Galaxies are only included if they pass our weak lensing selection and if they are located within the area covered by the WFC3 UVIS and IR observations. In the *right*-hand panel the vertical lines indicate the z_p ranges of our statistical correction for the redshift distribution.

imaging. These 'grism-zs' (z_g) significantly enlarge the sample of high-z ($z > 1$) galaxies with high-quality redshift estimates, where typical errors of the grism-zs are $\sigma_z \approx 0.003 \times (1 + z)$ (Brammer et al. 2012; Momcheva et al. 2016).

We compare the CANDELS photo-zs to the HUDF $z_{s/g}$ estimates in the left-hand panel of Fig. 6. The majority of the data points closely follow the diagonal, suggesting that the 3D-HST photo-zs are overall well calibrated as needed for unbiased estimates of the redshift distribution. However, we note the presence of two relevant systematic features: first, there are three catastrophic outliers that are at high $z_{s/g} \simeq 2.2$, but are assigned a low $z_p \simeq 0.07$. Secondly, there is an increased, asymmetric scatter at $1.2 \lesssim z_p \lesssim 1.7$. Most notably, many galaxies with an assigned photometric redshift $1.4 \lesssim z_p \lesssim 1.6$ are actually at higher redshift. This is likely the result of redshift focusing effects (e.g. Wolf 2009) caused by the broad-band *HST* filters. While this comparison allows us to identify these issues, the matched catalogue is insufficient to derive a robust statistical correction for our full photometric sample given the incompleteness of the $z_{s/g}$ sample.

To overcome this limitation of incompleteness, we use deep photometric redshifts computed by Rafelski et al. (2015) using HUDF data as a second comparison sample. Compared to the CANDELS photo-zs they benefit from much deeper *HST* optical (Beckwith et al. 2006) and NIR imaging (Koekemoer et al. 2013), and additionally incorporate new *HST*/UVIS Near UV imaging from the UVUDF project (Teplitz et al. 2013) taken in the *F*225*W*, *F*275*W* and *F*336*W* filters. These bands probe the Lyman break in the redshift range $1.2 \lesssim z \lesssim 2.7$, which contains most of our weak lensing source galaxies. At these redshifts, the NIR imaging additionally probes the location of the Balmer/4000 Å break. Hence, we expect that the resulting photo-z should be highly robust against catastrophic outliers. We test this by comparing them to the $z_{s/g}$ redshifts in the middle panel of Fig. 6. Here we use the photo-z estimates z_{BPZ} obtained by Rafelski et al. (2015) using BPZ as it yields the highest robustness against catastrophic outliers in their analysis. Note that the comparison of z_{BPZ} and $z_{s/g}$ suggests that z_{BPZ} slightly overestimates the redshifts for the colour-selected sample in the redshift intervals $1.0 \lesssim z_{BPZ} \lesssim 1.7$ and $2.6 \lesssim z_{BPZ} \lesssim 3.7$, with median redshift offsets of 0.071 and 0.171, respectively. We have therefore subtracted these offsets in the corresponding redshift intervals, yielding $z_{BPZ,fix}$, which is shown in Fig. 6. As visible in the middle panel of Fig. 6, $z_{BPZ,fix}$ correlates tightly with $z_{s/g}$. In particular, the three catastrophic outliers from the left-hand panel are now correctly placed at high redshifts. Likewise, the redshift focusing effects are basically removed. The remaining scatter with one moderate outlier has negligible impact on our results. For example, $\langle\beta\rangle$ agrees to 0.4 per cent between $z_{BPZ,fix}$ and $z_{s/g}$ for the matched catalogue and clusters at $z_l = 1.0$ (we include this in the systematic error budget of Section 6.3.2). This suggests that $z_{BPZ,fix}$ provides a sufficiently accurate approximation for the true redshift. Hence, we use $z_{BPZ,fix}$ as a reference to obtain a statistical correction for the systematic features of the CANDELS photo-zs.

We compare the 3D-HST photo-zs z_p in the HUDF to $z_{BPZ,fix}$ in the right-hand panel of Fig. 6, again showing the previously identified catastrophic outliers at $z_p < 0.3$ and redshift focusing effects at $1.4 \lesssim z_p \lesssim 1.6$, but now at the full depth of our photometric sample. The catastrophic outliers with $z_p < 0.3$ are dominated by blue $V_{606} - I_{814} < 0.2$ galaxies, for which 9 out of 12 galaxies appear to be truly at high redshifts. In order to implement a statistical correction for these outliers for the full CANDELS catalogue, we note the 12 redshift offsets $(z_{BPZ,fix} - z_p)_i$. We bootstrap this empirically defined distribution to define the correction: for each CANDELS galaxy with $z_p < 0.3$ and $V_{606} - I_{814} < 0.2$ we add a randomly drawn offset to its z_p. Likewise, we apply a statistical correction for the redshift focusing within the redshift range $1.4 \leq z_p \leq 1.6$ for galaxies with $V_{606} - I_{814} < 0.1$ (which are most strongly affected, see Fig. 6), again randomly sampling from the corresponding $(z_{BPZ,\,fix} - z_p)_i$ offsets in the HUDF. For the latter correction we split the galaxies into two magnitude ranges ($24 < V_{606} < 25.5$ and $25.5 < V_{606} < 26.5$) given that the fainter galaxies appear to suffer from the redshift focusing effects more strongly. We show the resulting distribution of statistically corrected redshifts z_f as magenta dashed histograms in the top panels of Fig. 5. As expected, it has a lower fraction of low-z galaxies compared to the uncorrected z_p distribution, as well as a reduction of the redshift focusing peak at $1.4 \leq z_p \leq 1.6$. Both effects are compensated by a higher fraction of high-z galaxies, where we also note that the local minimum at

$z_p \simeq 2$, which likely results from the redshift focusing (see also Section 6.3.3), is reduced.

Averaged over our full cluster sample, and accounting for the magnitude-dependent effects explained in the following sections (e.g. shape weights), the application of this correction scheme leads to a 12 per cent decrease of the resulting cluster masses. Of this, 10 per cent originate from the correction for catastrophic outliers, and 2 per cent from the correction for redshift focusing.

6.3.2 Uncertainty of the statistical correction of the redshift distribution

The statistical correction of the redshift distribution explained in Section 6.3.1 has a non-negligible impact on our analysis. Therefore it is important to quantify its uncertainty. We consider a number of effects that affect the uncertainty: first, we estimate the statistical uncertainty originating from the limited size of the HUDF catalogue by generating bootstrapped versions of it, which are then used to generate the $(z_{\rm BPZ,fix} - z_p)_i$ offset samples. This yields a small, 0.5 per cent uncertainty regarding the average masses. Secondly, our correction scheme assumes that the relative effects seen in the HUDF are representative for the full CANDELS area. However, some previous studies suggest that the GOODS-South field, which contains the HUDF, could be somewhat underdense at lower redshifts compared to the cosmic mean (e.g. Schrabback et al. 2007; Hartlap et al. 2009). To obtain a worst case estimate of the impact this could have, we assume that the GOODS-South field could be underdense at low redshifts by a factor 3 compared to the cosmic mean. Hence, we artificially boost the number of HUDF galaxies with $z_p < 0.3$ that are truly at low-z by a factor 3 for the generation of the offset pool. On average this leads to a 3 per cent increase of the cluster masses. Thirdly, we note that our correction for redshift focusing incorporates most but not all of the corresponding outliers in the right-hand panel of Fig. 6. We assume a conservative 50 per cent relative uncertainty on the 2 per cent correction, corresponding to an absolute 1 per cent uncertainty. Adding all individual systematic uncertainties identified here and in Section 6.3.1 in quadrature yields a combined systematic uncertainty for the systematic corrections to the photometric redshifts of 3.3 per cent in the average cluster mass.

6.3.3 Consistency checks using spectroscopic and grism redshifts in the CANDELS fields

In Section 6.3.1 we obtained a statistical correction for systematic features in the CANDELS photo-zs using very deep data available in the HUDF. Here we present cross-checks for this correction using the CANDELS redshift catalogue from Momcheva et al. (2016), which combines a compilation of high fidelity spectroscopic redshifts from S14 with redshift estimates derived from their joint analysis of slitless WFC3/NIR grism spectra from the 3D-*HST* project and the S14 photometric catalogues. These grism data are shallower than those available in the HUDF (see Section 6.3.1) but cover a much wider area. We restrict the use of these grism-zs to relatively bright galaxies (NIR magnitude $JH_{\rm IR} < 24$). These galaxies were individually inspected by the 3D-HST team, allowing us to select galaxies classified to have robust redshift estimates. For these relatively bright galaxies the continuum emission is comfortably detected in the grism data, yielding high-quality redshift estimates with a typical redshift error of $\sigma_z \approx 0.003 \times (1 + z)$ (Momcheva et al. 2016), which we can neglect compared to the photo-z uncertainties.

For the combined sample of galaxies with spec-zs and grism-zs we compare the colour-selected histogram of spec-zs/grism-zs ($z_{\rm s/g}$, using z_s in case both are available) to the histogram of their photo-zs in the bottom panels of Fig. 5. Here we note two points: First, the spec-zs/grism-zs confirm that the colour selection indeed provides a very efficient removal of galaxies at our targeted cluster redshifts. Secondly, the high-z galaxies are distributed in a relatively symmetric, unimodal peak that has a maximum at $z \simeq 1.9$ according to spec-zs/grism-zs. In contrast, the photo-z histogram shows two slight peaks ($z \simeq 1.5$ and $z \simeq 2.3$). This is consistent with the conclusion from Section 6.3.1 that the peaks in the photo-z histogram of the full photometric sample (top panels of Fig. 5) at these redshifts are a result of redshift focusing effects and not true large-scale structure peaks in the galaxy distribution.

As a further cross-check we reconstruct the redshift distribution of the photometric sample by exploiting its spatial cross-correlation with the spec-zs/grism-z sample, applying the technique developed by Newman (2008), Schmidt et al. (2013) and Ménard et al. (2013). Specifically, we use the implementation in THE-WIZZ[6] redshift recovery code (Morrison et al. 2017). We provide the details of this analysis in Appendix C, showing that it independently confirms the presence of the catastrophic redshift outliers and redshift focusing effects.

6.3.4 Limitations of the averaged posterior probability distribution

Past weak lensing studies suggest that a better approximation of the true source redshift distribution may be given by the average photometric redshift posterior probability distribution $p(z)$ of all sources compared to a histogram of the best-fitting (or peak) photometric redshifts (see e.g. Heymans et al. 2012; Benjamin et al. 2013; Bonnett 2015). To test this we recompute the $p(z)$ using EAZY from the S14 photometric catalogues, which is necessary as the $p(z)$ are not reported in the S14 catalogues.

As visible in Fig. 5, the redshift distribution inferred from the averaged $p(z)$ is relatively similar to the normalized histogram of the peak photometric redshifts z_p. We note that the redshift focusing peak at $z_p \simeq 1.5$ and local minimum at $z_p \simeq 2$ are slightly less pronounced in the averaged $p(z)$, but they do not reach the level suggested by the corrected z_f histogram. More severely, the averaged $p(z)$ overpredicts the fraction of low-z galaxies compared to the z_f distribution similarly to the z_p histogram. We therefore conclude that the use of the averaged $p(z)$ instead of the z_p histogram is insufficient to account for the systematic features identified in Section 6.3.1.

6.4 Source selection in the presence of photometric scatter

Outside the area of the central $F814W$ ACS tile we only have single band $F606W$ observations from *HST*. For the colour selection we therefore have to combine the $F606W$ data with the VLT/FORS2 I-band imaging (see Section 4.2). We measure colours between these images as described in Appendix D1. However, VLT/FORS2 I-band observations are not available in all CANDELS fields. We therefore need to accurately map the selection in the ACS+FORS2-based $V_{606,\rm con} - I_{\rm FORS2}$ colour to the $V_{606} - I_{814}$ colour available in CANDELS. We empirically obtain this mapping through the

[6] Available at https://github.com/morriscb/The-wiZZ.

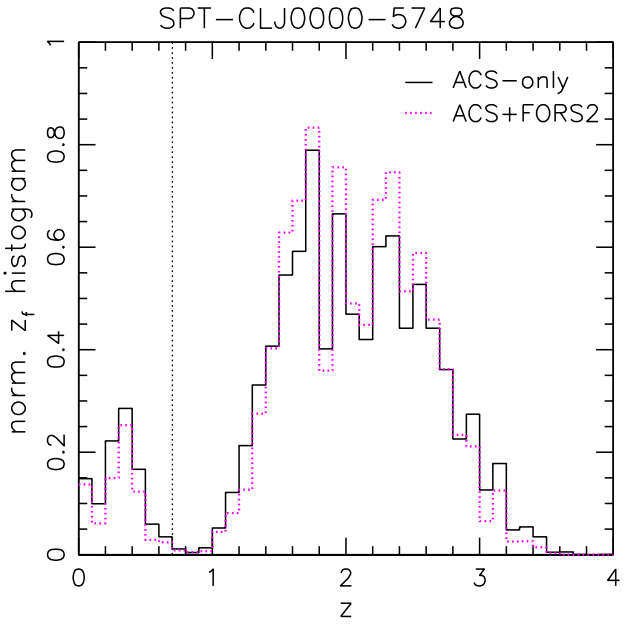

Figure 7. Normalized histogram of the statistically corrected photometric redshift estimates z_f for all galaxies in our CANDELS catalogues that pass the weak lensing cuts and the colour selection after adding noise to mimic the properties of the SPT-CL J0000−5748 data, both for the ACS+FORS2 (magenta dotted) and the ACS-only (black solid) selection. The vertical dotted line indicates the cluster redshift, at which both selections achieve an efficient suppression also in the presence of noise.

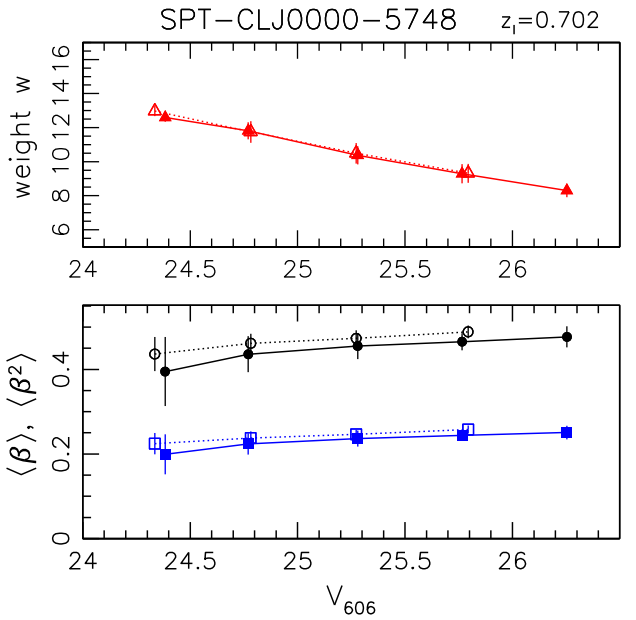

Figure 8. Analysis of SPT-CL J0000−5748 as a function of V_{606} magnitude, where the solid (open) symbols correspond to the ACS-only (ACS+FORS2) analysis. *Top:* Mean weak lensing shape weight w with error-bars indicating the dispersion from all selected galaxies in the magnitude bin. *Bottom:* $\langle\beta\rangle$ (circles) and $\langle\beta^2\rangle$ (squares) with error-bars showing the dispersion of their estimates computed from all CANDELS sub-patches (see Section 6.6), thus indicating the expected line-of-sight variations for the field sizes of our cluster observations.

comparison of both colour estimates in the inner cluster regions, where both are available (see Appendix D2).

As described in Appendix D3 we add photometric scatter to the catalogues from the CANDELS fields to mimic the noise properties of the cluster fields for the colour selection. In particular, we apply an empirical model for the (non-Gaussian) scatter between the ACS-only and the ACS+FORS2 colours derived from the comparison of the colour measurements in the inner cluster regions. The ACS-only colour selection has higher signal-to-noise, allowing us to include galaxies with $V_{606} < 26.5$ in the analysis. In contrast, the ACS+FORS2 colour selection is more noisy, which is why we have to employ shallower magnitude limits (dependent on the depth of the VLT data, see Table 2) and more stringent colour cuts (see Table D1 in Appendix D). Fig. 7 demonstrates that this approach leads to a robust removal of galaxies at the cluster redshift despite the presence of noise. Here we show the histogram of the statistically corrected redshift estimates z_f for the CANDELS galaxies passing the colour selection for SPT-CL J0000−5748 after application of the photometric scatter. Averaged over the cluster sample we find that 98.9 per cent (98.1 per cent) of the CANDELS galaxies with $|z_f - z_l| \leq 0.025$ are removed by the ACS+FORS2 (ACS-only) colour selection scheme when the noise is taken into account. As shown in Appendix F this translates into a negligible expected cluster member contamination in the weak lensing analysis. In addition, we will show in Section 6.8 that the total source density and the source density profiles provide limits on the residual cluster member contamination, which are consistent with no contamination.

6.5 Analysis in magnitude bins

As shown in Fig. 8, $\langle\beta\rangle$ increases moderately within the magnitude range $24 < V_{606} < 26.5$, which is due to a larger fraction of high-redshift galaxies passing the colour selection at fainter magnitudes. We only include galaxies with $V_{606} > 24$ in our analysis as brighter galaxies contain only a low fraction of background sources. We split the source galaxies into subsets according to V_{606} magnitude (0.5 mag-wide bins) in order to optimize the S/N of our measurement. This allows us to adequately weight the bins in the analysis accounting for not only the shape weight w, but also the geometric lensing efficiency.

6.6 Accounting for line-of-sight variations

There is statistical uncertainty on how well we can estimate the cosmic mean $\langle\beta\rangle$ in a magnitude bin (given our lensing and colour selection) due to sampling variance and the finite sky-coverage of CANDELS. Furthermore, the actual redshift distribution along the line of sight to each of our clusters will be randomly sampled from this cosmic mean distribution, leading to additional statistical scatter, see e.g. Hoekstra et al. (2011b), who show that this is particularly relevant for high-z clusters.

To account for the statistical scatter in our weak lensing mass analysis (Section 7), we subdivide the CANDELS fields into sub-patches that match the size of our cluster field observations (single ACS tiles for the ACS-only colour selection and 2×2 mosaics for the ACS+FORS2 selection) and compute $\langle\beta\rangle_i$ and $\langle\beta^2\rangle_i$ from the redshift distribution of each sub-patch i. From the scatter of these quantities between all sub-patches we compute the resulting scatter in the mass constraints in Section 7.2.

Furthermore, we need to investigate if the uncertainty on the estimate of the cosmic mean $\langle\beta\rangle$ due to the finite sky-coverage of CANDELS adds a significant systematic uncertainty in our error budget. For this, we first compute the uncertainty on the mean $\langle\beta\rangle$ from the variance of the $\langle\beta\rangle_i$. Assuming all N sub-patches

were statistically independent, we find a very small relative uncertainty $\frac{\Delta\langle\beta\rangle}{\langle\beta\rangle} = \sigma_{\langle\beta\rangle_i}/(\langle\beta\rangle\sqrt{N-1}) = 0.3$ per cent (0.6 per cent) for our lowest redshift cluster SPT-CL J2331−5051 at $z_\mathrm{l} = 0.576$ and 0.4 per cent (1.1 per cent) for our highest redshift cluster SPT-CL J2106−5844 at $z_\mathrm{l} = 1.132$ using the ACS-only (ACS+FORS2) colour selection combining all magnitude bins. However, due to large-scale structure the $\langle\beta\rangle_i$ within each CANDELS field will be correlated. A more conservative estimate can be obtained by computing $\langle\beta\rangle_i$ for each CANDELS field (without sub-patches) and estimating $\frac{\Delta\langle\beta\rangle}{\langle\beta\rangle}$ from the variation between the five fields.[7] This yields $\frac{\Delta\langle\beta\rangle}{\langle\beta\rangle} = 0.6$ per cent (0.6 per cent) for SPT-CL J2331−5051 and 0.4 per cent (1.0 per cent) for SPT-CL J2106−5844, again employing the ACS-only (ACS+FORS2) colour selection. This uncertainty is taken into account in our systematic error budget in Section 7.5, but we note that it is very small compared to our statistical errors in all cases.

6.7 Accounting for magnification

In addition to the shear, the weak lensing effect of the clusters magnifies background sources by a factor $\mu(z)$ given by equation (15). This effectively alters the source redshift distribution, but this effect has typically been ignored in previous studies. For our analysis this has three effects: first, it increases the fluxes of sources by a factor $\mu(z)$, which may place them into brighter magnitude bins, thus increasing the total source density by including galaxies which are intrinsically too faint to be included. Secondly, it reduces the source sky area we observe, diluting the number density of sources by a factor $\mu(z)$. Finally, the magnification of object sizes may lead to the inclusion of some small galaxies which would otherwise be excluded by the lensing size cut. However, the large majority of our galaxies are well-resolved with *HST*, so we will ignore this latter effect (but it may be more relevant for data with lower image quality).

We estimate the impact of the first and second effect from a colour-selected[8] S14 CANDELS catalogue (lensing is achromatic). Here we restrict the analysis to the deeper GOODS fields, initially including galaxies down to $V_{606} = 27.5$. For this part of the analysis we do not require a matching entry in our 1 orbit-depth shape catalogue in order to maximize the completeness at the faint end. We include the statistical correction for catastrophic redshift outliers and redshift focusing from Section 6.3.1, where we apply the same scheme also for one additional magnitude bin with $26.5 < V_{606} < 27.5$. For each cluster redshift we compute $\beta(z_i)$ for each galaxy i (using $z_i = z_{\mathrm{f},i}$) in the CANDELS catalogue and approximate the magnification as

$$\mu(z) - 1 \simeq \frac{\beta(z)}{\beta_0}\,(\mu_0 - 1)\,, \tag{23}$$

where μ_0 indicates the magnification at an arbitrary fiducial β_0, for which we use $\beta_0 = 0.3$ close to the mean β for our higher redshift clusters (compare Table 3). The scaling in equation (23) is adequate in the weak lensing limit ($|\kappa| \ll 1$, $|\gamma| \ll 1$), in which case equation (15) simplifies to

$$\mu(z) = \frac{1}{1 - 2\frac{\beta(z)}{\beta_0}\kappa_0 + \left(\frac{\beta(z)}{\beta_0}\right)^2(\kappa_0^2 - |\gamma_0|^2)} \simeq 1 + 2\frac{\beta(z)}{\beta_0}\kappa_0, \tag{24}$$

where κ_0 and γ_0 are the convergence and shear for $\beta = \beta_0$. In practice we find that the assumed linear scaling with β in equation (23) is sufficiently accurate for all of our clusters within the considered radial range of the tangential reduced shear profile fits (see Section 7.2).

For each galaxy in the CANDELS catalogue we compute $\mu(z_i)$ for a range of μ_0. We then estimate the magnified magnitude $V_{606,i}^{\mathrm{lensed}} = V_{606,i} - 2.5\log_{10}\mu(z_i)$ for each galaxy, and keep track of the reduced sky area through a weight $W_i = 1/\mu(z_i)$. By binning in $V_{606,i}^{\mathrm{lensed}}$ we then compute the lensed number density

$$n_{\mathrm{gal}}^{\mathrm{lensed}} = \sum_{\mathrm{galaxies}} W_i/\mathrm{area} \tag{25}$$

and the mean lensed geometric lensing efficiency

$$\langle\beta\rangle^{\mathrm{lensed}} = \sum_{\mathrm{galaxies}} W_i\beta_i \Big/ \sum_{\mathrm{galaxies}} W_i\,, \tag{26}$$

where the summations are over all galaxies with lensed magnitudes falling into the corresponding bin. In Fig. 9 we plot the ratio of these quantities to their not-lensed counterparts n_{gal} and $\langle\beta\rangle$ computed in V_{606} bins with uniform weight.[9] This shows that magnification has only a minor net effect at magnitudes $V_{606} \simeq 25.5$–26, which contain a large fraction of our source galaxies. In contrast, it significantly boosts both quantities for brighter magnitudes $V_{606} \lesssim 25$. The net impact of magnification on high-z cluster mass estimates therefore

Table 3. Summary geometric lensing efficiency and source densities.

Cluster	$\langle\beta\rangle$	$\langle\beta^2\rangle$	$\sigma_{\langle\beta\rangle_i}/\langle\beta\rangle$	n_{gal} (arcmin^{-2})	
				ACS-only	ACS+FORS2
SPT-CL J0000−5748	0.466	0.243	0.053	18.2	7.2
SPT-CL J0102−4915	0.374	0.163	0.068	20.4	3.6
SPT-CL J0533−5005	0.368	0.159	0.062	19.7	5.4
SPT-CL J0546−5345	0.303	0.107	0.083	13.1	2.9
SPT-CL J0559−5249	0.505	0.288	0.064	18.2	4.0
SPT-CL J0615−5746	0.334	0.132	0.075	18.0	2.3
SPT-CL J2040−5725	0.344	0.141	0.077	16.2	3.5
SPT-CL J2106−5844	0.282	0.093	0.087	9.2	2.0
SPT-CL J2331−5051	0.522	0.308	0.059	16.2	8.3
SPT-CL J2337−5942	0.425	0.205	0.059	18.3	7.6
SPT-CL J2341−5119	0.320	0.122	0.067	19.1	9.3
SPT-CL J2342−5411	0.300	0.105	0.082	15.8	2.5
SPT-CL J2359−5009	0.423	0.204	0.055	16.6	8.7

Note. Column 1: Cluster designation. *Columns 2–4:* $\langle\beta\rangle$, $\langle\beta^2\rangle$ and $\sigma_{\langle\beta\rangle_i}/\langle\beta\rangle$ averaged over both colour selection schemes and all magnitude bins that are included in the NFW fits according to their corresponding shape weight sum. *Columns 5–6:* Density of selected sources in the cluster fields for the ACS-only and the ACS+FORS2 colour selection schemes, respectively.

[7] Here we want to investigate how well we can estimate the cosmic mean redshift distribution from CANDELS, for which sub-patches are not needed. The sub-patches are needed to estimate the line-of-sight scatter in $\langle\beta\rangle$ between the different cluster fields, as discussed in the second paragraph of this subsection.

[8] Here we account for the magnitude-dependence of our colour cut (see Table D1 in Appendix D), by basing it on the lensed magnitude.

[9] When computing the *relative* impact of magnification on the number density and mean lensing efficiency we deliberately do not include the shape weights, as we would otherwise need to account for the increase in S/N and thus w due to the magnification. Since we perform the full analysis in magnitude bins, with very little variation in w within a bin, our approach constitutes a very good approximation.

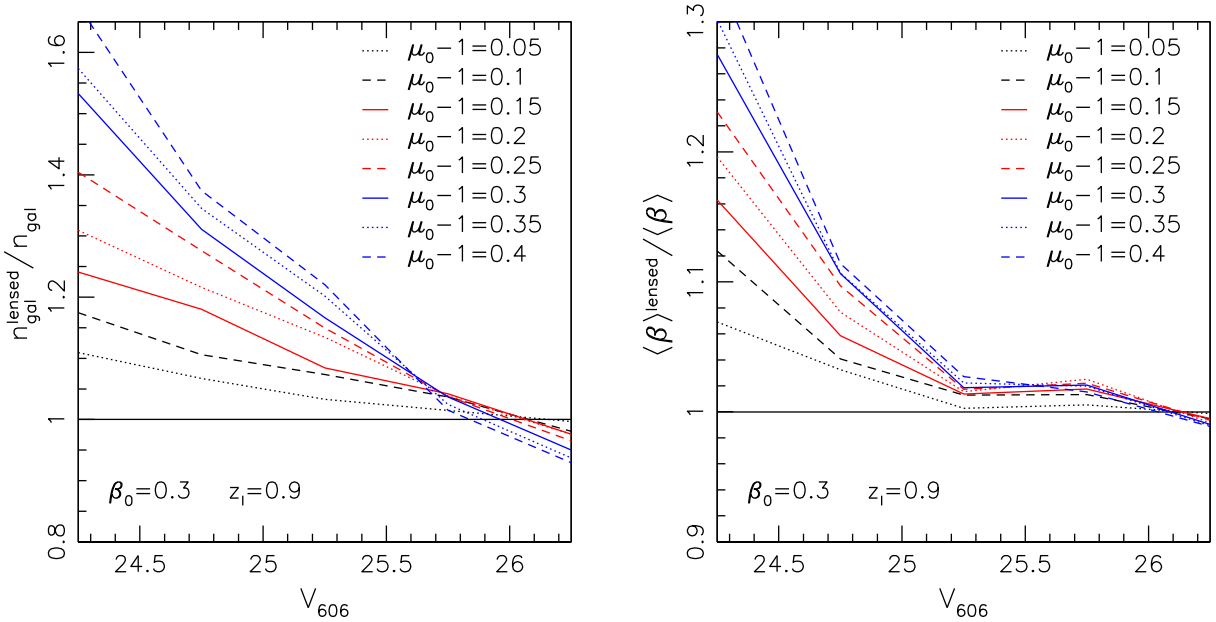

Figure 9. Relative change in the source density $n_{\rm gal}$ (*left*) and the average geometric lensing efficiency $\langle\beta\rangle$ (*right*) for galaxies in the GOODS-South and GOODS-North fields as a function of the V_{606} aperture magnitude when applying an artificial weak lensing magnification for $z_{\rm l} = 0.9$, $\beta_0 = 0.3$, and a range of μ_0 as indicated in the legend, assuming the linear scaling from equation (23). This analysis uses the statistically corrected photometric redshift estimates $z_{\rm f}$ for all galaxies in the S14 GOODS-South and GOODS-North catalogues which are located within the ACS+WFC3 area and pass our ACS-only colour selection.

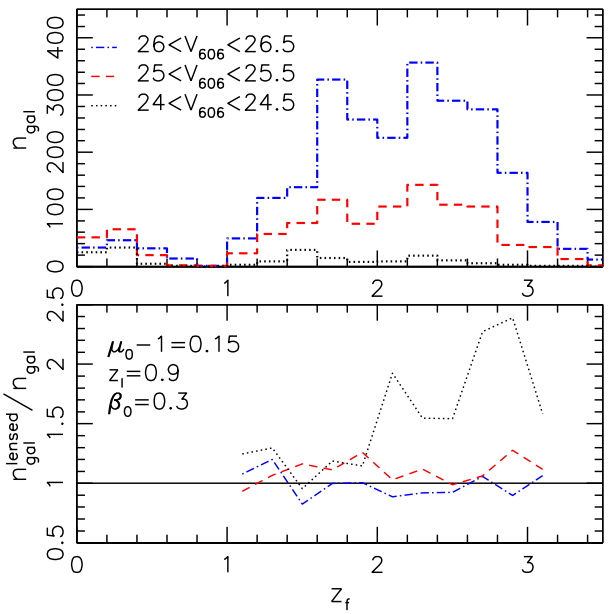

Figure 10. *Top:* Distribution of the statistically corrected photometric redshifts $z_{\rm f}$ for galaxies in the S14 GOODS-South and GOODS-North catalogues located in the ACS+WFC3 area when applying our ACS-only colour selection. The different histograms correspond to three different V_{606} magnitude bins. *Bottom:* Relative change in those redshift distributions after application of weak lensing magnification for a lens at $z_l = 0.9$ with $\mu_0 - 1 = 0.15$ and $\beta_0 = 0.3$.

strongly depends on the depth of the observations. For illustration we also show the redshift distributions within three magnitude bins and their relative change after applying magnification with $z_{\rm l} = 0.9$ and $\mu_0 - 1 = 0.15$ in Fig. 10.

Previous weak lensing magnification studies have made the simplifying assumptions that sources are located at a single redshift and that the source counts can be described as a power law. Under these assumptions the ratio of the lensed and non-lensed cumulative source densities above a magnitude $m_{\rm cut}$

$$\frac{n(<m_{\rm cut})}{n_0(<m_{\rm cut})} = \mu^{2.5s-1} \tag{27}$$

depends only on the magnification μ and the slope of the logarithmic cumulative number counts

$$s = \frac{{\rm d}\log_{10} n(<m)}{{\rm d}m} \tag{28}$$

(e.g. Broadhurst, Taylor & Peacock 1995; Chiu et al. 2016b), where slopes $s > 0.4$ ($s < 0.4$) lead to a net increase (decrease) of the counts. As an illustration we estimate s from our colour ($V_{606} - I_{814} < 0.3$) and shape-selected GOODS-South and GOODS-North catalogue, finding that it can approximately be described as

$$s(V_{606}) \simeq +0.88 \pm 0.03 - (0.15 \pm 0.02)(V_{606} - 24) \tag{29}$$

for $24 < V_{606} < 26.5$. Under these simplifying assumptions we therefore expect a significant boost in the source density at bright magnitudes ($V_{606} \simeq 24$–25) where the slope of the number counts is steep, and only a small boost towards fainter magnitudes ($V_{606} \simeq 26.5$), where the slope of the number counts is shallower. This roughly agrees with the more accurate results shown in Fig. 9, but there are noticeable differences, such as the slight net decrease in the source density at $V_{606} \simeq 26.5$ in Fig. 9. As our sources are not at a single redshift, the simplifying assumptions are clearly not met, which is why we base our analysis on the more accurate approach described above.

When fitting the reduced cluster shear profiles with NFW models in Section 7, we compute a $\mu(\langle\beta\rangle_j, r)$ profile for magnitude bin j and a given mass from the NFW model predictions for both $\kappa(\langle\beta\rangle_j, r)$ and $\gamma(\langle\beta\rangle_j, r)$ according to equation (15). Employing equation (23) with $\beta = \langle\beta\rangle_j$ we compute the corresponding $(\mu_0 - 1)(r)$ profile and obtain radius-dependent corrections $\langle\beta\rangle_j^{\rm lensed}/\langle\beta\rangle_j(r)$ and $n_{{\rm gal},j}^{\rm lensed}/n_{{\rm gal},j}(r)$ by interpolating our CANDELS-based estimates

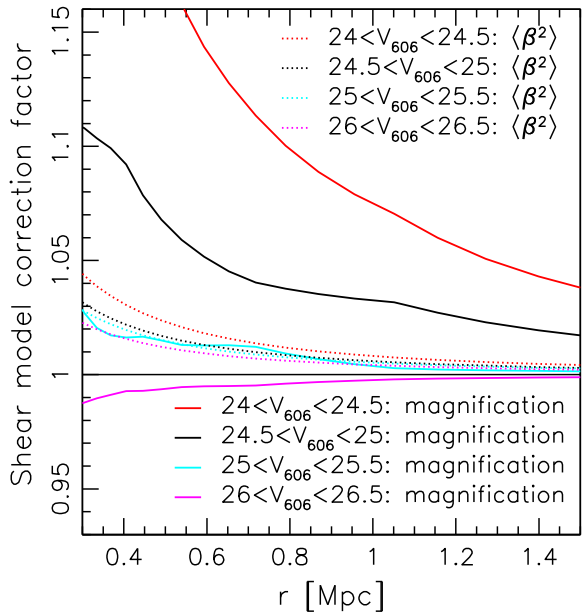

Figure 11. Correction factors to the reduced shear profile model of a $M_{200c} = 7 \times 10^{14}\,M_\odot$ galaxy cluster at $z_l = 0.87$ due to the impact of magnification on the source redshift distribution (solid curves) and the finite width of the redshift distribution ($\langle\beta^2\rangle$, see equation 12, dotted). The different colours correspond to different bins in the V_{606} aperture magnitude.

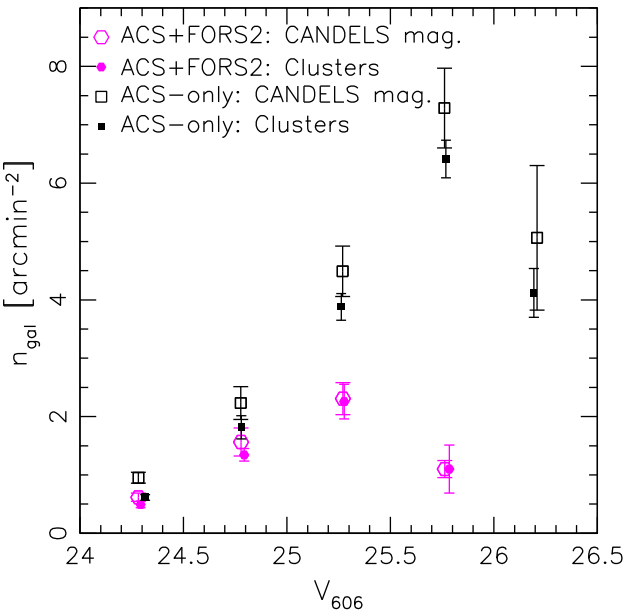

Figure 12. Selected source density $n_{\rm gal}$ as a function of V_{606} accounting for masks and averaged over all the cluster fields (small solid symbols) and the corresponding source density averaged over the CANDELS fields when mimicking the same selection and accounting for photometric scatter and magnification (large open symbols). Black squares show the ACS-only selection, while magenta hexagons correspond to the ACS+FORS2 selection. We include only galaxies located within the fit range of the shear profiles (see Section 7.2) to avoid limitations of the magnification correction at small radii. The error-bars show the uncertainty on the mean from the variation between the contributing cluster fields or the five CANDELS fields, respectively, assuming Gaussian scatter. They are correlated between magnitude bins due to large-scale structure.

(Fig. 9) between the discrete μ_0 values. The fact that we compute the magnification in the NFW prediction from both κ and γ is our primary motivation to conduct the interpolation in terms of μ_0 and not κ_0. This provides a more accurate correction than if the shear contribution is ignored, even though we assume the linear scaling in β in equation (23) to simplify the CANDELS analysis.

On average the application of the correction for magnification-induced changes in the redshift distribution reduces our estimated cluster masses by 3 per cent. This net impact is relatively small since the majority of our sources are at $V_{606} > 25$, requiring small corrections. Also, we exclude the cluster cores, where the correction is the largest (see Fig. 11), from our tangential shear profile fits (see Section 7). However, we emphasize that weak lensing studies of high-z clusters using shallower data will be affected more strongly and should adequately model this effect.

We note a subtle limitation of our modelling approach for magnification which results from our choice to conduct the analysis as a function of aperture magnitude. Here we ignore the fact that the increase in size due to magnification will lead to a larger fraction of the flux being outside the fixed aperture radius than without magnification. As a test we also conducted the magnification analysis using aperture-corrected magnitudes from CANDELS,[10] finding similar models as in Fig. 9 but shifted to brighter magnitudes, with $\langle\beta\rangle^{\rm lensed}/\langle\beta\rangle$ reaching unity at $V_{606}^{\rm tot} \simeq 25.0$–$25.5$. Given the very minor impact magnification has for our data compared to the statistical uncertainties, the described subtle limitation can safely be ignored for the current study. In the future this can be avoided by computing aperture corrections in the filter used for shape measurements both for CANDELS and the cluster fields.

[10] The 3D-HST CANDELS catalogue provides aperture magnitudes, which we can directly compare to our measurements, plus an aperture correction based on the H band, which is however not available for our cluster fields.

6.8 Number density consistency tests

The measurements of the total source density and its radial dependence can be used to test the cluster member removal and our procedure to consistently select galaxies in the cluster and CANDELS fields in the presence of noise (Section 6.4). When computing the source density we account for masks and apply an approximate correction[11] for the impact of obscuration by cluster members (Simet & Mandelbaum 2015). We also account for the impact of cluster magnification, employing the corresponding radius-dependent magnification model for each cluster from Section 6.7.

6.8.1 Total source density

In Fig. 12 we compare the average density of selected source galaxies in the cluster fields as a function of V_{606} to the corresponding

[11] Here we approximate the sky area blocked by a galaxy through the $N_{\rm pix}$ parameter from SOURCE EXTRACTOR. Hoekstra et al. (2015) present a more detailed treatment using image simulations, finding that obscuration by cluster members is a relatively minor effect for their analysis. Our cluster galaxies are at higher redshift and are thus more strongly dimmed, leading to an even smaller impact of obscuration by cluster members. Our pipeline automatically masks the image region around bright and very extended galaxies. With this applied we find that accounting for the sky area blocked by unmasked brighter galaxies via the $N_{\rm pix}$ parameter leads to $\lesssim$1 per cent changes in the source density even for the faintest galaxies considered in our analysis.

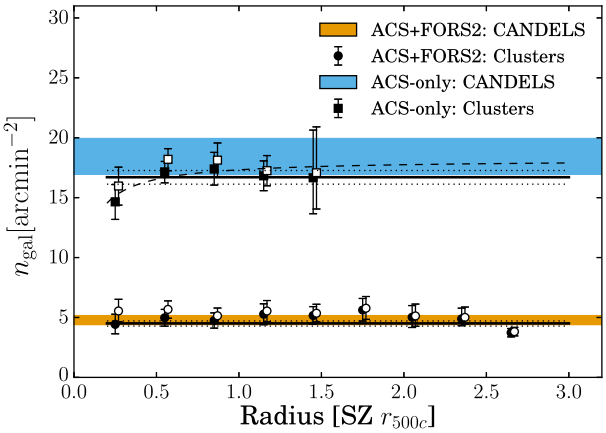

Figure 13. Density of sources $n_{\rm gal}$ for the ACS-only and ACS+FORS2 colour selections as a function of cluster-centric distance around the X-ray centre in units of r_{500c}, as estimated from the SZ signal. The profiles account for both masks and obscuration by cluster members. Solid (open) symbols include (do not include) the correction for magnification. The coloured regions indicate the 1σ constraints on the mean background density from the five CANDELS fields. Black solid and dotted lines show the maximum likelihood and 68 per cent uncertainty range for a constant density model. The dashed curve shows for the ACS-only selection the maximum likelihood contamination model following a $1/r$ functional form.

average density in the CANDELS fields corrected for the expected influence of magnification given our best-fitting NFW cluster mass models. There is very good agreement for the ACS+FORS2 selection and reasonable agreement for the ACS-only selection (errorbars are correlated because of large-scale structure). Fig. 12 also visualizes that the ACS-only analysis (with two ACS bands) provides a substantially higher average total density of selected sources in the cluster fields of 16.8 galaxies arcmin^{-2} compared to 5.2 galaxies arcmin^{-2} for the ACS+FORS2 colour selection (see Table 3 for the total source densities in each field). This shows that either substantially deeper ground-based imaging or ACS-based colours for the full imaging area would be required for the colour selection in order to adequately exploit the full depth of the ACS shape catalogues.

6.8.2 Source density profile

As an additional cross-check for the removal of cluster galaxies and our magnification model we plot the radial source density profiles for the ACS-only and ACS+FORS2 selected samples in Fig. 13, averaged over all clusters, as a function of cluster-centric distance from the X-ray centre (nearly identical results are obtained when using the SZ peak location, see Section 7) in units of their corresponding r_{500c} as estimated from the SZ signal. Here we compare the case with and without applying the magnification correction. The difference is small given that the magnification is relatively weak for the majority of the clusters. Also, most of the source galaxies have faint magnitudes, where the net impact of magnification is small (see Fig. 9). The net difference is strongest for the inner cluster regions where the magnification is strongest.

In the case of a complete cluster galaxy removal and an accurate correction for magnification we expect to measure a number density that is consistent with being flat as a function of radius. To test this, we perform a model comparison test using the cluster sample-averaged number density profiles, assuming that errors were independent and Gaussian distributed. Each radial bin used in the test is the average of at least three clusters, while uncertainties are determined by bootstrapping the cluster sample. With these measures, the χ^2 statistic should be a crude yet useful approximation to the true uncertainty distribution, while allowing us to use analytic model quality of fit and comparison tests.

As expected for adequate removal of cluster members and magnification correction, we find that the source density profiles are consistent with being flat. For the ACS-only case, the maximum likelihood constant number density model returns a $\chi^2 = 2.54$ with 4 degrees of freedom, while a $1/r$ inverse-r profile with two parameters, the contamination fraction f_{500} at r_{500c} and the background number density (Hoekstra 2007), returns a $\chi^2 = 0.74$ with 3 degrees of freedom. Both are acceptable models at $p > 0.05$, where the improvement in χ^2 is consistent with random according to an F-test ($p > 0.05$). The rather low χ^2 values might be due to our assumption of independent errors between bins, which neglects the effects of large-scale structure. For the ACS+FORS2 selection, the constant number density model returns $\chi^2 = 9.02$ with 8 degrees of freedom, while an exponential model (see Appendix E and Applegate et al. 2014), which is preferred over the inverse-r model in this case, returns $\chi^2 = 7.74$ with 7 degrees of freedom, again suggesting that a flat number density model is sufficient ($p > 0.05$ from the F-test). For a general test for the consistency of the combined number density profile being flat we allow for negative f_{500} in these fits, which could for example be mimicked by an incorrect magnification correction. The maximum likelihood parameter value for the inverse-r contamination model fit to the ACS-only number density profile indeed peaks at slightly, but not significantly negative values, $f_{500} = -0.050^{+0.038}_{-0.052}$. Fig. 13 shows the measured number density profiles and maximum-likelihood model fits for both selections.

7 WEAK LENSING CONSTRAINTS AND MASS ANALYSIS

We reconstruct the projected mass distribution in our cluster fields in Section 7.1 and constrain the cluster masses via fits to the tangential reduced shear profile in Section 7.2. In Section 7.3 we compare the stacked shear profiles from all clusters for the different centres used in the analysis and investigate the consistency of the data with different concentration–mass relations. In Section 7.4 we detail on the simulations used to calibrate the mass estimates. We discuss the systematic error budget in Section 7.5.

7.1 Mass maps

The weak lensing shear and convergence are linked as they are both based on second-order derivatives of the lensing potential. Therefore, a reconstruction of the convergence field can be obtained from the shear field up to a constant (Kaiser & Squires 1993), which is the mass-sheet degeneracy (Schneider & Seitz 1995). Motivated by the different colour-selected source densities in the inner and outer regions of our clusters we employ a Wiener filter for the convergence reconstruction using an implementation described in McInnes et al. (2009) and Simon, Taylor & Hartlap (2009). This code estimates the convergence on a grid taking the spatial variation in the source number density into account; it applies more smoothing where the number density of sources is lower. The smoothing in the Wiener filtered map employs the shear two-point correlation function $\xi_+(\theta)$ (e.g. Schneider 2006) as a prior on the angular correlation of the convergence, which affects the degree of smoothing. For this, we measure $\xi_+(\theta)$ in the cluster fields and find that it is on average approximately described by the fitting function

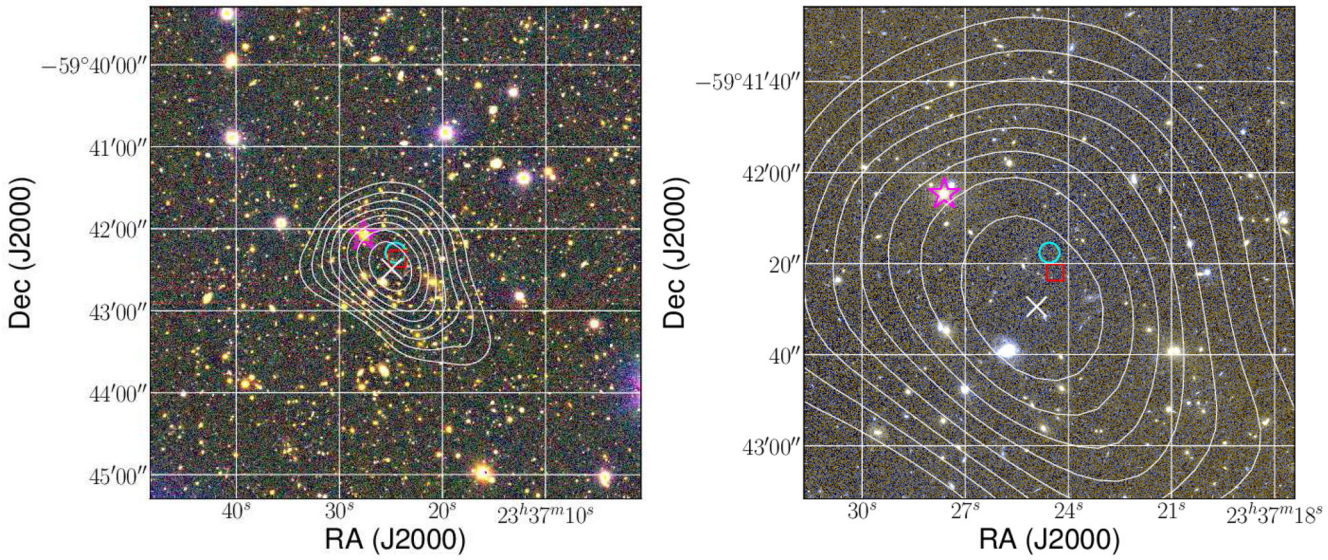

Figure 14. Contours of the signal-to-noise map of the Wiener-filtered mass reconstruction for SPT-CL J2337−5942, starting at 2σ in steps of 0.5σ with the cross indicating the peak location, overlaid on a VLT/FORS2 *BIz* colour image (*left*, 6 arcmin × 6 arcmin), as well as a 1.8 arcmin × 1.8 arcmin cut-out of the ACS imaging (*right*, using *F*606*W* as blue, *F*814*W* as red and the sum $F606W + 2 \times F814W$ as green channel). The cyan circle, red square and magenta star indicate the positions of the SZ peak, X-ray centroid and BCG, respectively. The corresponding plots of the other clusters are shown in Appendix G.

Table 4. Locations (α, δ) of the mass map signal-to-noise peaks of the clusters, their positional uncertainty ($\Delta\alpha$, $\Delta\delta$) estimated by bootstrapping the galaxy catalogue and the peak signal-to-noise ratio.

Cluster	α (deg J2000)	δ (deg J2000)	$\Delta\alpha$ (arcsec)	$\Delta\delta$ (arcsec)	$\Delta\alpha$ (kpc)	$\Delta\delta$ (kpc)	$(S/N)_{peak}$
SPT-CL J0000−5748	0.251 95	−57.808 75	1.9	2.2	14	16	5.7
SPT-CL J0102−4915	15.717 43	−49.254 58	7.1	7.9	55	61	5.7
SPT-CL J0533−5005	83.397 72	−50.099 84	10.2	8.0	79	62	3.0
SPT-CL J0546−5345	86.653 96	−53.758 61	5.1	3.7	41	30	3.6
SPT-CL J0559−5249	89.928 75	−52.822 97	4.3	3.6	29	24	5.0
SPT-CL J0615−5746	93.965 62	−57.779 79	4.3	2.8	34	23	5.1
SPT-CL J2040−5725	310.063 89	−57.422 32	5.0	5.1	40	40	3.1
SPT-CL J2106−5844	316.522 10	−58.743 36	7.2	4.5	59	37	2.9
SPT-CL J2331−5051	352.965 21	−50.863 60	1.8	2.3	12	15	5.1
SPT-CL J2337−5942	354.353 84	−59.708 19	1.8	2.5	13	19	6.0
SPT-CL J2341−5119	355.300 57	−51.330 15	5.3	5.4	42	44	3.3
SPT-CL J2342−5411	355.693 05	−54.180 43	3.7	9.9	30	80	3.1
SPT-CL J2359−5009	359.932 13	−50.168 22	4.7	4.8	35	35	5.2

$\xi_+^{\rm fit}(\theta) = 0.012(1 + \theta/{\rm arcmin})^{-2}$. We fix the mass-sheet degeneracy by setting the average convergence inside each cluster field to zero. While this underestimates the overall convergence for our relatively small cluster fields, this is irrelevant as we use the reconstructions to study positional offsets and the signal-to-noise (S/N) ratio of relative mass distributions, but not to obtain quantitative mass constraints. To compute the S/N mass maps we generate 500 noise maps for each cluster by randomizing the ellipticity phases and repeating the mass reconstruction. We then define the S/N mass map as the ratio of the reconstruction from the actual data and the r.m.s. image of the noise reconstructions.

As an example, Fig. 14 shows an overlay of the S/N contours of the mass reconstruction (starting at 2σ in steps of 0.5σ) for SPT-CL J2337−5942 on a FORS2 *BIz* colour image (left) as well as a 1.8 arcmin × 1.8 arcmin cut-out of the ACS imaging (right). Here we also indicate the locations of the X-ray centroid, BCG and SZ peak (see Table 1). The corresponding figures for the other clusters are shown in Appendix G.

For all clusters the weak lensing reconstruction shows a mass peak associated with the cluster, with a peak signal-to-noise ratio $(S/N)_{peak}$ between 2.9σ and 6.0σ. Typically, the mass reconstructions follow the distribution of red cluster galaxies well, especially for the clusters with $(S/N)_{peak} \gtrsim 4$. Of these clusters, SPT-CL J0000−5748 and SPT-CL J2331−5051 show relatively symmetric mass reconstructions consistent with more relaxed morphologies, while SPT-CL J0102−4915, SPT-CL J0559−5249, SPT-CL J0615−5746, SPT-CL J2337−5942 and SPT-CL J2359−5009 show more elongated or perturbed morphologies. In particular, SPT-CL J0102−4915 is known to be an extreme merger (Menanteau et al. 2012), for which our mass reconstruction separates both main components well (see also the independent weak lensing analysis by Jee et al. 2014).

In the mass signal-to-noise maps we determine the position of the mass peak of the corresponding cluster by identifying the pixel with the highest S/N within 90 arcsec from the SZ peak location. We report these positions in Table 4 along with estimates of their

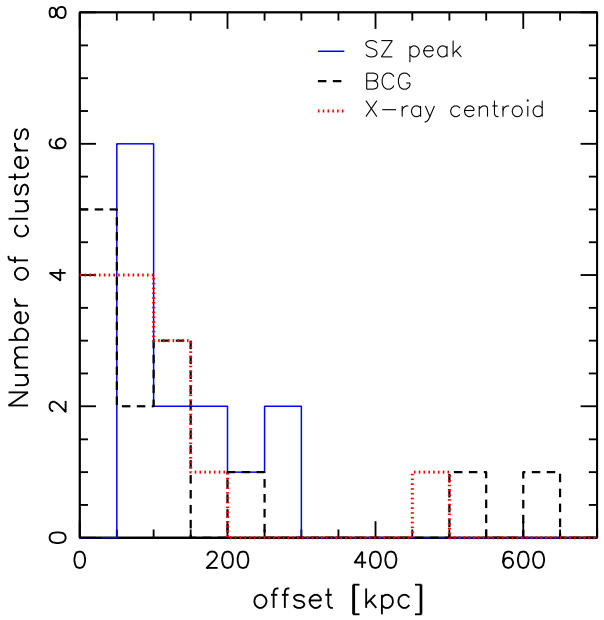

Figure 15. Histograms of spatial offsets between the peak in the mass reconstruction signal-to-noise map and the indicated centres.

uncertainty and peak signal to noise $(S/N)_{peak}$. The positional uncertainties are estimated by generating 500 bootstrap samples of the source catalogue for which we repeat the reconstruction and identification of the peak location. The average r.m.s. positional uncertainty (including both directions) for the full sample is 59 kpc.

Dietrich et al. (2012) investigate the origin of offsets between peaks in weak lensing mass reconstructions and the projected position of the 3D centre (defined as the minimum of the potential) of cluster-scale dark matter haloes in the Millennium Simulation (Springel et al. 2005; Hilbert et al. 2009). Without shape noise and smoothing applied in the mass reconstruction they find very small offsets: their analysis using sources at $z = 3.06$ is most similar to the set-up of our study, yielding a 90th percentile offset of $5.6\,h^{-1}$ kpc. Hence, projection effects and large-scale structure have a negligible impact for the measured offsets in typical observing scenarios. Dietrich et al. (2012) find that smoothing and shape noise increase the offsets substantially, where the addition of shape noise has the biggest impact unless unnecessarily large smoothing kernels are used. Our bootstrap analysis provides an estimate for the positional uncertainty due to shape noise. However, the analysis likely underestimates the true positional uncertainty with respect to the 3D cluster centre as it does not explicitly account for the impact of smoothing. Nevertheless, we can use the distribution of offsets between the peaks in the mass signal-to-noise maps and different proxies for the cluster centre, namely the X-ray centroid, SZ peak, and BCG position, to test if these are similarly good proxies for the true 3D cluster centre position. Fig. 15 shows a histogram of the corresponding offset distributions. We also summarize the average, r.m.s., and median of these offset distributions in Table 5, where errors indicate the dispersion of the corresponding values when bootstrapping the cluster sample. While the X-ray centroids yield the smallest average and median offsets, their r.m.s. offset is similar to the one for the SZ peaks. The distribution of offsets between the BCG locations and the mass signal-to-noise peaks has the largest r.m.s. offset, resulting from two outliers: the largest offset occurs for the merger SPT-CL J0102−4915 (642 kpc), where the BCG is located in the south-eastern component while the highest signal-to-noise peak in the mass reconstruction coincides with the north-western cluster component, which also shows a strong concentration of galaxies but has a less bright BCG (see Fig. G2). In contrast, both the SZ peak and the X-ray centroid are located between the two cluster cores and peaks of the mass reconstruction, resulting in smaller offsets. SPT-CL J0533−5005 also shows a large (522 kpc) offset between the BCG and the mass signal-to-noise peak, while the latter is broadly consistent with the SZ peak, X-ray centroid and strongest galaxy concentration. This could also be explained with a merger scenario, where a smaller component hosting a brighter BCG is falling into the main cluster.

Table 5. Average, r.m.s. and median of the offsets (kpc) between the peaks in the mass reconstruction signal-to-noise maps and the SZ peak, X-ray centroid and BCG location.

Centre	Average	r.m.s.	Median
SZ peak	137 ± 21	158 ± 23	100 ± 13
X-ray centroid	105 ± 30	154 ± 50	64 ± 44
BCG location	159 ± 52	249 ± 72	80 ± 54

Note. Errors indicate the dispersion of the values when bootstrapping the cluster sample.

7.2 Individual shear profile analysis

We compute profiles of the tangential reduced shear (equation 7) around the cluster centres in 14 linearly spaced bins of transverse physical separation between 300 kpc and 1.7 Mpc (100 kpc-wide bins), but note that we restrict the fit range to 500 kpc $\leq r \leq$ 1.5 Mpc when deriving mass constraints. Smaller scales are more susceptible to the impact of miscentring, cluster substructure, uncertainties in the concentration–mass relation and shear calibration, while larger scales suffer from an increasingly incomplete azimuthal coverage, where 1.5 Mpc (1.3 Mpc) equals the largest radius with full azimuthal coverage at the median (lowest) redshift of the targeted clusters. We repeat the analysis for the different proxies for the cluster centre (X-ray centroid, SZ peak and BCG position), but regard the measurements using the X-ray centroids as our primary (default) results, given that they yield the smallest average and median offsets from the peaks in the mass signal-to-noise maps (Section 7.1).

We compute separate tangential reduced shear profiles for each magnitude bin and colour selection scheme, where we use the ACS-only selection in the inner cluster regions where both ACS bands are available, and the ACS+FORS2 selection in the outer cluster regions. Each magnitude bin for both colour selection schemes has a separate value for $\langle\beta\rangle$ and $\langle\beta^2\rangle$ (see Section 6.5, Fig. 8, and average values reported in Table 3), which we correct for magnification as a function of cluster-centric distance as described in Section 6.7. We fit the profiles from all ACS-only magnitude bins plus those ACS+FORS2 bins that have sufficiently low photometric scatter (Section 6.4 and Appendix D3.2) jointly with a reduced shear profile model (see equation 12) according to Wright & Brainerd (2000), assuming spherical mass distributions that follow the NFW density profile (Navarro, Frenk & White 1997). Here we use a fixed concentration–mass ($c(M)$) relation, where we by default employ the $c(M)$ relation from Diemer & Kravtsov (2015), but also test the consistency of the data with other relations in Section 7.3. While the mass distributions in individual clusters may well deviate from an NFW profile, we account for the net impact on an ensemble of clusters in Section 7.4. Due to the fixed concentration–mass relation we fit a one-parameter model to each cluster. Here we perform a χ^2 minimization using M_{200c} as free parameter,

Table 6. Weak lensing mass constraints from the NFW fits to the reduced shear profiles using scales 500 kpc $< r <$ 1.5 Mpc and the Diemer & Kravtsov (2015) $c(M)$ relation for two different overdensities $x \in \{200c, 500c\}$. Columns 2–5 correspond to the default analysis centring around the X-ray centroids, while columns 6–9 list results for the analysis centring around the SZ peaks. $M_x^{\mathrm{biased,ML}}$ are the maximum likelihood mass estimates in $10^{14}\,\mathrm{M}_\odot$ *without* bias correction applied. All errors are statistical 68 per cent uncertainties, listing the contributions from shape noise (asymmetric errors), uncorrelated large-scale and line-of-sight variations in the redshift distribution. Systematic uncertainties are listed in Table 8. The factor b_x indicates the expected mass bias factor for the scaling relation analysis when the full likelihood distribution of the mass constraints is used.

	X-ray centres				SZ centres			
Cluster	$M_{200c}^{\mathrm{biased,ML}}$	b_{200c}	$M_{500c}^{\mathrm{biased,ML}}$	b_{500c}	$M_{200c}^{\mathrm{biased,ML}}$	b_{200c}	$M_{500c}^{\mathrm{biased,ML}}$	b_{500c}
SPT-CL *J*0000−5748	$6.2^{+2.6}_{-2.4} \pm 1.1 \pm 0.5$	0.91	$4.2^{+1.8}_{-1.6} \pm 0.7 \pm 0.3$	0.88	$6.5^{+2.6}_{-2.5} \pm 1.1 \pm 0.5$	0.80	$4.5^{+1.8}_{-1.7} \pm 0.7 \pm 0.3$	0.82
SPT-CL *J*0102−4915	$11.1^{+2.9}_{-2.8} \pm 1.2 \pm 1.1$	0.86	$7.9^{+2.2}_{-2.1} \pm 0.9 \pm 0.8$	0.88	$14.4^{+2.8}_{-2.8} \pm 1.2 \pm 1.5$	0.79	$10.3^{+2.1}_{-2.1} \pm 0.9 \pm 1.1$	0.79
SPT-CL *J*0533−5005	$4.3^{+2.7}_{-2.4} \pm 1.0 \pm 0.4$	0.88	$2.9^{+1.9}_{-1.6} \pm 0.7 \pm 0.3$	0.87	$2.4^{+2.4}_{-1.8} \pm 1.0 \pm 0.2$	0.80	$1.6^{+1.7}_{-1.2} \pm 0.7 \pm 0.1$	0.81
SPT-CL *J*0546−5345	$5.4^{+3.7}_{-3.3} \pm 1.1 \pm 0.7$	0.86	$3.7^{+2.6}_{-2.3} \pm 0.8 \pm 0.5$	0.85	$2.6^{+3.5}_{-2.4} \pm 1.1 \pm 0.3$	0.72	$1.8^{+2.4}_{-1.6} \pm 0.8 \pm 0.2$	0.73
SPT-CL *J*0559−5249	$8.0^{+3.1}_{-2.9} \pm 1.0 \pm 0.8$	0.79	$5.4^{+2.2}_{-2.0} \pm 0.7 \pm 0.5$	0.81	$4.7^{+2.9}_{-2.5} \pm 1.0 \pm 0.5$	0.84	$3.2^{+2.0}_{-1.7} \pm 0.7 \pm 0.3$	0.85
SPT-CL *J*0615−5746	$6.8^{+2.9}_{-2.6} \pm 1.0 \pm 0.8$	0.88	$4.7^{+2.0}_{-1.8} \pm 0.7 \pm 0.5$	0.85	$5.8^{+2.8}_{-2.5} \pm 1.0 \pm 0.7$	0.82	$3.9^{+1.9}_{-1.7} \pm 0.7 \pm 0.4$	0.80
SPT-CL *J*2040−5726	$2.1^{+2.9}_{-1.9} \pm 0.8 \pm 0.2$	0.87	$1.4^{+2.0}_{-1.3} \pm 0.6 \pm 0.2$	0.81	$2.1^{+2.9}_{-2.0} \pm 0.8 \pm 0.2$	0.80	$1.4^{+2.0}_{-1.3} \pm 0.6 \pm 0.2$	0.80
SPT-CL *J*2106−5844	$8.8^{+5.0}_{-4.6} \pm 1.5 \pm 1.1$	0.85	$6.1^{+3.7}_{-3.3} \pm 1.1 \pm 0.8$	0.86	$8.2^{+5.0}_{-4.3} \pm 1.5 \pm 1.1$	0.81	$5.7^{+3.6}_{-3.1} \pm 1.1 \pm 0.7$	0.78
SPT-CL *J*2331−5051	$3.8^{+2.5}_{-2.1} \pm 1.1 \pm 0.3$	0.85	$2.6^{+1.7}_{-1.4} \pm 0.7 \pm 0.2$	0.92	$4.0^{+2.5}_{-2.1} \pm 1.1 \pm 0.4$	0.85	$2.7^{+1.7}_{-1.4} \pm 0.7 \pm 0.2$	0.87
SPT-CL *J*2337−5942	$10.5^{+2.9}_{-2.8} \pm 1.3 \pm 0.9$	0.88	$7.2^{+2.1}_{-2.0} \pm 0.9 \pm 0.7$	0.91	$10.0^{+2.9}_{-2.8} \pm 1.3 \pm 0.9$	0.82	$6.9^{+2.1}_{-2.0} \pm 0.9 \pm 0.6$	0.83
SPT-CL *J*2341−5119	$2.4^{+2.5}_{-1.9} \pm 1.1 \pm 0.2$	0.91	$1.6^{+1.7}_{-1.3} \pm 0.7 \pm 0.2$	0.89	$2.3^{+2.5}_{-1.8} \pm 1.1 \pm 0.2$	0.80	$1.5^{+1.7}_{-1.2} \pm 0.7 \pm 0.1$	0.80
SPT-CL *J*2342−5411	$8.6^{+3.8}_{-3.5} \pm 1.4 \pm 1.1$	0.87	$6.0^{+2.8}_{-2.5} \pm 1.0 \pm 0.7$	0.84	$7.0^{+3.7}_{-3.4} \pm 1.4 \pm 0.9$	0.79	$4.8^{+2.7}_{-2.4} \pm 1.0 \pm 0.6$	0.81
SPT-CL *J*2359−5009	$5.0^{+3.0}_{-2.6} \pm 1.1 \pm 0.4$	0.91	$3.4^{+2.1}_{-1.8} \pm 0.8 \pm 0.3$	0.92	$5.7^{+2.8}_{-2.5} \pm 1.1 \pm 0.5$	0.83	$3.9^{+1.9}_{-1.7} \pm 0.8 \pm 0.3$	0.84

comparing the predicted reduced shear values to the measured values in each contributing bin in radius and magnitude. Given its dominance we employ a pure shape noise covariance matrix derived from our empirical weighting scheme (see Appendix A). In the fit we also allow for $M_{200c} < 0$, which we model by switching the sign of the tangential reduced shear profile. For the calibration of scaling relations we make use of the full likelihood distribution (see Section 8). In addition, we identify the maximum likelihood (minimum χ^2) location and the $\Delta\chi^2 = 1$ points, which we report in Table 6, where conversions between overdensity masses use the assumed $c(M)$ relation. Note that the derived mass constraints are expected to be biased due to effects in the mass modelling such as miscentring, which will be addressed in Section 7.4.

The statistical errors in Table 6 include two additional minor noise sources. The first source is given by line-of-sight variations in the redshift distribution, which we estimate from the dispersion $\sigma_{\langle\beta\rangle_i}$ of the estimates $\langle\beta\rangle_i$ from the CANDELS subpatches (see Section 6.6). In Table 3 we report $\sigma_{\langle\beta\rangle_i}/\langle\beta\rangle$, which introduces an additional relative noise in the mass estimates of $\sigma_{M,z}/M \simeq 1.5\sigma_{\langle\beta\rangle_i}/\langle\beta\rangle$, where $M \in \{M_{200c}, M_{500c}\}$. Further statistical noise is added by projections of uncorrelated large-scale structure (Hoekstra 2001). To estimate it we compute 500 random realizations of the cosmic shear field per cluster for our reference cosmology and the colour-selected source redshift distribution as detailed in appendix B of Simon (2012), with the non-linear matter power spectrum estimated following Takahashi et al. (2012).[12] We add these cosmic shear field realizations to the measured shear field in the SPT cluster fields and repeat the cluster mass analysis for each realization. The dispersion in the best-fitting mass estimates then yields an estimate for the large-scale structure noise. We find that it amounts to 30−50 per cent of the statistical errors from shape noise. Additional scatter between profile-fitted weak lensing mass estimates and halo masses defined via spherical overdensities is caused by halo triaxiality, variations in cluster density profiles and correlated large-scale structure (e.g. Gruen et al. 2015; Umetsu et al. 2016). This scatter typically amounts to ∼20 per cent for massive clusters (Becker & Kravtsov 2011) and is not explicitly listed in Table 6. Instead, we absorb it in the intrinsic scatter accounted for in the scaling relation analysis (see Section 8 and Dietrich et al. 2017).

For visualization we show profiles in Figs 16 and G1–G12, where we have combined shear estimates from the different magnitude bins and colour selections for the analysis using the X-ray centroids as centres. Here we stack all profiles scaled to the same average $\langle\beta\rangle$ of all magnitude bins of the cluster as

$$\langle g_\mathrm{t}\rangle_\mathrm{comb}(r_k) = \sum_{j\in\mathrm{mag\,bins}} \langle g_\mathrm{t}\rangle_j(r_k)\frac{\langle\beta\rangle}{\langle\beta\rangle_j} W_{j,k} \Big/ \sum_{j\in\mathrm{mag\,bins}} W_{j,k}, \qquad (30)$$

where k indicates the radial bin, j the magnitude bin and colour selection scheme, and $W_{j,k} = (\langle\beta\rangle_j/\langle\beta\rangle)^2\sum w_i$ is the rescaled sum of the shape weights of the contributing galaxies.

7.3 Stacked signal and constraints on the average cluster concentration

Miscentring reduces the shear signal at small radii. To test if our data show signs for this, we compare the stacked signal for the different centres (top panel of Fig. 17). To stack the signal from clusters at different redshifts and lensing efficiencies we employ the differential surface mass density $\Delta\Sigma(r)$ (see equation 14), where we compute Σ_crit using the $\langle\beta\rangle$ of the corresponding magnitude bin and colour

[12] This approach generates Gaussian random shear fields based on the matter power spectrum. Comparing the resulting scatter in cluster mass estimates, Hoekstra et al. (2011b) show that approaches using the shear power spectrum provide good approximations to more accurate estimates from a ray-tracing analysis through the Millennium Simulation (Springel et al. 2005; Hilbert et al. 2009).

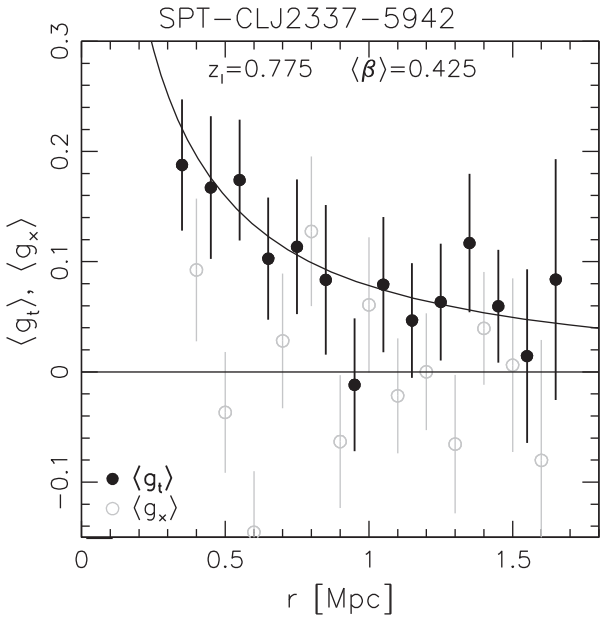

Figure 16. Tangential reduced shear profile (black solid circles) of SPT-CL J2337−5942 centred on the X-ray centroid and obtained by combining the profiles of all contributing magnitude bins of the ACS-only plus the ACS+FORS2 selection (see Section 7.2). The curve shows the correspondingly combined best-fitting NFW model prediction obtained by fitting the data in the range 500 kpc $\leq r \leq$ 1.5 Mpc and using the Diemer & Kravtsov (2015) concentration–mass relation. The grey open circles indicate the 45 degrees-rotated reduced cross-shear component, which is a test for systematics, shifted by d$r = -0.05$ Mpc for clarity. The corresponding plots of the other clusters are shown in Appendix G.

selection scheme. Our clusters span a significant range in mass. Here we expect higher $\Delta\Sigma(r)$ profiles for the more massive clusters. Before stacking, we therefore scale them to approximately the same signal amplitude. For this we compute a theoretical NFW model for the differential surface mass density $\Delta\Sigma_{\rm model}$ for each cluster assuming its mass inferred from the SZ signal $M_{\rm 500c,SZ}$ (Bleem et al. 2015)[13] and a fixed $c_{\rm 200c} = 4$, and then scale the cluster signal as

$$\Delta\Sigma^*(r) = s\Delta\Sigma(r) \equiv \frac{\langle\Delta\Sigma_{\rm model}(800\,{\rm kpc})\rangle}{\Delta\Sigma_{\rm model}(800\,{\rm kpc})}\Delta\Sigma(r). \quad (31)$$

We evaluate the theoretical model at an intermediate scale $r = 800$ kpc, but note that the exact choice is not important as we are only interested in an approximate rescaling to optimize the weighting. We then compute the weighted average

$$\langle\Delta\Sigma^*\rangle(r_j) = \sum_{i\in{\rm clusters}} \Delta\Sigma_i^*(r_j)\hat{W}_{ij} \Big/ \sum_{i\in{\rm clusters}} \hat{W}_{ij}, \quad (32)$$

with $\hat{W}_{ij} = \left(s\sigma(\Delta\Sigma(r_j))\right)^{-2}$ and $\sigma(\Delta\Sigma(r_j))$ indicating the 1σ uncertainty of $\Delta\Sigma(r_j)$.

The results are shown in the top panel of Fig. 17. We first note that the stacked profiles are fairly similar for the different centre definitions. This is also the case for an analysis centred on the peaks in the weak lensing mass reconstruction. Such an analysis should not suffer from miscentring, but is rather expected to deliver shear estimates that are biased high (see e.g. Dietrich et al. 2012). The

[13] We weight according to the SZ mass and not the lensing-inferred mass. The latter is more noisy and would give higher weight to clusters for which the lensing mass estimate scatters up.

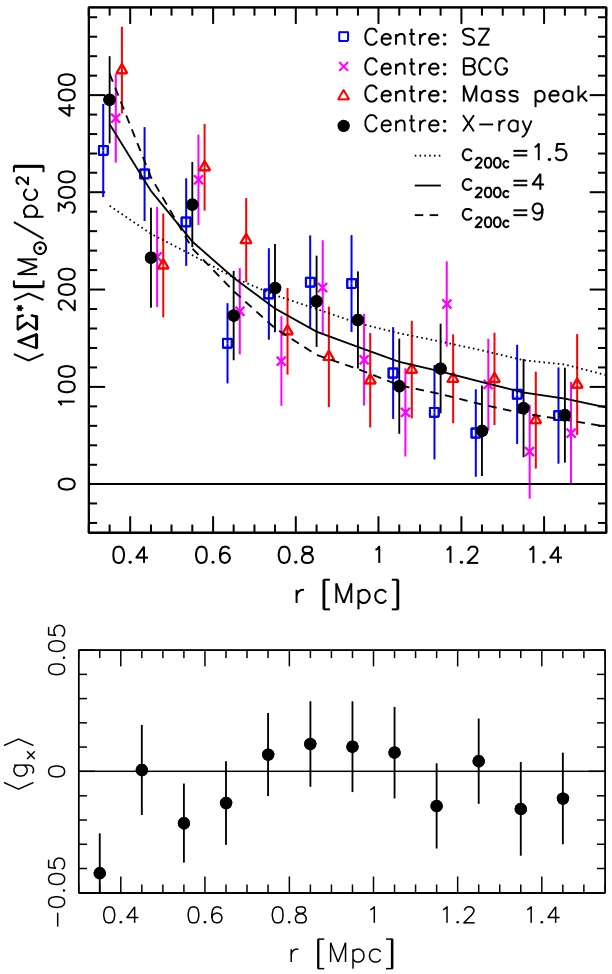

Figure 17. *Top:* Weighted average of the rescaled differential surface mass density profiles from all clusters. The circles, squares, crosses and triangles show the signal measured around the X-ray centroids, the SZ peak positions, the BCG locations and the weak lensing mass peaks, respectively. The circles showing the signal around the X-ray centroids are displayed at the correct radius, while the other symbols are shown with a horizontal offset for clarity. The curves show the correspondingly averaged best-fitting model predictions for different fixed concentrations for the analysis employing the X-ray centres and using an extended fit range 300 kpc to 1.5 Mpc, which increases the sensitivity for constraints on the average concentration. *Bottom:* Profile of the stacked reduced cross-shear component of all clusters measured with respect to their X-ray centres.

similarity of the shear profiles suggests that, for the sample as a whole, miscentring appears to have relatively minor impact at the radial scales considered in our analysis.

We also fit the reduced shear profiles of all clusters using models with different fixed concentrations. For three of these fixed concentrations and the analysis using the X-ray centres we show the averaged best-fitting models from all clusters in Fig. 17, using the same scale factors and weights as used for the data. In these fits we use an extended fit range 300 kpc to 1.5 Mpc to increase the sensitivity of the data for constraints on the concentration, which are mostly derived from the change in the slope between small and large radii (compare Fig. 17). Adding the χ^2 from the individual clusters with equal weights we compute the total $\chi^2_{\rm tot}$ of the sample as a function of the fixed concentration, allowing us to place constraints

on the average concentration of the sample[14] to $c_{200c} = 5.6^{+3.7}_{-1.8}$ using the X-ray centres (χ^2_{tot}/d.o.f. = 747.3/744), $c_{200c} = 4.9^{+3.1}_{-1.7}$ using the SZ centres (χ^2_{tot}/d.o.f. = 754.6/712), $c_{200c} = 5.5^{+3.5}_{-1.8}$ using the BCG centres (χ^2_{tot}/d.o.f. = 754.2/774) and $c_{200c} = 6.5^{+3.6}_{-2.2}$ (χ^2_{tot}/d.o.f. = 749.2/752) when centring on the weak lensing mass peaks. We stress that the fitting was conducted for each cluster separately (see Section 7.2), and that the stacked signal shown in Fig. 17 is for illustrative purposes only. This is important given that the scaling is only approximate, while the individual analyses account for all effects (e.g. reduced shear).

Due to miscentring the estimates using the X-ray, SZ and BCG centres may be slightly biased low, while the estimate based on the mass peak centre is likely biased high. Given that all constraints are well consistent within the uncertainties, we conclude that miscentring has a negligible impact for the constraints on concentration at the current statistical precision. These estimates are consistent with predictions from recent numerical simulations. In particular, the $c(M)$ relation from Diemer & Kravtsov (2015), which corresponds to our default analysis, yields average concentrations $3.5 \lesssim c_{200c} \lesssim 4.6$ (average 3.8) in our mass and redshift range, fully consistent with our constraints. Accordingly, it is not surprising that it provides similarly good fits to the data as the best-fitting fixed concentrations, e.g. we obtain χ^2_{tot}/d.o.f. = 748.8/745 with the Diemer & Kravtsov (2015) $c(M)$ relation for the analysis using the X-ray centres. For comparison, the $c(M)$ relation from Duffy et al. (2008) yields lower average concentrations $2.4 \lesssim c_{200c} \lesssim 3.0$ (average 2.7) in our mass and redshift range, which agrees with our constraints at the $\sim 2\sigma$ level only.

In the bottom panel of Fig. 17 we additionally show the stacked profile of the reduced cross-shear component of all clusters measured with respect to their X-ray centres (computed without rescaling). We find that it is consistent with zero, providing another consistency check for our analysis.

7.4 Calibration of the mass estimates with simulations and consistency checks in the data

We have adopted a simplistic model for the mass distribution in clusters, namely a spherical NFW halo with a known centre and a concentration fixed by a concentration–mass relation. However, effects such as choosing an improper cluster centre ('miscentring'), variations in cluster density profiles and noise bias in statistical estimators can introduce substantial biases in the mass constraints derived from fits of such a model to cluster weak lensing shear profiles (e.g. Becker & Kravtsov 2011; Gruen et al. 2015). To estimate and correct for these biases in our analysis we apply our measurement procedure to a large sets of simulated cluster weak lensing data based on the Millennium XXL simulation (Angulo et al. 2012) and the simulations created by Becker & Kravtsov (2011, henceforth BK11). The details of this analysis will be presented in Applegate et al. (in preparation). Here we only summarize the most important points relevant to this analysis.

7.4.1 Simulations

The two simulations considered for our calibration differ in the redshifts of the available snapshots, in the cluster mass range and the input cosmology. The difference in cosmology alters the concentration–mass relation in the simulation (e.g. Diemer & Kravtsov 2015), but this is small compared to the range of $c(M)$ relations we consider (Section 7.4.4). Likewise, the calibration does not depend on mass to a level that is important for this analysis when the full likelihood distribution of the mass constraints is used (see Applegate et al., in preparation). We find that the bias has some dependence on redshift and therefore interpolate between the two available snapshots that match our observations best. For the calibration of both M_{200c} and M_{500c} these are the $z = 0.5$ snapshot of the BK11 simulation and the $z = 1$ snapshot of the Millennium XXL simulation. Note that both simulations yield consistent bias calibrations at $z = 0.25$, where data are available from both simulations (see Applegate et al., in preparation).

For the BK11 $z = 0.5$ snapshot we include 788 haloes with $M_{500c} > 1.5 \times 10^{14}\,M_\odot\,h^{-1}$, providing a good match to the SPT cluster mass range. Since the sample provided to us by BK11 is selected in M_{500c}, we are only able to measure the bias in M_{200c} at $M_{500c} > 4 \times 10^{14}\,M_\odot\,h^{-1}$, above which we are still complete. For the MXXL $z = 1$ snapshot we include the 2100 most massive haloes, corresponding to $M_{200c} > 3.5 \times 10^{14}\,M_\odot\,h^{-1}$. For the calibration of weak lensing estimates for M_{500c} this sample is complete for $M_{500c} \gtrsim 3.2 \times 10^{14}\,M_\odot\,h^{-1}$, matching the mass range of the studied SPT clusters well (compare to Table 1).

The generation of simulated shear fields from the underlying N-body simulations is described in BK11. In short, all particles within $400\,h^{-1}$ Mpc along the line of sight to each cluster are projected on to a common plane to produce a κ map, from which a fast Fourier transform can compute the shear field on a regular grid. The procedure is similar for MXXL, except that particles within $200\,h^{-1}$ Mpc are used, and three orthogonal projection directions are employed.

We create mock observations matching each cluster in our observed sample. We first select a profile centre location by randomly choosing an offset from the true cluster centre, which is defined as the position of the most-bound particle in the simulation, according to different probability distributions reflecting our assumptions on the miscentring distributions of SZ and X-ray centres (Section 7.4.2). We then bin and azimuthally average the simulated reduced shear grid, matching the binning in the observed shear profile and add Gaussian random noise to each bin matching the observed noise levels. We fit the cluster masses from these simulated weak lensing data as done for the real clusters, calculating scans of χ^2 versus M_{meas}. To obtain a bias calibration for the scaling relation analysis (see Section 8) we model the ratio M_{meas}/M_{true}, where M_{true} denotes the corresponding halo mass, as a log-normal distribution. We associate the mean of the log-normal distribution as the inferred average bias and the width of the distribution as the intrinsic scatter from cluster triaxiality, substructure and line-of-sight projections. We fit the log-normal distribution to the population of clusters in each snapshot, marginalizing over the statistical uncertainty for each cluster (see Applegate et al., in preparation). While we perform the analysis in bins of true halo mass to check for mass-dependence of the bias, we instead only use one all-encompassing mass bin to determine the bias for this analysis.

We repeat the whole procedure for a number of miscentring distributions and $c(M)$ relations. We list individual bias numbers for each cluster for the X-ray and SZ miscentring distributions and the Diemer & Kravtsov (2015) $c(M)$ relation in Table 6, and sample-averaged values for a number of configurations in Table 7. We stress that the quoted bias numbers are adequate for quantitative analyses that take the full likelihood distribution of the mass constraints

[14] An alternative approach to constrain cluster concentrations from weak lensing data is to fit both mass and concentration simultaneously for each cluster. These individual constraints are however very weak due to shape noise, and they are strongly affected by large-scale structure projections (e.g. Hoekstra 2003).

Table 7. Mass recovery bias factors for the analysis taking the full likelihood distribution into account, averaged over all of our clusters, for different miscentring distributions and concentration–mass relations. The statistical uncertainty of the bias correction ranges from 1.5 per cent for our lower redshift clusters to 2.5 per cent for our highest redshift clusters.

Miscentring	$c(M)$ rel.	$\langle b_{200c}\rangle$	$\langle b_{500c}\rangle$
None	Diemer+15	0.95	0.96
X-ray-hydro	Diemer+15	0.87	0.87
SZ-hydro	Diemer+15	0.81	0.81
SZ-hydro	$c_{200c} = 4$	0.79	0.81
SZ-hydro	$c_{200c} = 3$	0.89	0.86
SZ-hydro	$c_{200c} = 5$	0.73	0.77

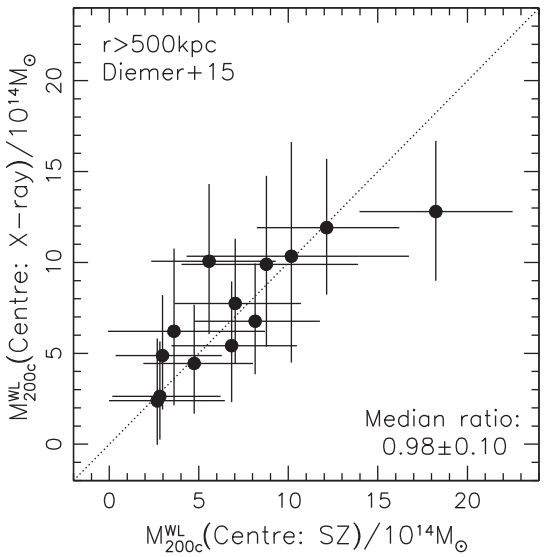

Figure 18. Comparison of the bias-corrected weak lensing mass estimates using the X-ray versus the SZ centres. The high-mass outlier is the merger SPT-CL *J*0102−4915, for which the location of the SZ peak is closer to the centre between the two peaks of the mass reconstruction (see Fig. G2), resulting in a higher mass estimate.

into account, as done in our scaling relation analysis presented in Section 8. We correct the mass estimates as

$$M_x^{\rm WL} = \frac{M_x^{\rm biased}}{b_x}. \quad (33)$$

As an approximation we also apply these bias correction factors to the maximum likelihood values and confidence intervals indicated in Figs 18–20 in the following sections. However, note that the bias factors may differ at some level for the maximum likelihood estimates and the fits that use the full likelihood distribution due to differences in the impact of noise bias. We plan to investigate this issue further in Applegate et al. (in preparation).

7.4.2 *Miscentring distributions*

For the SPT clusters we have proxies for the cluster centres, where we in particular use the X-ray centroids and SZ peaks for the mass analysis. These need to be related to the cluster centres defined by halo finding algorithms used to predict the cluster mass function from simulations. These offsets will typically lower the measured shear from the expected NFW signal at small radii (e.g. Johnston et al. 2007; George et al. 2012). To mimic this effect in the BK11 and MXXL *N*-body simulations, where we have neither mock SZ nor mock X-ray observations, we employ offset distributions derived from the Magneticum Pathfinder Simulation (Dolag, Komatsu & Sunyaev 2016; see also Bocquet et al. 2016), which is a large volume, high-resolution cosmological hydrodynamical simulation. It includes simulated SZ and X-ray observations, where we make use of SPT mock catalogues (Saro et al. 2014; Gupta et al. 2017) that include the full SPT cluster detection procedure. We find that the most relevant parameter regarding the centring uncertainty when using the SZ centres is the smoothing scale θ_c used for the cluster detection (see Bleem et al. 2015). We therefore use the actual distribution of θ_c values for our clusters from Bleem et al. (2015) for the generation of the miscentring distribution.

7.4.3 *Impact and uncertainty of the miscentring correction*

Using the default $c(M)$ relation (Diemer & Kravtsov 2015) and comparing the analyses using the miscentring distributions from the hydrodynamical simulation to the case without miscentring, we estimate that miscentring on average introduces a moderate mass bias of 8–9 per cent when using the X-ray centres, and a more substantial bias of 14–15 per cent using the SZ centres (see Table 7). The SZ measurements less accurately determine the cluster centre, which on-average increases the bias correction. This result is consistent with the smaller average offsets from the mass peaks found for the X-ray centres (Section 7.1).

As a consistency check for the miscentring correction we compare the bias-corrected mass estimates using the X-ray and SZ centres in Fig. 18. Their median ratio 0.98 ± 0.10, with an uncertainty estimated by bootstrapping the clusters, is consistent with unity as expected in the case of accurate bias correction. We, however, note that the small sample size leads to a significant uncertainty of this median ratio, making it not a very stringent test for the accuracy of the bias correction.

The accurate correction for mass modelling biases such as the one introduced by miscentring is an active field of research (e.g. LSST Dark Energy Science Collaboration 2012). Our analysis using a miscentring distribution based on a hydrodynamical simulation is a step forward in this respect, but we acknowledge that it is still simplistic. In particular, it ignores that positional offsets are not always in a random direction. This is prominently demonstrated by the merger SPT-CL *J*0102−4915, for which the location of the SZ peak is closer to the centre between the two peaks of the mass reconstruction (see Fig. G2), leading to an increased mass estimate (compare Fig. 18). Due to this simplification in our current analysis, we conservatively assign a large uncertainty for the miscentring correction, which amounts to 50 per cent of the correction, corresponding to a 4 per cent uncertainty in mass when using the X-ray centres and 7 per cent when using the SZ centres. Future analyses can reduce this uncertainty by simulating all observables including the weak lensing data from the same hydrodynamical simulation (see Section 9.5).

7.4.4 *Uncertainties in the concentration–mass relation*

For the case of SZ miscentring Table 7 lists average bias numbers for the $c(M)$ relation from Diemer & Kravtsov (2015), as well as fixed concentrations $c_{200c} \in \{3, 4, 5\}$. Our bias correction procedure effectively maps the $c(M)$ relation used for the fit to the observed $c(M)$ relations in the simulations that are used for the bias correction (BK11, Millennium XXL). The remaining question is how well

the $c(M)$ relations in these simulations resemble the true average $c(M)$ relation in the Universe, especially regarding the impact of baryons. Duffy et al. (2010) show that the impact of baryon physics appears to have only a relatively minor ($\lesssim$10 per cent) influence on the concentrations of very massive clusters. De Boni et al. (2013) find similar numbers at low redshifts (for complete halo samples), and slightly stronger effects at $z = 1$ ($\sim$15–20 per cent). Interpolating between the $\langle b_{500c} \rangle$ values in Table 7 we estimate that a 10–20 per cent uncertainty on the concentration around $c_{200c} = 4$ leads to a $\sim$2–4 per cent systematic uncertainty for the constraints on M_{500c}, where we conservatively adopt the larger number in our systematic error budget (see Section 7.5).

De Boni et al. (2013) note that differences in the definition of the concentration can lead to shifts in the values measured from N-body simulations of up to 20 per cent. This is not a concern for our analysis, as we directly estimate the calibration from the simulated weak lensing data, and therefore do not rely on concentration measurements in the simulations.

7.5 Statistical precision versus systematic uncertainty

We summarize the identified sources of systematic uncertainty for our study in Table 8, pointing to their corresponding sections, and listing their associated relative uncertainties in the measured weak lensing signal and mass constraints. Combining all systematic error contributions in quadrature, we estimate an overall systematic mass uncertainty of 9 per cent (11 per cent) for the analysis using the X-ray (SZ) centres. This can be compared to the combined statistical mass signal-to-noise ratio of the sample, which we approximate as

$$(\mathrm{S/N})^{\mathrm{sample}}_{\mathrm{mass}} = \sqrt{\sum_{\mathrm{clusters}} \left(M_{500c,i}/\Delta M^{\mathrm{stat.}}_{500c,i}\right)^2} \simeq 7.3, \qquad (34)$$

which corresponds to a $\sim$14 per cent precision, ignoring the impact of intrinsic scatter, e.g. from cluster triaxiality. Accordingly, our total uncertainty is dominated by statistical measurement noise and not systematic uncertainties.

For the analysis of larger future data sets with improved statistical precision it will be important to further reduce systematic uncertainties. When discussing the individual sources of systematic uncertainty we have already suggested strategies how their influence can be reduced in the future. The largest contributions to the systematic error budget currently come from the shear calibration, miscentring corrections, and uncertainties in the $c(M)$ relation. All of these can be reduced with better simulations. For the latter two issues the weak lensing data can themselves provide information that help to reduce these uncertainties (see also Section 9.5). As a rough guess we expect that it should be possible to cut the systematic uncertainties associated with the mass modelling by half in the coming years with moderate effort (compare Table 8), and note that some improved shape measurement techniques have already reached significantly higher accuracy (e.g. Bernstein et al. 2016; Fenech Conti et al. 2017). We further discuss the strategies to reduce systematic uncertainties in Section 9.

8 CONSTRAINTS ON THE M–T_{X} SCALING RELATION

In the self-similar model (e.g. Kaiser & Silk 1986) galaxy clusters form through the gravitational collapse of the most overdense regions in the early Universe. In this model the cluster baryons are heated through gravitational processes only, leading to predictions for cluster scaling relations. Deviations from self-similarity, e.g. regarding the slope of the X-ray luminosity–temperature relation (e.g. Arnaud & Evrard 1999), suggest that non-gravitational effects, such as heating by active galactic nuclei or radiative cooling, provide non-negligible contributions to the energy budget of clusters. However, the redshift evolution of cluster X-ray observables

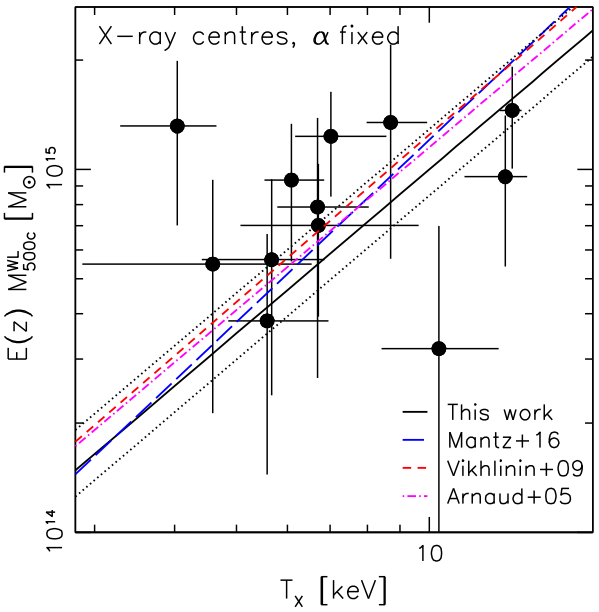

Figure 19. Core-excised X-ray temperatures measured in the range (0.15–1) $\times$ r_{500c} based on *Chandra* data versus $E(z)M^{\mathrm{WL}}_{500c}$ from the weak lensing analysis using the X-ray centroids and assuming the $c(M)$ relation from Diemer & Kravtsov (2015). The solid black line shows our best-fitting estimate of the scaling relation when assuming a fixed slope $\alpha = 3/2$. The dotted lines correspond to normalizations that are lower or higher by 1σ, combining the statistical and systematic uncertainties of our constraints. The dashed and dashed-dotted lines indicate best-fitting estimates derived by Arnaud et al. (2005), Vikhlinin et al. (2009a) and Mantz et al. (2016).

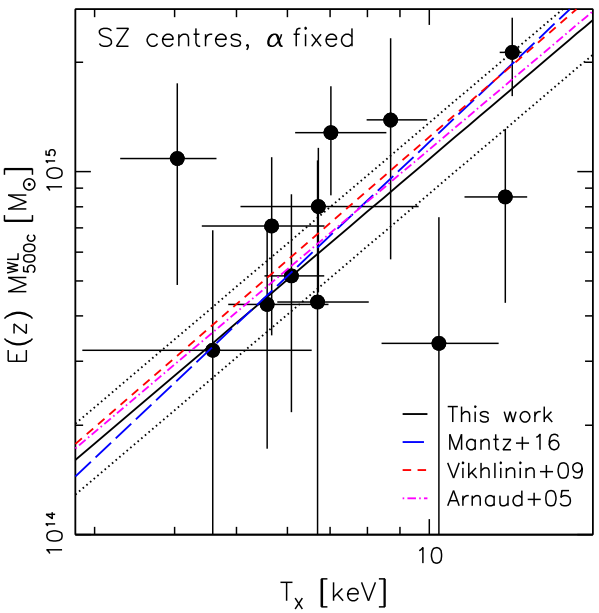

Figure 20. As Fig. 19, but employing the weak lensing results for the SZ centres.

Table 8. Systematic error budget for our current study and our expectation for what can be achieved in similar studies in the near future with moderate analysis improvements.

Source	Current		Near future		Sect./ App.	Improve via
	Rel. error signal	Rel. error M_{500c}	Rel. error signal	Rel. error M_{500c}		
Shape measurements:						
Shear calibration	4 per cent	6 per cent	1 per cent	1.5 per cent	5	Image simulations
Redshift distribution:						
$\langle\beta\rangle$ sys. photo-z	2.2 per cent	3.3 per cent	1.5 per cent	2.2 per cent	6.3.2	Improved priors + $p(z)$
$\langle\beta\rangle$ cosmic variance	1 per cent	1.5 per cent	1 per cent	1.5 per cent	6.6	More reference fields
$\langle\beta\rangle$ deblending	0.5 per cent	0.8 per cent	0 per cent	0 per cent	B	F606W-detected photo-zs
$\langle\beta\rangle$ LCBG contamination	0.9 per cent	1.4 per cent	0.5 per cent	0.8 per cent	F	Apply model
Mass model:						
Miscentring for X-ray (SZ) centres		4 per cent (7 per cent)		2 per cent (3.5 per cent)	7.4.2	Hydro sims, weak lensing
$c(M)$ relation		4 per cent		2 per cent	7.4.4	Hydro sims, weak lensing
Total for X-ray (SZ) centres**:**		9.2 per cent (10.8 per cent)		4.2 per cent (5.1 per cent)		

appears to be consistent with self-similar predictions (e.g. Maughan et al. 2006), suggesting that non-gravitational effects have a similar impact at low and high redshifts. If this 'weak self-similarity' (e.g. Bower 1997) also applies to cluster masses, we expect a scaling between temperature and mass in the form

$$M_x E(z) \propto T^{\alpha}, \tag{35}$$

(e.g. Mathiesen & Evrard 2001; Böhringer, Dolag & Chon 2012), where

$$E(z) = \frac{H(z)}{H_0} = \sqrt{\Omega_{\rm m}(1+z)^3 + \Omega_{\Lambda}} \tag{36}$$

indicates the redshift dependence of the *Hubble* parameter, here assuming a flat ΛCDM cosmology, and $\alpha = 3/2$ corresponds to the self-similar prediction for the slope of the relation.

The main constraints on cluster scaling relations from our sample will be presented in a forthcoming paper (Dietrich et al. 2017) that combines our measurements with a complementary sample of clusters at lower redshifts with Magellan/Megacam observations and accounts for the SPT selection function, which is especially important when calibrating SZ scaling relations. However, here we already combine our measurements with core-excised *Chandra* X-ray temperature estimates $T_{\rm X}$ that are available for 12 clusters in our sample. Details of the specific measurements are provided in McDonald et al. (2013), with the analysis pipeline adapted based on Vikhlinin et al. (2006). In short, *Chandra* ACIS-I data were reduced using CIAO v4.7 and CALDB v4.7.1. All exposures were initially filtered for flares, before applying the latest calibrations and determining the appropriate epoch-based blank-sky background. Point sources were identified via an automated wavelet decomposition technique (Vikhlinin et al. 1998) and masked. Spectra were extracted in a core-excised region from (0.15–1) × r_{500c} (McDonald et al. 2013) and fit over 0.5–10.0 keV using a combination of an absorbed, optically thin plasma (PHABS × APEC), an absorbed hard background component (PHABS × BREMSS), and a soft background (APEC), see McDonald et al. (2013) for details.

Figs 19 and 20 show the bias-corrected $M^{\rm WL}_{500c}E(z)$ using the Diemer & Kravtsov (2015) $c(M)$ relation as a function of the core-excised $T_{\rm X}$ estimates (Table 9) for the analyses centring on the X-ray centroids or SZ peaks, respectively. For comparison we show best-fitting estimates for the scaling relation derived by Arnaud, Pointecouteau & Pratt (2005, based on their $T_{\rm X} > 3.5$ keV sample), Vikhlinin et al. (2009a) and Mantz et al. (2016) using clusters at lower and intermediate redshifts ($z \lesssim 0.6$).

Table 9. Core-excised *Chandra* X-ray temperatures used for our constraints on the M–$T_{\rm X}$ scaling relation.

Cluster	$T_{\rm X}$ (keV)
SPT-CL J0000−5748	$6.7^{+2.9}_{-1.6}$
SPT-CL J0102−4915	$13.5^{+0.5}_{-0.6}$
SPT-CL J0533−5005	$4.6^{+2.0}_{-1.7}$
SPT-CL J0546−5345	$6.7^{+1.4}_{-0.9}$
SPT-CL J0559−5249	$6.1^{+0.8}_{-0.6}$
SPT-CL J0615−5746	$13.1^{+1.1}_{-1.8}$
SPT-CL J2106−5844	$8.7^{+1.2}_{-0.7}$
SPT-CL J2331−5051	$5.6^{+1.4}_{-0.7}$
SPT-CL J2337−5942	$7.0^{+1.6}_{-0.9}$
SPT-CL J2341−5119	$10.4^{+2.5}_{-1.9}$
SPT-CL J2342−5411	$4.0^{+0.6}_{-0.8}$
SPT-CL J2359−5009	$5.7^{+1.2}_{-1.3}$

To obtain quantitative constraints on the scaling relation, we assume the functional form

$$\ln\left(E(z)M_{500c}/10^{14}\,{\rm M}_{\odot}\right) = A + \alpha\left[\ln\left(kT/7.2{\rm keV}\right)\right], \tag{37}$$

where the temperature pivot point roughly corresponds to the mean temperature of the sample. Our fitting method is based on the approach of Kelly (2007), which incorporates measurement errors in the x- and y- coordinates and has been extended to include log-normal intrinsic scatter. The method has been generalized to use the exact likelihood from the lensing analysis, and a two-piece normal approximation to the X-ray likelihood (Applegate et al. 2016). For this analysis we use the lensing likelihood based on the dominant shape noise only and absorb the minor contributions from

large-scale structure projections and line-of-sight variations in the redshift distribution (see Section 7.2) in the intrinsic scatter $\sigma_{\rm M}$.

We fix the slope of the scaling relation to the self-similar prediction ($\alpha = 3/2$) for the current analysis, given the limited sample size and mass range. We then obtain constraints $(A, \sigma_{\rm M}) = (1.81^{+0.24}_{-0.14}, 0.05^{+0.32}_{-0.05})$ for our default analysis using the X-ray centres. When alternatively using the SZ peaks as centre for the weak lensing analysis we obtain consistent results $(A, \sigma_{\rm M}) = (1.89^{+0.20}_{-0.19}, 0.31^{+0.04}_{-0.31})$. In addition to these statistical uncertainties there is a 9 per cent (11 per cent) systematic uncertainty for the analysis using the X-ray (SZ) centres, directly propagating into the normalization of the scaling relation (see Section 7.5). The obtained constraints are consistent with the aforementioned results from lower redshift samples when assuming self-similar redshift evolution within 1σ (see Figs 19 and 20).

Jee et al. (2011) present an *HST* weak lensing analysis for 27 galaxy clusters at $0.83 \leq z \leq 1.46$, using a heterogeneous sample that includes optically, NIR- and X-ray-selected clusters. Their analysis suggests a possible evolution in the M_{2500c}–$T_{\rm X}$ scaling relation in comparison to self-similar extrapolations from lower redshifts. For example, at $T_{\rm X} = 5$ keV their estimated scaling relation has a lower amplitude by 27 ± 7 per cent [statistical uncertainty from Jee et al. (2011) only] compared to the best-fitting relation from Arnaud et al. (2005). We do not find significant indications for a similar evolution for the M_{500c}–$T_{\rm X}$ scaling relation, but note that our statistical uncertainties are significantly larger given our smaller sample size and more conservative radial fit range. There are various additional differences in the analyses, such as different samples for the calibration of the source redshift distribution, our more conservative removal of cluster galaxies, and our calibration of modelling biases on simulations, making the direct comparison difficult. Importantly, both studies use different overdensities for the scaling relation constraints.[15] Furthermore, Jee et al. (2011) use X-ray temperature estimates from the literature that typically do not exclude the core regions. Including the cores should, on average, reduce the temperatures in the presence of cool-core clusters. This would, however, aggravate the tension between the Jee et al. (2011) results and the self-similar extrapolations from lower redshift samples.

9 DISCUSSION

In our analysis we have introduced a number of new aspects and systematic investigations for weak lensing studies of high-redshift clusters. Here we discuss their relevance also in the context of future weak lensing programmes. Our study using *HST* and VLT data provides a demonstration for future weak lensing science investigations that combine deep high-resolution space-based shape measurements, e.g. from *Euclid* (Laureijs et al. 2011) or WFIRST (Spergel et al. 2015), with deep photometry, e.g. from LSST (LSST Science Collaboration et al. 2009).

9.1 The benefits and challenges of using faint blue galaxies for weak lensing

For deep weak lensing surveys conducting shape measurements at optical wavelengths the majority of the high-redshift ($z \sim 1.5$–3) sources are blue star-forming galaxies observed at rest-frame UV wavelengths with blue observed optical colours (see the top left panel of Fig. 5). These galaxies are useful as the source sample in weak lensing studies of high-redshift clusters both because of their high source density and high geometric lensing efficiency, but also because they can be readily distinguished from both blue and red cluster galaxies using optical colours (see Section 6.2). This enables a nearly complete removal of cluster galaxies from the weak lensing source sample, which is important both in order to minimize modelling uncertainties regarding cluster member contamination (Appendix E), and to ensure that intrinsic alignments of galaxies within the targeted clusters cannot bias mass constraints (but note that this appears to be a negligible effect at the precision of current samples, see Sifón et al. 2015).

To exploit these benefits, a number of challenges need to be overcome. Here we first stress that high signal-to-noise optical photometry is needed to robustly select these galaxies in colour space. In the case of our study a well-matched colour selection was possible in areas covered by ACS in both *F*606*W* and *F*814*W*. However, outside the *F*814*W* footprint we had to rely on the combination of *F*606*W* ACS imaging and VLT $I_{\rm FORS2}$ images, which, despite a good VLT integration time $\langle t_{\rm exp} \rangle = 2.4$ ks and $\lesssim$0.8 arcsec seeing, delivered a density of usable sources that is only 32 per cent of the density from the ACS-only $V_{606} - I_{814}$ selection (Section 6.8.1). This highlights that future weak lensing programmes and surveys should carefully tune the relative depth of their bands (regarding both red and blue filters) to maximize the science output of their data.

While our analysis is based on simple colour cuts due to the limited data available in our cluster fields, we expect that similar conclusions apply for surveys that aim at computing individual photometric redshifts for the weak lensing source galaxies. Photometric redshift selections correspond to higher dimensional cuts in colour–colour space. However, depending on the survey characteristics, the large population of blue high-z galaxies may only be detected in a few of the bluer optical pass bands, effectively reducing photo-z cuts to a selection in a relatively small colour–colour space. As a result, individual photometric redshift estimates for faint blue galaxies have typically large uncertainties unless deep photometry is available over a very broad wavelength range (in particular including deep u-band and NIR observations). For cluster weak lensing studies noise in individual photometric redshifts is not a problem as long as cluster galaxies can be removed robustly and the overall source redshift distribution can be modelled accurately.

9.2 Robust estimates of the source redshift distribution

We employ a statistically consistent selection of source galaxies matched in filter, magnitude, colour and shape properties in our cluster fields and observations of the CANDELS fields. This allows us to estimate the average source redshift distribution and its statistical variation between lines of sights using the CANDELS data and apply this information for the cluster weak lensing analyses. At depths similar to our data, the CANDELS fields are currently among the extragalactic fields that are best studied both photometrically and spectroscopically. We have shown that they cover enough sky area to reduce the cosmic variance contribution to the uncertainty on the mean lensing efficiency at our cluster redshifts to the $\sim$1 per cent level (Section 6.6), which is much smaller than current statistical weak lensing uncertainties. Therefore, we expect that the CANDELS fields will remain to be an important calibration sample for estimates of the source redshift distribution in deep weak lensing data in the near future.

[15] We do not report M_{2500c} masses as these are not available in the BK11 simulation, preventing us to compute accurate bias corrections for this overdensity.

As revealed by our comparison to HUDF data (Section 6.3.1) and confirmed via spatial cross-correlations with spectroscopic/grism redshifts (Appendix C), the 3D-HST CANDELS photo-zs suffer from catastrophic redshift outliers (primarily galaxies at $2 \lesssim z \lesssim 3$ that are assigned a low photometric redshift $z_{\rm p} < 0.3$) and redshift focusing effects at $z_{\rm p} \simeq 1.5$. Together these would on average bias our mass estimates high by 12 per cent if not accounted for. For our current study we have implemented an empirical correction for these systematics effects. We plan to investigate this issue and its causes in detail in a future paper (Raihan et al., in preparation). Given the high photometric quality, depth and broad wavelength coverage of the CANDELS data, we speculate that some other current photometric redshift data sets might suffer from similar effects. This is supported by the weak lensing analyses of S10 and Heymans et al. (2012) as discussed in Section 6.3. We therefore expect that also other weak lensing programmes will have to implement similar correction schemes or improved photometric redshift algorithms, and apply these either to deep field data in case of colour cut analyses, or their survey data in case of individual photo-z estimates. Surveys that obtain individual photo-zs can also attempt to identify and remove galaxies in problematic $z_{\rm p}$ ranges at the cost of reduced sensitivity. We stress that the use of the average redshift posterior probability distribution instead of the peak photometric redshift estimates is not sufficient to cure the identified issue for the 3D-HST photo-zs (Section 6.3.4).

One route to calibrate photo-zs is via very deep spectroscopy for representative galaxy samples. At present, such spectroscopic samples are very incomplete at the depth of our analysis, which is why we resorted to the comparison of the CANDELS photo-zs to photometric redshifts for the HUDF (Rafelski et al. 2015), which are based on deeper data and a broader wavelength coverage. We find that this is a viable approach at the precision of current and near-term high-z cluster samples with weak lensing measurements, but it is likely not of sufficient accuracy for the calibration of very large future data sets. To prepare for the analyses of such data sets it is vital and timely to obtain larger spectroscopic calibration samples, including both highly complete deep samples for direct calibration, but also very large, potentially shallower and less complete samples (Newman et al. 2015). The later can be used to infer information on the redshift distribution via spatial cross-correlations (e.g. Newman 2008; Matthews & Newman 2010; Schmidt et al. 2013; Rahman et al. 2015, 2016), for which we provide one of the first practical applications in the context of weak lensing measurements (see Appendix C and Hildebrandt et al. 2017).

As an important ingredient for our modelling of the redshift distribution we carefully matched the selection criteria and noise properties between our cluster field data and the CANDELS data to ensure that consistent galaxy populations are selected between both data sets (see Section 6.4 and Appendix D3). For the colours obtained from the combination of ACS $F606W$ and VLT $I_{\rm FORS2}$ data we empirically estimated the net scatter distribution by comparing to the colours estimated in the inner cluster regions from ACS $F606W$ and $F814W$ data. We note that systematic effects such as residuals from the PSF homogenization can add scatter which may well deviate from Poisson noise distributions that are often assumed, e.g. in photometric redshift codes. As we empirically sample from the actual scatter distribution such effects are automatically accounted for in our analysis. For future surveys that vary in data quality we recommend to obtain repeated imaging observations of spectroscopic reference fields that span the full range of varying observing conditions, in order to generate similar empirical models for the impact of the actual noise properties.

9.3 Accounting for magnification

The impact of weak lensing magnification on the source redshift distribution has typically been ignored in past weak lensing studies. Our investigation of this effect in Section 6.7 indicates that the net effect is small for our study given the depth of our data. However, shallower programmes such as Dark Energy Survey (DES; The Dark Energy Survey Collaboration 2005) or KiDS (Kuijken et al. 2015), which aim to calibrate high-z cluster masses by combining measurements from a large number of clusters, will need to carefully account for the resulting boost in the average lensing efficiency $\langle\beta\rangle$. For example, Fig. 11 illustrates that the impact of magnification on the source redshift distribution has a larger impact on the reduced shear profile at brighter magnitudes than the typically applied correction for the finite width of the source redshift distribution.

We point out that knowledge of the redshift distribution is needed at fainter magnitudes than the targeted depth limit of a survey in order to be able to compute the actual correction for the impact of magnification (Section 6.7). Accordingly, it is necessary to obtain spectroscopic redshift samples for photo-z calibration to greater depth than the targeted survey depth. The difference in depth depends on the maximum magnification that is considered, and therefore the magnitude limit, the cluster redshift and mass, as well as the considered fit range.

We also note that it is important to take magnification into account when using the source density and the density profiles as validation tests for the cluster member removal (see Section 6.8). Programmes with ground-based resolution will also need to account for the change in source sizes due to magnification as a function of redshift, cluster-centric distance and mass, as shape cuts could otherwise introduce redshift- and mass-dependent selection biases.

9.4 Shape measurement biases

Currently the shear calibration uncertainty constitutes the largest individual contribution to the systematic error budget of our study (4 per cent for the shear calibration corresponding to a 6 per cent mass uncertainty). This is due to the fact that we base the calibration on simulations from the STEP project (Section 5.1) which lack faint galaxies that influence the bias calibration (Hoekstra et al. 2015) and do not probe shears as high as those used in our analysis. However, this source of systematics can easily be reduced through image simulations that resemble real galaxy populations and cluster-regime shears more accurately, and which can be generated with recent tools such as GALSIM (Rowe et al. 2015). We therefore expect that shear measurement biases in cluster weak lensing studies will soon be reduced to the levels reached in cosmic shear measurements (e.g. Fenech Conti et al. 2017). Also see Hoekstra et al. (2017), whose results suggest that the impact of the higher density of sources in cluster regions on shape measurement biases should be negligible for current data.

In addition, additive shape measurement biases can be relevant for cluster weak lensing in particular for pointed follow-up programmes where the clusters are always centred at similar detector positions. An example for such a potential source of bias can be CTI residuals. However, through a new null test we have shown that our data show no significant CTI-like residuals within the current statistical uncertainty (Section 5.2).

9.5 Accounting for biases in the mass modelling

We have calibrated our mass estimates using reduced shear profile fits to simulated cluster weak lensing data from *N*-body simulations (see Section 7.4). One important source for bias is miscentring of the reduced shear profile. As we do not know the location of the centre of the 3D cluster potential we have to rely on observable proxies for the cluster centre, leading to a suppression of the expected reduced shear signal at small radii. Based on the work from Dietrich et al. (2012) we expect that the peaks in the reconstructed weak lensing mass maps of the clusters (see Section 7.1) should provide a tight tracer for the centre of the 3D cluster potential, but we do not use these centres for our mass constraints in our current analysis as they are expected to yield masses that are biased high. By studying the offset distributions between the mass peaks and the other proxies for the cluster centre we find that the X-ray centroids provide the smallest average offsets, closely followed by the SZ peak locations. Hence, they also provide good proxies for the cluster centre, which is why we employ them as centres for our mass constraints. To account for the expected remaining bias caused by miscentring, we randomly misplace the centre in the simulated weak lensing data based on offset distributions measured between the 3D cluster centre and the SZ peak location or X-ray centroid in hydrodynamical simulations (see Section 7.4.2). Future studies could further advance this approach by simulating all observables including SZ, X-ray and weak lensing data from the same hydrodynamical simulation, in order to also account for possible covariances between these observables. Our analysis of the prominent merger SPT-CL J0102−4915 demonstrates that such covariances exist, as both the X-ray centroid and SZ peak are located between the two peaks of the mass reconstruction (see Fig. G2). Hence, the misplacement is not in a random direction. To validate the accuracy of the employed simulations, the measured offset distributions between the mass peaks and the different proxies for the centre can be compared between the real data and the simulations. This approach could be further expanded by explicitly accounting for the miscentring in the fitted reduced shear profile model (e.g. Johnston et al. 2007; George et al. 2012; also see Köhlinger, Hoekstra & Eriksen 2015 for the impact of miscentring in stacked Stage IV analyses).

A further uncertainty for the mass constraints arises from uncertainties in the assumed $c(M)$ relation. The applied calibration procedure essentially maps the measurements on to the $c(M)$ relation of the simulation. Remaining uncertainties reflect our ability to simulate the true $c(M)$ relation of the Universe, especially with respect to the impact of baryons. These uncertainties are expected to shrink with further advances in simulations, in particular thanks to the recent advent of large hydrodynamical simulations (e.g. Dolag et al. 2016). In addition, the weak lensing measurements themselves can be used to test if the inferred reduced shear profiles are consistent with the simulation-based priors on the $c(M)$ relation, in particular if information from the inner reduced shear profiles is incorporated. Using the X-ray centroids our analysis yields a best-fitting fixed concentration for the sample of $c_{200c} = 5.6^{+3.7}_{-1.8}$ when including scales >300 kpc (Section 7.3). This is fully consistent with recent results for the $c(M)$ relation from simulations (e.g. Diemer & Kravtsov 2015), but higher than earlier results from Duffy et al. (2008), which, however, are based on a *WMAP5* cosmology (Komatsu et al. 2009) with lower Ω_m and σ_8, reducing the resulting concentrations. We note that future studies that aim to obtain tighter constraints on the concentration will have to account for the impact of miscentring and stronger shears in the inner cluster regions, which we could ignore for this part of our analysis given the statistical uncertainties.

10 CONCLUSIONS

We have presented a weak gravitational lensing analysis of 13 high-redshift clusters from the SPT-SZ Survey, based on shape measurements in high resolution *HST*/ACS data and colour measurements that also incorporate VLT/FORS2 imaging. We have introduced new methods for the weak lensing analysis of high redshift clusters and carefully investigated the impact of systematic uncertainties as discussed in Section 9 in the context of future programmes. In particular, we select blue galaxies in $V_{606} - I_{814}$ colour to achieve a nearly complete removal of cluster galaxies, while selecting most of the relevant source galaxies at $1.4 \lesssim z \lesssim 3$ (see Section 6.2). Carefully matching our selection criteria we estimate the source redshift distribution using data from CANDELS, where we apply a statistical correction for photometric redshift outliers. This correction is derived from the comparison to deep spectroscopic and photometric data from the HUDF (see Section 6.3), and checked using spatial cross-correlations (see Appendix C). We account for the impact of lensing magnification on the source redshift distribution, which we find is especially important for shallower surveys (see Section 6.7). We also introduce a new test for residual contamination of galaxy shape estimates from CTI, which is in particular applicable for pointed cluster follow-up observations (see Section 5.2). Finally, we account for biases in the mass modelling through simulations (see Section 7.4).

At present, our weak lensing mass constraints are limited by statistical uncertainties given the small cluster sample and the limited depth of the data for the colour selection in the cluster outskirts. For the current study the total systematic uncertainty on the cluster mass scale at high-z is at the ∼9 per cent level, where the largest contributions come from the shear calibration and mass modelling. As discussed in Section 7.5 we have identified strategies how this can be reduced to the ∼4 per cent level in the near future based on existing calibration data and improved simulations. This is particularly relevant for near-term studies using larger *HST* data sets.

We have used our measurements to derive updated constraints on the M_{500c}–T_X scaling relation for massive high-z clusters in combination with *Chandra* observations. Compared to scaling relations calibrated at lower redshifts we find no indication for a significant deviation from self-similar redshift evolution at our current ∼20 per cent precision (see Section 8). Our measurements will additionally be used in companion papers to derive updated constraints on additional mass–observable scaling relations, where we also incorporate weak lensing measurements at lower redshifts from Magellan/Megacam (Dietrich et al. 2017) and the DES (The Dark Energy Survey Collaboration 2005; Stern et al., in preparation), and to derive improved cosmological constraints from the SPT-SZ cluster sample.

We investigate the offset distributions between different proxies for the cluster centre and the weak lensing mass reconstruction, where we find that the X-ray centres provide the smallest average offsets (see Section 7.1). Our analysis constrains the average concentration of the cluster sample to $c_{200c} = 5.6^{+3.7}_{-1.8}$ (Section 7.3) when using the X-ray centres and including information from smaller scales (300 kpc $< r <$ 500 kpc), which are excluded for the conservative mass constraints.

With the advent of the next generation of deep cluster surveys such as SPT-3G (Benson et al. 2014), the Dark Energy Survey (DES The Dark Energy Survey Collaboration 2005), Hyper-Suprimecam (HSC Miyazaki et al. 2012), eROSITA (Merloni et al. 2012) and Advanced ACTPol (Henderson et al. 2016) it will be vital to further tighten the weak lensing calibration of cluster masses in order to exploit these surveys for constraints on cosmology and cluster astrophysics. At low and intermediate redshifts, weak lensing surveys such as DES, HSC and KiDS are expected to soon calibrate cluster masses at the few per cent level, especially if large numbers of clusters can be reliably selected down to lower masses and if their weak lensing signatures are combined statistically (e.g. Rozo, Wu & Schmidt 2011). Such surveys will also provide some statistical weak lensing constraints for clusters out to $z \sim 1$ (e.g. van Uitert et al. 2016), but it still needs to be demonstrated how reliably such measurements can be conducted from the ground as most of the distant background galaxies are poorly resolved. At such cluster redshifts *HST* is currently unique with its capabilities to measure robust individual cluster masses with good signal-to-noise ratio. Clusters at high redshifts and high masses are very rare. As a result, stacking analyses of shallower wide-area surveys cannot compete in terms of precision for their mass calibration with a large *HST* programme that obtains pointed follow-up observations for all of them. Our current study is an important pathfinder towards such a program. For comparison, stacked analyses tend to be more powerful for lower mass clusters, which are too numerous to be followed up individually. The combination of deep pointed follow-up for high-mass clusters and stacked shallower measurements for lower mass clusters is therefore particularly powerful for obtaining constraints on the slope of mass–observable scaling relations. In addition, good signal-to-noise ratios for individual clusters, as provided by deep pointed follow-up, are needed for constraints on intrinsic scatter.

In the 2020s weak lensing Stage IV dark energy experiments such as *Euclid* (Laureijs et al. 2011), LSST (LSST Science Collaboration et al. 2009) and WFIRST (Spergel et al. 2015) are expected to provide a precise calibration of cluster masses over a wide range in redshift (for a forecast for *Euclid* see Köhlinger et al. 2015). To reach their weak lensing science goals they will require highly accurate calibrations for the redshift distribution and shear estimation. Further efforts will be needed to fully exploit these calibrations and weak lensing data sets for cluster mass estimation. For example, the shear calibration needs to be extended towards stronger shear, and magnification has to be taken into account when estimating the source redshift distribution (Section 9.3). We also stress that it will be vital to pair such observational studies with analyses of large sets of hydrodynamical simulations, in order to accurately calibrate the weak lensing mass estimates and account for covariances with other observables (see Section 9.5).

LSST and *Euclid* will still have significantly lower densities of high-redshift background source galaxies compared to *HST* observations. In order to extend the mass calibration for massive clusters out to very high redshifts ($z \gtrsim 1.3$), large pointed *HST* and subsequently JWST programmes may therefore remain the most effective approach until similarly deep data become available from WFIRST.

ACKNOWLEDGEMENTS

This work is based on observations made with the NASA/ESA *HST*, using imaging data from the SPT follow-up GO programmes 12246 (PI: C. Stubbs) and 12477 (PI: F. W. High), as well as archival data from GO programmes 9425, 9500, 9583, 10134, 12064, 12440 and 12757, obtained via the data archive at the Space Telescope Science Institute, and catalogues based on observations taken by the 3D-HST Treasury Program (GO 12177 and 12328) and the UVUDF Project (GO 12534, also based on data from GO programmes 9978, 10086, 11563, 12498). STScI is operated by the Association of Universities for Research in Astronomy, Inc. under NASA contract NAS 5-26555. It is also based on observations made with ESO Telescopes at the La Silla Paranal Observatory under programmes 086.A-0741, 088.A-0796, 088.A-0889, 089.A-0824. The scientific results reported in this article are based in part on observations made by the *Chandra* X-ray Observatory (ObsIDs 9332, 9333, 9334, 9335, 9336, 9345, 10851, 10864, 11738, 11739, 11741, 11742, 11748, 11799, 11859, 11864, 11870, 11997, 12001, 12002, 12014, 12091, 12180, 12189, 12258, 12264, 13116, 13117, 14017, 14018, 14022, 14023, 14349, 14350, 14351, 14437, 15572, 15574, 15579, 15582, 15588, 15589, 18241).

It is a pleasure to thank Gabriel Brammer, Pieter van Dokkum, Mattia Fumagalli, Ivelina Momcheva, Rosalind Skelton and the 3D-HST team for helpful discussions and for making their photometric and grism redshift catalogues available to us prior to public release. We thank Matthew Becker and Andrey Kravtsov for making results from their *N*-body simulation (Becker & Kravtsov 2011) available to us, Raul Angulo for providing data from the Millennium XXL Simulation, and Klaus Dolag for providing access to data from the Magneticum Pathfinder Simulation.

TS, DA and SFR acknowledge support from the German Federal Ministry of Economics and Technology (BMWi) provided through DLR under projects 50 OR 1210, 50 OR 1308, 50 OR 1407 and 50 OR 1610. TS also acknowledges support from NSF through grant AST-0444059-001 and SAO through grant GO0-11147A. This work was supported in part by the Kavli Institute for Cosmological Physics at the University of Chicago through grant NSF PHY-1125897 and an endowment from the Kavli Foundation and its founder Fred Kavli. HH acknowledges support from NWO VIDI grant number 639.042.814 and ERC FP7 grant 279396. Work at Argonne National Laboratory was supported under U.S. Department of Energy contract DE-AC02-06CH11357. The Munich group acknowledges the support by the DFG Cluster of Excellence 'Origin and Structure of the Universe' and the Transregio program TR33 'The Dark Universe'. RJF is supported in part by fellowships from the Alfred P. Sloan Foundation and the David and Lucile Packard Foundation. TdH is supported by a Miller Research Fellowship. PS acknowledges support by the European DUEL Research-Training Network (MRTN-CT-2006-036133) and by the Deutsche Forschungsgemeinschaft under the project SCHN 342/7-1. CBM acknowledges the support of the DFG under Emmy Noether grant Hi 1495/2-1. BAB is supported by the Fermi Research Alliance, LLC under Contract No. De-AC02-07CH11359 with the United States Department of Energy. CLR acknowledges support from the Australian Research Council's Discovery Projects scheme (DP150103208). The Dark Cosmology Centre is funded by the Danish National Research Foundation.

REFERENCES

Albrecht A. et al., 2006, preprint (arXiv:astro-ph/0609591)
Allen S. W., Evrard A. E., Mantz A. B., 2011, ARA&A, 49, 409
Andersson K. et al., 2011, ApJ, 738, 48
Angulo R. E., Springel V., White S. D. M., Jenkins A., Baugh C. M., Frenk C. S., 2012, MNRAS, 426, 2046
Applegate D. E. et al., 2014, MNRAS, 439, 48
Applegate D. E. et al., 2016, MNRAS, 457, 1522
Arnaud M., Evrard A. E., 1999, MNRAS, 305, 631

Arnaud M., Pointecouteau E., Pratt G. W., 2005, A&A, 441, 893
Bartelmann M., Schneider P., 2001, Phys. Rep., 340, 291
Battaglia N. et al., 2016, JCAP, 8, 013
Bayliss M. B. et al., 2016, ApJS, 227, 3
Becker M. R., Kravtsov A. V., 2011, ApJ, 740, 25
Beckwith S. V. W. et al., 2006, AJ, 132, 1729
Benítez N., 2000, ApJ, 536, 571
Benítez N. et al., 2004, ApJS, 150, 1
Benjamin J. et al., 2013, MNRAS, 431, 1547
Benson B. A. et al., 2013, ApJ, 763, 147
Benson B. A. et al., 2014, in Holland W. S., Zmuidzinas J., eds, Proc. SPIE Conf. Ser. Vol. 9153, Millimeter, Submillimeter, and Far-Infrared Detectors and Instrumentation for Astronomy VII. SPIE, Bellingham, p. 91531P
Bernstein G. M., Armstrong R., Krawiec C., March M. C., 2016, MNRAS, 459, 4467
Bertin E., Arnouts S., 1996, A&AS, 117, 393
Bleem L. E. et al., 2015, ApJS, 216, 27
Bocquet S. et al., 2015, ApJ, 799, 214
Bocquet S., Saro A., Dolag K., Mohr J. J., 2016, MNRAS, 456, 2361
Böhringer H., Dolag K., Chon G., 2012, A&A, 539, A120
Bonnett C., 2015, MNRAS, 449, 1043
Bower R., 1997, MNRAS, 288, 355
Brammer G. B., van Dokkum P. G., Coppi P., 2008, ApJ, 686, 1503
Brammer G. B. et al., 2012, ApJS, 200, 13
Brammer G. B., van Dokkum P. G., Illingworth G. D., Bouwens R. J., Labbé I., Franx M., Momcheva I., Oesch P. A., 2013, ApJ, 765, L2
Broadhurst T. J., Taylor A. N., Peacock J. A., 1995, ApJ, 438, 49
Brodwin M. et al., 2010, ApJ, 721, 90
Bruzual G., Charlot S., 2003, MNRAS, 344, 1000
Carlstrom J. E. et al., 2011, PASP, 123, 568
Chiu I. et al., 2016a, MNRAS, 455, 258
Chiu I. et al., 2016b, MNRAS, 457, 3050
Coe D., Benítez N., Sánchez S. F., Jee M., Bouwens R., Ford H., 2006, AJ, 132, 926
Crawford S. M., Bershady M. A., Glenn A. D., Hoessel J. G., 2006, ApJ, 636, L13
Crawford S. M., Wirth G. D., Bershady M. A., Hon K., 2011, ApJ, 741, 98
Crawford S. M., Wirth G. D., Bershady M. A., 2014, ApJ, 786, 30
Crawford S. M., Wirth G. D., Bershady M. A., Randriamampandry S. M., 2016, ApJ, 817, 87
Davis M. et al., 2007, ApJ, 660, L1
De Boni C., Ettori S., Dolag K., Moscardini L., 2013, MNRAS, 428, 2921
de Haan T. et al., 2016, ApJ, 832, 95
Diemer B., Kravtsov A. V., 2015, ApJ, 799, 108
Dietrich J. P., Böhnert A., Lombardi M., Hilbert S., Hartlap J., 2012, MNRAS, 419, 3547
Dietrich J. P. et al., 2017, MNRAS, preprint (arXiv:1711.05344)
Dolag K., Komatsu E., Sunyaev R., 2016, MNRAS, in press
Duffy A. R., Schaye J., Kay S. T., Dalla Vecchia C., 2008, MNRAS, 390, L64
Duffy A. R., Schaye J., Kay S. T., Dalla Vecchia C., Battye R. A., Booth C. M., 2010, MNRAS, 405, 2161
Erben T., Van Waerbeke L., Bertin E., Mellier Y., Schneider P., 2001, A&A, 366, 717
Erben T. et al., 2005, Astron. Nachr., 326, 432
Fenech Conti I., Herbonnet R., Hoekstra H., Merten J., Miller L., Viola M., 2017, MNRAS, 467, 1627
Foley R. J. et al., 2011, ApJ, 731, 86
Ford J. et al., 2015, MNRAS, 447, 1304
George M. R. et al., 2012, ApJ, 757, 2
Giavalisco M. et al., 2004, ApJ, 600, L93
Grogin N. A. et al., 2011, ApJS, 197, 35
Gruen D. et al., 2014, MNRAS, 442, 1507
Gruen D., Seitz S., Becker M. R., Friedrich O., Mana A., 2015, MNRAS, 449, 4264
Gupta N., Saro A., Mohr J. J., Dolag K., Liu J., 2017, MNRAS, 469, 3069
Haiman Z., Mohr J. J., Holder G. P., 2001, ApJ, 553, 545
Hartlap J., Schrabback T., Simon P., Schneider P., 2009, A&A, 504, 689
Hasselfield M. et al., 2013, JCAP, 7, 8
Henderson S. W. et al., 2016, J. Low Temp. Phys, in press
Heymans C. et al., 2006, MNRAS, 368, 1323
Heymans C. et al., 2012, MNRAS, 427, 146
High F. W. et al., 2012, ApJ, 758, 68
Hilbert S., Hartlap J., White S. D. M., Schneider P., 2009, A&A, 499, 31
Hildebrandt H. et al., 2017, MNRAS, 465, 1454
Hinshaw G. et al., 2013, ApJS, 208, 19
Hoekstra H., 2001, A&A, 370, 743
Hoekstra H., 2003, MNRAS, 339, 1155
Hoekstra H., 2007, MNRAS, 379, 317
Hoekstra H., Franx M., Kuijken K., Squires G., 1998, ApJ, 504, 636
Hoekstra H., Franx M., Kuijken K., 2000, ApJ, 532, 88
Hoekstra H., Donahue M., Conselice C. J., McNamara B. R., Voit G. M., 2011a, ApJ, 726, 48
Hoekstra H., Hartlap J., Hilbert S., van Uitert E., 2011b, MNRAS, 412, 2095
Hoekstra H., Mahdavi A., Babul A., Bildfell C., 2012, MNRAS, 427, 1298
Hoekstra H., Bartelmann M., Dahle H., Israel H., Limousin M., Meneghetti M., 2013, Space Science Rev., 177, 75
Hoekstra H., Herbonnet R., Muzzin A., Babul A., Mahdavi A., Viola M., Cacciato M., 2015, MNRAS, 449, 685
Hoekstra H., Viola M., Herbonnet R., 2017, MNRAS, 468, 3295
Ilbert O. et al., 2009, ApJ, 690, 1236
Israel H., Erben T., Reiprich T. H., Vikhlinin A., Sarazin C. L., Schneider P., 2012, A&A, 546, A79
Jee M. J., Blakeslee J. P., Sirianni M., Martel A. R., White R. L., Ford H. C., 2007, PASP, 119, 1403
Jee M. J. et al., 2011, ApJ, 737, 59
Jee M. J., Hughes J. P., Menanteau F., Sifón C., Mandelbaum R., Barrientos L. F., Infante L., Ng K. Y., 2014, ApJ, 785, 20
Johnston D. E. et al., 2007, preprint (arXiv:0709.1159)
Kaiser N., Silk J., 1986, Nature, 324, 529
Kaiser N., Squires G., 1993, ApJ, 404, 441
Kaiser N., Squires G., Broadhurst T., 1995, ApJ, 449, 460
Kannawadi A., Mandelbaum R., Lackner C., 2015, MNRAS, 449, 3597
Kelly B. C., 2007, ApJ, 665, 1489
Kettula K. et al., 2015, MNRAS, 451, 1460
Kinney A. L., Calzetti D., Bohlin R. C., McQuade K., Storchi-Bergmann T., Schmitt H. R., 1996, ApJ, 467, 38
Koekemoer A. M., Fruchter A. S., Hook R. N., Hack W., 2003, in Arribas S., Koekemoer A., Whitmore B., eds, The 2002 HST Calibration Workshop: Hubble after the Installation of the ACS and the NICMOS Cooling System. Space Telescope Science Institute, Baltimore, MD, p. 337
Koekemoer A. M. et al., 2011, ApJS, 197, 36
Koekemoer A. M. et al., 2013, ApJS, 209, 3
Köhlinger F., Hoekstra H., Eriksen M., 2015, MNRAS, 453, 3107
Komatsu E. et al., 2009, ApJS, 180, 330
Koo D. C., Bershady M. A., Wirth G. D., Stanford S. A., Majewski S. R., 1994, ApJ, 427, L9
Koo D. C., Guzmán R., Gallego J., Wirth G. D., 1997, ApJ, 478, L49
Kuijken K. et al., 2015, MNRAS, 454, 3500
Laureijs R. et al., 2011, preprint (arXiv:1110.3193)
Le Brun A. M. C., McCarthy I. G., Schaye J., Ponman T. J., 2014, MNRAS, 441, 1270
Leauthaud A. et al., 2007, ApJS, 172, 219
Leauthaud A. et al., 2010, ApJ, 709, 97
Lieu M. et al., 2016, A&A, 592, A4
Lin Y., Mohr J. J., Stanford S. A., 2004, ApJ, 610, 745
LSST Dark Energy Science Collaboration, 2012, preprint (arXiv:1211.0310)
LSST Science Collaboration et al., 2009, preprint (arXiv:0912.0201)
Luppino G. A., Kaiser N., 1997, ApJ, 475, 20
Mantz A., Allen S. W., Ebeling H., Rapetti D., Drlica-Wagner A., 2010, MNRAS, 406, 1773
Mantz A. B. et al., 2015, MNRAS, 446, 2205
Mantz A. B. et al., 2016, MNRAS, 463, 3582
Massey R. et al., 2007, MNRAS, 376, 13
Massey R. et al., 2014, MNRAS, 439, 887

Mathiesen B. F., Evrard A. E., 2001, ApJ, 546, 100
Matthews D. J., Newman J. A., 2010, ApJ, 721, 456
Maughan B. J., Jones L. R., Ebeling H., Scharf C., 2006, MNRAS, 365, 509
McDonald M. et al., 2013, ApJ, 774, 23
McInnes R. N., Menanteau F., Heavens A. F., Hughes J. P., Jimenez R., Massey R., Simon P., Taylor A., 2009, MNRAS, 399, L84
Medezinski E., Broadhurst T., Umetsu K., Oguri M., Rephaeli Y., Benítez N., 2010, MNRAS, 405, 257
Melchior P. et al., 2017, MNRAS, 469, 4899
Menanteau F. et al., 2012, ApJ, 748, 7
Ménard B., Scranton R., Schmidt S., Morrison C., Jeong D., Budavari T., Rahman M., 2013, preprint (arXiv:1303.4722)
Merloni A. et al., 2012, preprint (arXiv:1209.3114)
Miller L. et al., 2013, MNRAS, 429, 2858
Miyazaki S. et al., 2012, in McLean I. S., Ramsay S. K., Takami H., eds, Proc. SPIE Conf. Ser. Vol. 8446, Ground-based and Airborne Instrumentation for Astronomy IV. SPIE, Bellingham, p. 84460Z
Momcheva I. G. et al., 2016, ApJS, 225, 27
Morrison C. B., Hildebrandt H., Schmidt S. J., Baldry I. K., Bilicki M., Choi A., Erben T., Schneider P., 2017, MNRAS, 467, 3576
Navarro J. F., Frenk C. S., White S. D. M., 1997, ApJ, 490, 493
Newman J. A., 2008, ApJ, 684, 88
Newman J. A. et al., 2015, Astropart. Phys., 63, 81
Oguri M., Bayliss M. B., Dahle H., Sharon K., Gladders M. D., Natarajan P., Hennawi J. F., Koester B. P., 2012, MNRAS, 420, 3213
Okabe N., Smith G. P., 2016, MNRAS, 461, 3794
Okabe N., Smith G. P., Umetsu K., Takada M., Futamase T., 2013, ApJ, 769, L35
Pickles A. J., 1998, PASP, 110, 863
Planck Collaboration XXVII, 2016a, A&A, 594, A27
Planck Collaboration XIII, 2016b, A&A, 594, A13
Planck Collaboration XXIV, 2016c, A&A, 594, A24
Rafelski M. et al., 2015, AJ, 150, 31
Rahman M., Ménard B., Scranton R., Schmidt S. J., Morrison C. B., 2015, MNRAS, 447, 3500
Rahman M., Mendez A. J., Ménard B., Scranton R., Schmidt S. J., Morrison C. B., Budavári T., 2016, MNRAS, 460, 163
Rapetti D., Allen S. W., Mantz A., Ebeling H., 2009, MNRAS, 400, 699
Rapetti D., Blake C., Allen S. W., Mantz A., Parkinson D., Beutler F., 2013, MNRAS, 432, 973
Reichardt C. L. et al., 2013, ApJ, 763, 127
Reichert A., Böhringer H., Fassbender R., Mühlegger M., 2011, A&A, 535, A4
Reiprich T. H., Böhringer H., 2002, ApJ, 567, 716
Rhodes J. D. et al., 2007, ApJS, 172, 203
Rix H.-W. et al., 2004, ApJS, 152, 163
Rowe B. T. P. et al., 2015, Astron. Comput., 10, 121
Rozo E. et al., 2010, ApJ, 708, 645
Rozo E., Wu H.-Y., Schmidt F., 2011, ApJ, 735, 118
Ruel J. et al., 2014, ApJ, 792, 45
Saro A. et al., 2014, MNRAS, 440, 2610
Schirmer M., 2013, ApJS, 209, 21
Schlegel D. J., Finkbeiner D. P., Davis M., 1998, ApJ, 500, 525
Schmidt F., Vikhlinin A., Hu W., 2009, Phys. Rev. D, 80, 083505
Schmidt S. J., Ménard B., Scranton R., Morrison C., McBride C. K., 2013, MNRAS, 431, 3307
Schneider P., 2006, in Meylan G., Jetzer P., North P., eds, Gravitational Lensing: Strong, Weak & Micro, Saas-Fee Advanced Course 33, Swiss Society for Astrophysics and Astronomy. Springer-Verlag, Berlin, p. 269
Schneider P., Seitz C., 1995, A&A, 294, 411
Schrabback T. et al., 2007, A&A, 468, 823
Schrabback T. et al., 2010, A&A, 516, A63
Scottez V. et al., 2016, MNRAS, 462, 1683
Scoville N. et al., 2007, ApJS, 172, 1
Sehgal N. et al., 2011, ApJ, 732, 44
Seitz C., Schneider P., 1997, A&A, 318, 687
Sifón C., Hoekstra H., Cacciato M., Viola M., Köhlinger F., van der Burg R. F. J., Sand D. J., Graham M. L., 2015, A&A, 575, A48
Simet M., Mandelbaum R., 2015, MNRAS, 449, 1259
Simet M., McClintock T., Mandelbaum R., Rozo E., Rykoff E., Sheldon E., Wechsler R. H., 2017, MNRAS, 466, 3103
Simon P., 2012, A&A, 543, A2
Simon P., Taylor A. N., Hartlap J., 2009, MNRAS, 399, 48
Skelton R. E. et al., 2014, ApJS, 214, 24
Spergel D. et al., 2015, preprint (arXiv:1503.03757)
Springel V. et al., 2005, Nature, 435, 629
Sunyaev R. A., Zel'dovich Y. B., 1970, Comments Astrophys. Space Phys., 2, 66
Sunyaev R. A., Zel'dovich Y. B., 1972, Comments Astrophys. Space Phys., 4, 173
Takahashi R., Sato M., Nishimichi T., Taruya A., Oguri M., 2012, ApJ, 761, 152
Teplitz H. I. et al., 2013, AJ, 146, 159
The Dark Energy Survey Collaboration, 2005, preprint (astro-ph/0510346)
Umetsu K. et al., 2014, ApJ, 795, 163
Umetsu K., Zitrin A., Gruen D., Merten J., Donahue M., Postman M., 2016, ApJ, 821, 116
van Uitert E., Gilbank D. G., Hoekstra H., Semboloni E., Gladders M. D., Yee H. K. C., 2016, A&A, 586, A43
Vanderlinde K. et al., 2010, ApJ, 722, 1180 (V10)
Vikhlinin A., McNamara B., Forman W., Jones C., Quintana H., Hornstrup A., 1998, ApJ, 502, 558
Vikhlinin A., Kravtsov A., Forman W., Jones C., Markevitch M., Murray S. S., Van Speybroeck L., 2006, ApJ, 640, 691
Vikhlinin A. et al., 2009a, ApJ, 692, 1033
Vikhlinin A. et al., 2009b, ApJ, 692, 1060
Viola M., Kitching T. D., Joachimi B., 2014, MNRAS, 439, 1909
von der Linden A. et al., 2014a, MNRAS, 439, 2
von der Linden A. et al., 2014b, MNRAS, 443, 1973
Weinberg D. H., Mortonson M. J., Eisenstein D. J., Hirata C., Riess A. G., Rozo E., 2013, Phys. Rep., 530, 87
Williamson R. et al., 2011, ApJ, 738, 139 (W11)
Wolf C., 2009, MNRAS, 397, 520
Wright C. O., Brainerd T. G., 2000, ApJ, 534, 34
Ziparo F., Smith G. P., Okabe N., Haines C. P., Pereira M. J., Egami E., 2016, MNRAS, 463, 4004

APPENDIX A: GALAXY ELLIPTICITY DISPERSION AND SHAPE MEASUREMENT WEIGHTS

As explained in Section 5 we processed ACS observations in the CANDELS fields to be able to mimic our source selection in the photometric redshift reference catalogues. These blank field data also enable us to study the galaxy ellipticity distribution as detailed in this appendix. On one hand this allows us to optimize our weighting scheme for the current study. In addition, these estimates can be used to optimize future weak lensing observing programmes and forecast their performance. For the latter purpose we have studied shape estimates from both ACS standard lensing filters *F*606*W* and *F*814*W*. This also updates earlier results on the intrinsic ellipticity dispersion estimated by Leauthaud et al. (2007) for *F*814*W* observations in the COSMOS Survey.

A1 Method

Our ellipticity measurements ϵ provide estimates for the reduced shear g. We model the measured dispersion of the galaxy ellipticity σ_ϵ with contributions from the intrinsic galaxy shapes $\sigma_{\rm int}$ and measurement noise $\sigma_{\rm m}$ as

$$\sigma_\epsilon^2 = \sigma_{\rm int}^2 + \sigma_{\rm m}^2. \quad \text{(A1)}$$

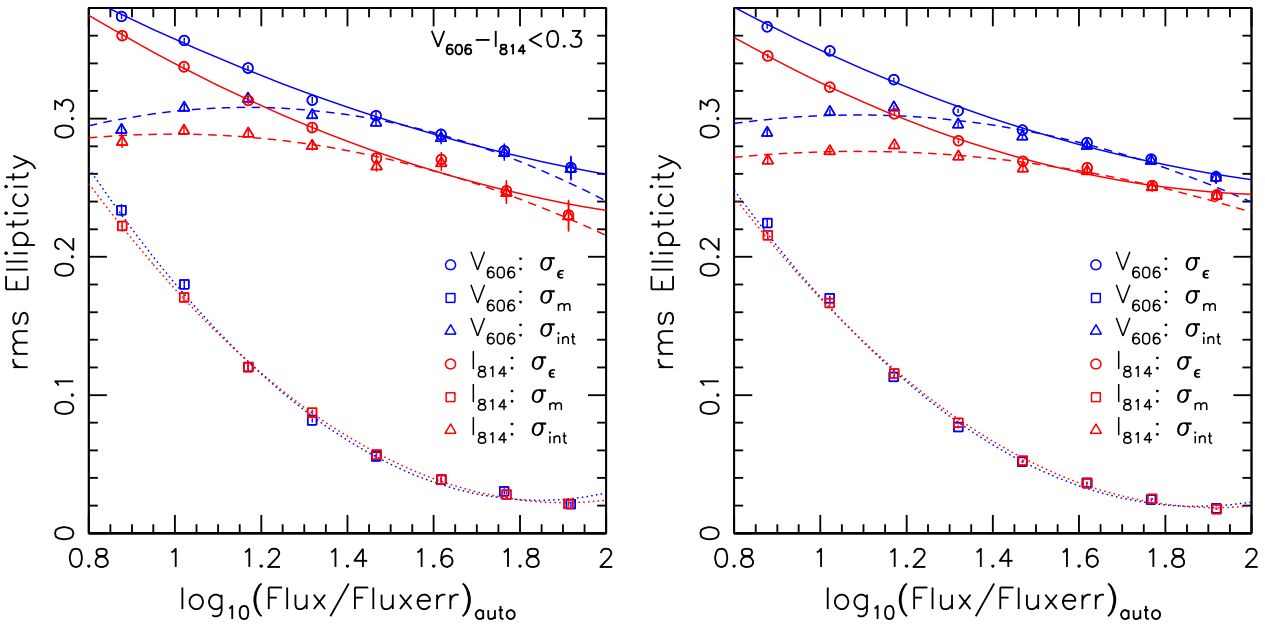

Figure A1. Galaxy ellipticity dispersion per ellipticity component as a function of the logarithmic flux signal-to-noise ratio (measured by SEXTRACTOR as FLUX_AUTO/FLUXERR_AUTO) with (*left*) and without colour selection (*right*), estimated from ACS *F*606*W* and *F*814*W* data in the CANDELS fields (see the text for details), and averaged over both ellipticity components. The circles show the r.m.s. of our KSB ellipticity estimates σ_ϵ, with polynomial interpolations indicated by the solid curves. The squares show the measurement noise σ_m estimated from the difference between the ellipticity estimates in overlapping tiles, with polynomial interpolations indicated by the dotted curves. The triangles show the estimate for intrinsic shape noise $\sigma_\mathrm{int} = \sqrt{\sigma_\epsilon^2 - \sigma_\mathrm{m}^2}$, with polynomial interpolations indicated by the dashed curves. The symbols mark the bin centres, and error-bars indicate the uncertainty estimated via bootstrapping.

The contribution from the cosmological shear in CANDELS is small compared to σ_ϵ, and for the purpose of this study we regard it as part of σ_int. To estimate σ_m we make use of the overlap region of neighbouring ACS tiles (that have similar noise properties), where we have two estimates (*a*,*b*) of the ellipticity of each galaxy with two independent realizations of the measurement noise for identical ϵ_int. After rotating the ellipticities to the same coordinate frame, the dispersion of their difference $\Delta\epsilon = \epsilon^\mathrm{a} - \epsilon^\mathrm{b}$ allows us to estimate

$$\sigma_\mathrm{m}^2 = \sigma_{\Delta\epsilon}^2/2, \tag{A2}$$

from which we compute σ_int according to (A1). Generally, we quote r.m.s. ellipticity values *per ellipticity component*, where we compute the average from both components as

$$\sigma_\epsilon^2 = \left(\sigma_{\epsilon,1}^2 + \sigma_{\epsilon,2}^2\right)/2. \tag{A3}$$

A2 Data

For this analysis we generated and analysed tile-wise *F*606*W* and *F*814*W* stacks of four ACS exposures each. We include the initial AEGIS ACS *F*606*W* and *F*814*W* observations (Davis et al. 2007, Proposal ID 10134). Similar to Schrabback et al. (2007) we generate *F*606*W* stacks in GOODS-South and GOODS-North that always combine two epochs of the observations from Giavalisco et al. (2004, Proposal IDs 9425, 9583). In GOODS-South we also include *F*606*W* observations from GEMS (Rix et al. 2004, Proposal ID 9500), which provides some additional overlap with the S14 WFC3/IR-detected catalogues. Generally, we limit our analysis to the overlap region with the S14 catalogues to enable the colour selection and provide constraints as a function of photometric redshift. For the COSMOS and UDS fields we use the *F*606*W* and *F*814*W* observations from CANDELS (Grogin et al. 2011, Proposal IDs 12440, 12064). Here, the tile-wise *F*606*W* stacks have slightly shorter integration times of 1.3–1.7 ks compared to our targeted ∼2 ks depth. For the constraints on the ellipticity dispersion we therefore include these observations only when studying the ellipticity dispersion as a function of flux signal-to-noise ratio, where the impact of the shallower depth is minimal.

A3 Discussion

We plot our estimates for the measured ellipticity dispersion σ_ϵ, the intrinsic ellipticity dispersion σ_int and the measurement noise σ_m for both ACS filters in Figs A1–A4.

We investigate the dependencies on the logarithmic flux signal-to-noise ratio $\log_{10}(\mathrm{Flux/Fluxerr})_\mathrm{auto}$, defined via the ratio FLUX_AUTO/FLUXERR_AUTO from SEXTRACTOR in Fig. A1, on the aperture magnitude in Fig. A2, and on the auto magnitude from SEXTRACTOR in Fig. A3, in all cases with (left-hand panels) and without (right-hand panels) applying our colour selection. As expected, the measurement noise σ_m increases steeply towards low signal-to-noise and fainter magnitudes. This is one of the reasons why σ_ϵ increases towards lower signal-to-noise and fainter magnitudes. Interestingly, we find that σ_int also increases towards fainter magnitudes. The analysis of COSMOS data by Leauthaud et al. (2007) also hinted at this trend with magnitude, but these authors discussed that it might be an artefact from their simplified estimator of the measurement error. We expect that our estimate of the measurement noise from overlapping tiles is fairly robust, and therefore suggest that this indeed appears to be a real effect, showing that intrinsically fainter galaxies have a broader ellipticity distribution.

As a function of the signal-to-noise ratio we largely observe the corresponding trend of an increasing σ_ϵ and σ_int towards lower $\log_{10}(\mathrm{Flux/Fluxerr})_\mathrm{auto}$, but note that our estimate for σ_int flattens at $\log_{10}(\mathrm{Flux/Fluxerr})_\mathrm{auto} \sim 1$–1.2 and eventually turns over to decreasing σ_int. Using stacks of different depth we verified that this

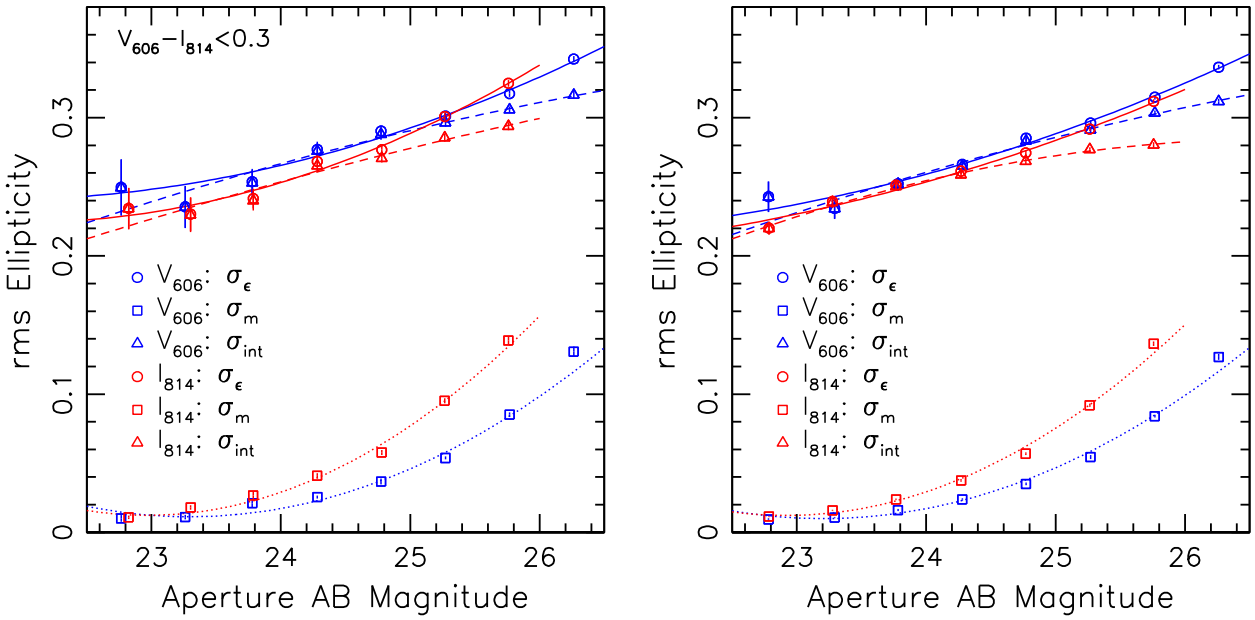

Figure A2. Galaxy ellipticity dispersion per ellipticity component as a function of AB magnitude V_{606} or I_{814}. See the caption of Fig. A1 for further details.

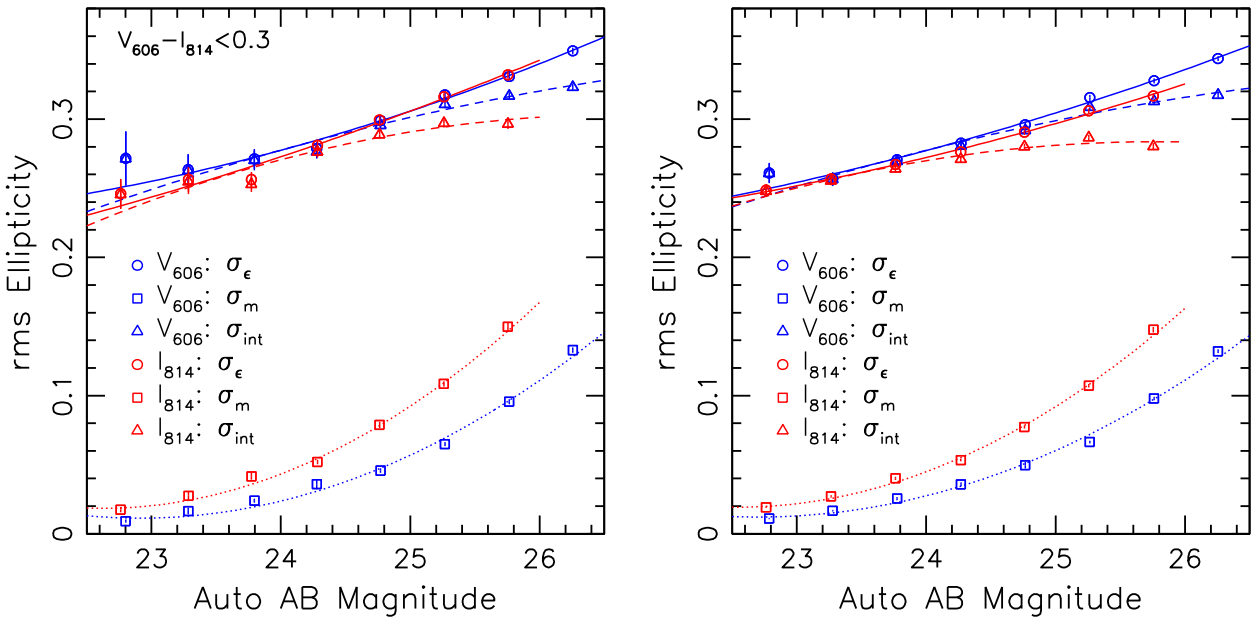

Figure A3. Galaxy ellipticity dispersion per ellipticity component as a function of AB auto magnitude from SExtractor. See the caption of Fig. A1 for further details.

flattening is not intrinsic to the galaxies. Instead, we expect that the validity of Equation (A1) breaks down for large σ_m. In addition, selection effects may have some influence, e.g. the cuts applied in size and $\mathrm{Tr}[P^g]/2$, as well as non-Gaussian tails in the measured ellipticity distribution at low signal-to-noise ratio.

Comparing the left- and right-hand panels in Figs A1–A3 we find that the application of our colour selection to remove cluster galaxies has only a relatively small impact on the ellipticity dispersion: Applying the colour selection $V_{606} - I_{814} < 0.3$ (which preferentially selects blue high-z background galaxies) increases σ_ϵ by 0.004 ± 0.002 (0.009 ± 0.002) and $\sigma_{\rm int}$ by 0.004 ± 0.002 (0.008 ± 0.002) at magnitudes $24 \leq \mathrm{mag_{aper}} \leq 26$ in the $F606W$ ($F814W$) filter. This can be compared to the dependence of the ellipticity dispersion on photometric redshift shown in Fig. A4, where we split the sample into bright (left-hand panel) and faint (right-hand panel) galaxies. Over the broad redshift range covered by the *HST* data the redshift dependence appears to be relatively weak. Most notably, the faint galaxies show an increase in σ_ϵ and $\sigma_{\rm int}$ between redshift 0 and ∼1. In principle, one expects such a trend, as galaxies at higher redshifts are observed at bluer rest-frame wavelengths, with stronger light contributions from sites of star formation. However note that it is more challenging to robustly infer conclusions on the redshift dependence of the shape distribution, as this is more strongly affected by large-scale structure variations (compare e.g. Kannawadi, Mandelbaum & Lackner 2015). We therefore suggest to investigate these trends further in the future with larger data sets.

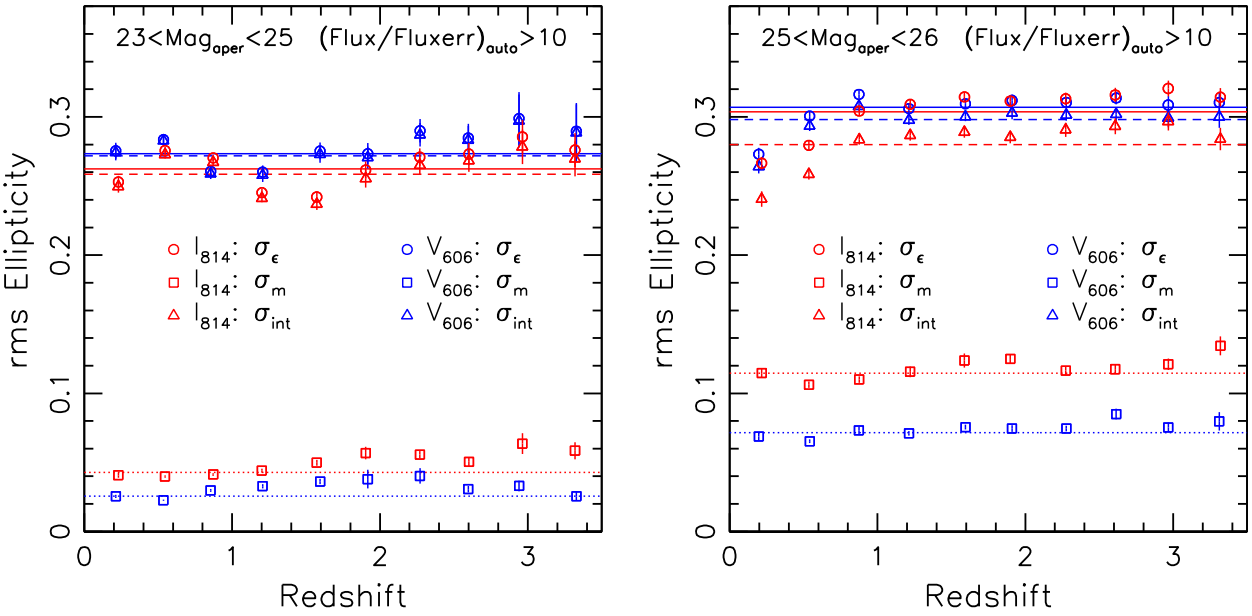

Figure A4. Galaxy ellipticity dispersion per ellipticity component as a function of photometric redshift for bright $23 < \mathrm{mag} < 25$ galaxies (*left*), and faint $25 < \mathrm{mag} < 26$ galaxies (*right*). The horizontal lines show the weighted averages. See the caption of Fig. A1 for further details.

A4 Comparing the weak lensing efficiency of *F*606*W* and *F*814*W*

In Figs A1–A3 $\sigma_{\rm int}$ is typically lower for the analysis of the *F*814*W* data than for the *F*606*W* images at a given signal-to-noise ratio or magnitude. However, when interpreting this one has to keep in mind that the bins do not contain identical sets of galaxies. To facilitate a fair direct comparison of the performance of both filters for weak lensing measurements we limit the analysis to the *F*606*W* and *F*814*W* AEGIS observations, which were taken under very similar conditions with similar exposure times. As a first test, we compare the ellipticity dispersions computed from those galaxies that have robust shape estimates and $(\mathrm{Flux/Fluxerr})_{\rm auto} > 10$ in *both* bands. Including the matched galaxies with $24 < V_{606} < 26$ we find that on average $\sigma_{\rm int}$ (σ_ϵ) is lower for the *F*814*W* shape estimates by 0.022 ± 0.003 (0.019 ± 0.003) compared to the *F*606*W* shapes when no colour selection is applied, and by 0.016 ± 0.006 (0.009 ± 0.004) when blue galaxies are selected with $V_{606} - I_{814} < 0.3$. Hence, we find that intrinsic galaxy shapes are slightly rounder when observed in the redder filter, which reduces their weak lensing shape noise. However, the quantity that actually sets the effective noise level for weak lensing studies is the effective source density after colour selection, which we define as

$$n_{\rm eff} = \sum_{\rm mag} n(\mathrm{mag}) \times \left(\frac{\sigma_\epsilon^{\rm ref}}{\sigma_\epsilon(\mathrm{mag})} \frac{\langle\beta\rangle(\mathrm{mag})}{\langle\beta\rangle^{\rm ref}} \right)^2 . \quad \text{(A4)}$$

For a cluster at $z_{\rm l} = 1.0$ we find from the AEGIS data that $n_{\rm eff}$ is higher by a factor 1.28 (1.06) for *F*606*W* compared to *F*814*W* when applying (when not applying) the colour selection with $V_{606} - I_{814} < 0.3$. Hence, if only a single band is observed with *HST*, *F*606*W* is slightly more efficient for the shape measurements than *F*814*W*. However, given that the ratio between the estimates is close to unity, we expect that programmes which have observations in *both* *F*606*W* and *F*814*W* can achieve a higher effective source density when jointly estimating shapes from both bands. Our work has shown the necessity for depth-matched colours for the cluster member removal. Therefore, we suggest that future *HST* weak lensing programmes of clusters at $0.7 \lesssim z_{\rm l} \lesssim 1.1$ should consider to split their observations between *F*606*W* and *F*814*W* to obtain both colour estimates and joint shape measurements from both bands.

A5 Fitting functions and shape weights

We compute second-order polynomial interpolations for the ellipticity dispersions $y \in \{\sigma_\epsilon, \sigma_{\rm int}, \sigma_{\rm m}\}$ as a function of logarithmic signal-to-noise and magnitude $x \in \{\log_{10}(\mathrm{Flux/Fluxerr})_{\rm auto}, \mathrm{Mag}_{\rm aper}, \mathrm{Mag}_{\rm auto}\}$ within limits $x_{\rm min} < x < x_{\rm max}$ as

$$y = a + b\hat{x} + c\hat{x}^2, \quad \text{(A5)}$$

where $\hat{x} = x - x_{\rm min}$. For our weak lensing analysis of SPT clusters we compute empirical shape weights for galaxy i as

$$w_i = \left[\sigma_\epsilon^{\rm fit}\left(\log_{10}(\mathrm{Flux/Fluxerr})_{{\rm auto},i}\right)\right]^{-2} \quad \text{(A6)}$$

from the interpolation of σ_ϵ as a function of the logarithmic signal-to-noise ratio for the $V_{606} - I_{814} < 0.3$ colour-selected CANDELS galaxies. We plot the best-fitting interpolations in Figs A1–A3 and list their polynomial coefficients in Table A1.

APPENDIX B: NON-MATCHING GALAXIES IN CANDELS

We have investigated the ∼2.4 per cent of non-matching galaxies between our CANDELS *F*606*W* shear catalogue and the S14 photo-*z* catalogue (see Section 6.1) by visually inspecting a random subset. Most of the non-matching galaxies can be explained through differences in the object detection or deblending given the different detection bands (optical *F*606*W* versus NIR *F*125*W*+*F*140*W*+*F*160*W*). For ∼0.7 per cent of the total galaxies centroid shifts prevent a match. These should not affect the source redshift distribution. For ∼1.2 per cent the S14 catalogue contains a single object which is associated with two deblended objects in our *F*606*W* shear catalogue. If such differences in the deblending would occur independent of

Table A1. Parameters and coefficients for the polynomial interpolation of the ellipticity dispersions σ_ϵ, $\sigma_{\rm int}$ and $\sigma_{\rm m}$ in CANDELS as a function of magnitude and logarithmic signal-to-noise ratio.

Band	Colour	x	$x_{\rm min}$	$x_{\rm max}$	y	a	b	c
I_{814}	All	$\log_{10}(\mathrm{Flux/Fluxerr})_{\rm auto}$	0.75	2	σ_ϵ	0.36777	−0.18359	0.06843
I_{814}	All	$\log_{10}(\mathrm{Flux/Fluxerr})_{\rm auto}$	0.75	2	$\sigma_{\rm int}$	0.27050	0.03504	−0.05252
I_{814}	All	$\log_{10}(\mathrm{Flux/Fluxerr})_{\rm auto}$	0.75	2	$\sigma_{\rm m}$	0.26390	−0.42101	0.18058
I_{814}	All	$\mathrm{Mag}_{\rm aper}$	22.5	26	σ_ϵ	0.22123	0.01644	0.00340
I_{814}	All	$\mathrm{Mag}_{\rm aper}$	22.5	26	$\sigma_{\rm int}$	0.21232	0.03411	−0.00402
I_{814}	All	$\mathrm{Mag}_{\rm aper}$	22.5	26	$\sigma_{\rm m}$	0.01480	−0.01211	0.01453
I_{814}	All	$\mathrm{Mag}_{\rm auto}$	22.5	26	σ_ϵ	0.24301	0.01649	0.00201
I_{814}	All	$\mathrm{Mag}_{\rm auto}$	22.5	26	$\sigma_{\rm int}$	0.23712	0.02839	−0.00433
I_{814}	All	$\mathrm{Mag}_{\rm auto}$	22.5	26	$\sigma_{\rm m}$	0.01925	−0.00090	0.01200
I_{814}	$V_{606}-I_{814}<0.3$	$\log_{10}(\mathrm{Flux/Fluxerr})_{\rm auto}$	0.75	2	σ_ϵ	0.38420	−0.19190	0.05716
I_{814}	$V_{606}-I_{814}<0.3$	$\log_{10}(\mathrm{Flux/Fluxerr})_{\rm auto}$	0.75	2	$\sigma_{\rm int}$	0.28447	0.03555	−0.07253
I_{814}	$V_{606}-I_{814}<0.3$	$\log_{10}(\mathrm{Flux/Fluxerr})_{\rm auto}$	0.75	2	$\sigma_{\rm m}$	0.27431	−0.43743	0.18966
I_{814}	$V_{606}-I_{814}<0.3$	$\mathrm{Mag}_{\rm aper}$	22.5	26	σ_ϵ	0.22602	0.00757	0.00698
I_{814}	$V_{606}-I_{814}<0.3$	$\mathrm{Mag}_{\rm aper}$	22.5	26	$\sigma_{\rm int}$	0.21238	0.02943	−0.00130
I_{814}	$V_{606}-I_{814}<0.3$	$\mathrm{Mag}_{\rm aper}$	22.5	26	$\sigma_{\rm m}$	0.01583	−0.01478	0.01571
I_{814}	$V_{606}-I_{814}<0.3$	$\mathrm{Mag}_{\rm auto}$	22.5	26	σ_ϵ	0.23050	0.02525	0.00195
I_{814}	$V_{606}-I_{814}<0.3$	$\mathrm{Mag}_{\rm auto}$	22.5	26	$\sigma_{\rm int}$	0.22288	0.03886	−0.00469
I_{814}	$V_{606}-I_{814}<0.3$	$\mathrm{Mag}_{\rm auto}$	22.5	26	$\sigma_{\rm m}$	0.01869	−0.00322	0.01307
V_{606}	All	$\log_{10}(\mathrm{Flux/Fluxerr})_{\rm auto}$	0.75	2	σ_ϵ	0.38882	−0.16903	0.05008
V_{606}	All	$\log_{10}(\mathrm{Flux/Fluxerr})_{\rm auto}$	0.75	2	$\sigma_{\rm int}$	0.29414	0.05089	−0.07555
V_{606}	All	$\log_{10}(\mathrm{Flux/Fluxerr})_{\rm auto}$	0.75	2	$\sigma_{\rm m}$	0.27001	−0.44604	0.19850
V_{606}	All	$\mathrm{Mag}_{\rm aper}$	22.5	26.5	σ_ϵ	0.22918	0.01439	0.00371
V_{606}	All	$\mathrm{Mag}_{\rm aper}$	22.5	26.5	$\sigma_{\rm int}$	0.21549	0.03276	−0.00186
V_{606}	All	$\mathrm{Mag}_{\rm aper}$	22.5	26.5	$\sigma_{\rm m}$	0.01564	−0.01605	0.01140
V_{606}	All	$\mathrm{Mag}_{\rm auto}$	22.5	26.5	σ_ϵ	0.24435	0.01885	0.00208
V_{606}	All	$\mathrm{Mag}_{\rm auto}$	22.5	26.5	$\sigma_{\rm int}$	0.23647	0.03082	−0.00233
V_{606}	All	$\mathrm{Mag}_{\rm auto}$	22.5	26.5	$\sigma_{\rm m}$	0.01257	−0.00372	0.00912
V_{606}	$V_{606}-I_{814}<0.3$	$\log_{10}(\mathrm{Flux/Fluxerr})_{\rm auto}$	0.75	2	σ_ϵ	0.39491	−0.16019	0.04158
V_{606}	$V_{606}-I_{814}<0.3$	$\log_{10}(\mathrm{Flux/Fluxerr})_{\rm auto}$	0.75	2	$\sigma_{\rm int}$	0.29096	0.08216	−0.09812
V_{606}	$V_{606}-I_{814}<0.3$	$\log_{10}(\mathrm{Flux/Fluxerr})_{\rm auto}$	0.75	2	$\sigma_{\rm m}$	0.28751	−0.48200	0.22022
V_{606}	$V_{606}-I_{814}<0.3$	$\mathrm{Mag}_{\rm aper}$	22.5	26.5	σ_ϵ	0.24319	0.00763	0.00486
V_{606}	$V_{606}-I_{814}<0.3$	$\mathrm{Mag}_{\rm aper}$	22.5	26.5	$\sigma_{\rm int}$	0.22404	0.03115	−0.00180
V_{606}	$V_{606}-I_{814}<0.3$	$\mathrm{Mag}_{\rm aper}$	22.5	26.5	$\sigma_{\rm m}$	0.01884	−0.01892	0.01190
V_{606}	$V_{606}-I_{814}<0.3$	$\mathrm{Mag}_{\rm auto}$	22.5	26.5	σ_ϵ	0.24607	0.01653	0.00295
V_{606}	$V_{606}-I_{814}<0.3$	$\mathrm{Mag}_{\rm auto}$	22.5	26.5	$\sigma_{\rm int}$	0.23311	0.03313	−0.00234
V_{606}	$V_{606}-I_{814}<0.3$	$\mathrm{Mag}_{\rm auto}$	22.5	26.5	$\sigma_{\rm m}$	0.01309	−0.00857	0.01044

redshift, there would be no net effect on the source redshift distribution. However, such differences might be more frequent for high-z ($z \gtrsim 1$) galaxies, where the *F*606*W* images probe rest-frame UV wavelengths and mostly detect sites of star formation, while the IR imaging probes the stellar content of the galaxies. Finally, ∼0.4 per cent of the total galaxies show clear isolated galaxies in our *F*606*W* shear catalogue that are missing in the S14 NIR-detected catalogue, possibly because they are too faint and too blue.

To obtain a rough estimate for the resulting uncertainty of these effects on our analysis, we assume a scenario where both the missing isolated galaxies (∼0.4 per cent) plus the excess half of the differently deblended galaxies (∼0.6 per cent) constitute an excess population of 100 per cent blue ($V_{606}-I_{814}<0.3$) galaxies at high redshifts ($z \simeq 2$). This scenario is pessimistic for the differently deblended galaxies as explained above (no impact if the effect is redshift independent). For the missing isolated galaxies the scenario is likely to be realistic, but we note that it would also overestimate the impact in case some of the galaxies are redder and removed by our $V_{606}-I_{814}<0.3$ colour selection. At our median cluster redshift $z_{\rm l}=0.88$ the scenario leads to a *relative* increase in $\langle\beta\rangle$ by only +0.5 per cent, thanks to our colour selection which already selects mostly $z>1$ galaxies.

APPENDIX C: CROSS-CHECK FOR THE REDSHIFT DISTRIBUTION USING SPATIAL CROSS-CORRELATIONS

A number of studies have explored the use of spatial cross-correlation techniques to constrain source redshift distributions (e.g. Newman 2008; Matthews & Newman 2010; Benjamin et al. 2013). In particular, Newman (2008), Matthews & Newman (2010), Schmidt et al. (2013), Rahman et al. (2015), Rahman et al. (2016) and Scottez et al. (2016) aim at reconstructing the redshift distribution of a sample with an unknown redshift distribution ('photometric sample') via its spatial cross-correlation with galaxies in redshift slices of an incomplete spectroscopic reference sample. The cross-correlation amplitude increases if a larger fraction of the photometric sample is located within the redshift range of the corresponding slice. As a result, information on the redshift distribution of the photometric sample can be inferred. When using photometric samples with a broad redshift distribution the accuracy of the method is limited by how well a potential redshift evolution of the relative galaxy bias between the populations can be accounted for (e.g. Rahman et al. 2015). However, the impact of this limitation can be reduced if the photometric sample can be split into subsamples with relatively narrow individual redshift distributions, as suggested

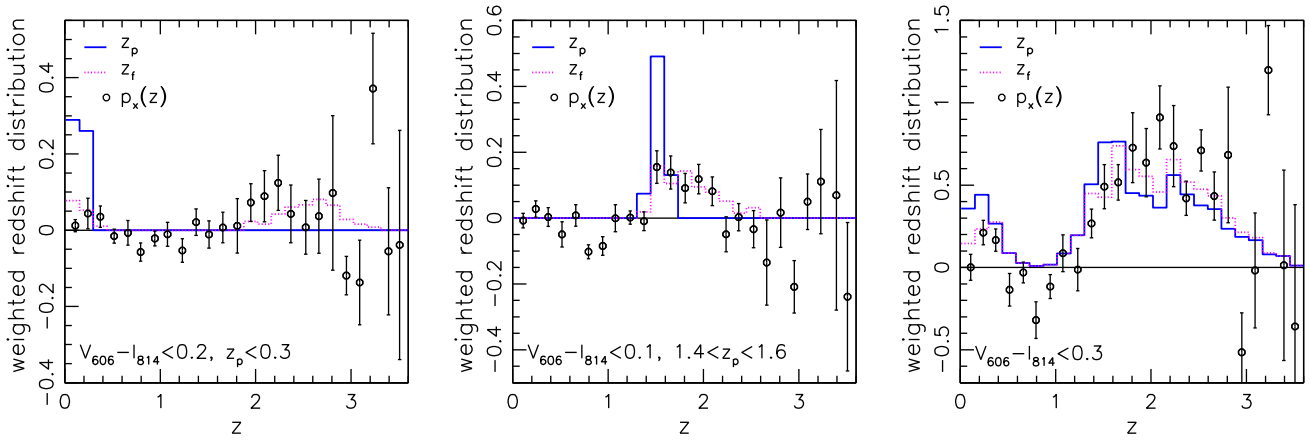

Figure C1. Comparison of the histograms of the 3D-HST peak photometric redshift z_p (blue solid) and the redshifts z_f that are statistically corrected based on the HUDF comparison (magenta dotted) to the reconstructed redshift distribution $p_x(z)$ inferred from the cross-correlation analysis (black circles) using colour-selected CANDELS galaxies with $24 < V_{606} < 26.5$ and applying shape weights from our CANDELS shear catalogue. The *left* and *middle* panels correspond to the galaxies for which we apply the corrections for catastrophic outliers or redshift focusing, respectively, while the *right*-hand panel includes the full sample. Error-bars show the dispersion of the $p_x(z)$ estimates when splitting the total sample into ten subareas and bootstrapping the contributing subareas. The large scatter at $z \gtrsim 2.8$ is caused by the small spectroscopic sample at these redshifts. The negative peak at $0.7 \lesssim z \lesssim 1.0$ is an artefact resulting from spatial density variations in the spectroscopic sample and the colour selection applied to the photometric sample.

by Schmidt et al. (2013) and Ménard et al. (2013), and applied to SDSS data in Rahman et al. (2016). The CANDELS data are well suited to employ this technique, as considerable spectroscopic (or grism) redshift samples are available (Section 6.3.3), and given that the 3D-HST photo-zs allow for a relatively clean subdivision into narrower redshift slice for most of the galaxies.

We employ the THE-WIZZ[16] implementation (Morrison et al. 2017) of the cross-correlation technique described in Schmidt et al. (2013) and Ménard et al. (2013) to obtain an independent cross-check for our estimate of the colour-selected CANDELS redshift distribution. For this we use the combined sample of high-fidelity spectroscopic and high-quality grism redshifts (see Section 6.3.3) as spectroscopic reference sample (without colour selection) and the colour-selected photo-z sample as photometric sample, splitting galaxies into 25 linear bins in $z_{s/g}$ or z_p, respectively, between $z = 0.01$ and $z = 3.6$. We compare the estimate for the redshift probability distribution $p_x(z)$ obtained from the cross-correlation analysis using physical separations between 30 and 300 kpc to the z_p and z_f histograms in Fig. C1 using galaxies with $24 < V_{606} < 26.5$ and the actual shape weights from our CANDELS shear catalogue.

The left-hand and the middle panels of Fig. C1 correspond to the subset of CANDELS galaxies for which we implemented statistical corrections (see Section 6.3) for catastrophic redshift outliers ($V_{606} - I_{814} < 0.2$, $z_p < 0.3$) or redshift focusing ($V_{606} - I_{814} < 0.1$, $1.4 < z_p < 1.6$), respectively. In both cases we find that the redshift distribution inferred from the cross-correlation analysis is largely consistent with the statistically corrected distribution based on the HUDF analysis (z_f), while it is clearly incompatible with the uncorrected distribution in the selected z_p ranges, providing an independent confirmation for the HUDF-based correction scheme. The right-hand panel of Fig. C1 shows the combined $p_x(z)$ reconstruction for the full colour selected sample ($V_{606} - I_{814} < 0.3$). Consistent with the other panels the reconstruction describes the z_f histogram better than the z_p histogram, both at low redshifts ($z < 0.3$) and around the broad peak at $z \sim 2$.

[16] https://github.com/morriscb/The-wiZZ

The statistical error-bars shown in Fig. C1 indicate the dispersion of the $p_x(z)$ reconstruction when splitting the combined CANDELS data set into 10 subareas of equal area and obtaining 1000 bootstrap resamples of the subareas included in the analysis. We expect that this yields a good approximation for the statistical uncertainty for most of the redshift range of interest. However, at the highest redshifts ($z \gtrsim 2.8$) the spectroscopic samples become very small (compare Fig. 5), likely introducing additional uncertainties that are not fully captured by the error-bars. This is also suggested by the large fluctuations of both the recovered $p_x(z)$ and the error-bars between neighbouring high-z bins.

We note the substantially negative $p_x(z)$ reconstructions at $0.7 \lesssim z \lesssim 1.0$ in the middle and right-hand panel of Fig. C1. At these redshifts the full spectroscopic sample contains a large number of galaxies (no colour selection applied to the spectroscopic sample). We therefore expect that the error-bars are robust and that the negative $p_x(z)$ estimates are indeed significant. We interpret these negative $p_x(z)$ values as a spurious effect caused by our colour selection, which explicitly removes galaxies at these redshifts from the photometric sample. Therefore, the photometric sample is spatially underdense in regions that are physically overdense at these redshifts. In contrast, the spectroscopic sample is spatially overrepresented in regions of physical overdensities at these redshifts. This results in a net anticorrelation between the samples and negative $p_x(z)$ estimates. As a possible solution to this problem Rahman et al. (2015) suggest to homogenize the spatial density of the spectroscopic sample by removing galaxies in overdense regions. However, as the spectroscopic sample employed in our analysis (14 472 galaxies) is already much smaller than the sample employed by Rahman et al. (2015) (791 546 galaxies) we do not follow this approach. As an approximate solution for this systematic effect we instead set the $p_x(z)$ values of the two bins in Fig. C1 at $0.7 < z < 1.0$ to zero when computing $\langle\beta\rangle$ as described in the next paragraph. This is justified by multiple tests presented in this work that suggest that the residual contamination by galaxies at these redshifts should be very low and close to zero (Sections 6.2, 6.3.3, 6.4 and 6.8). However, outside this redshift range we treat bins with negative $p_x(z)$ as negative contributions in the computation of $\langle\beta\rangle$. This is needed

in order to achieve unbiased results in the case of purely statistical scatter that has equal chance to be positive or negative.

For a quantitative comparison of the $p_x(z)$ distribution and the histograms shown in the right-hand panel of Fig. C1 we compute $\langle\beta\rangle$ for our median cluster redshift $z_l = 0.88$, dealing with negative $p_x(z)$ as explained in the previous paragraph, and generally limiting the considered redshift range to $z < 3.2$ to minimize the impact of the highest z data points for which the $p_x(z)$ recovery suffers the strongest from the small spectroscopic sample. The resulting $\langle\beta(p_x(z))\rangle = 0.403 \pm 0.017$ from the cross-correlation analysis (the error indicates the statistical scatter from the bootstrap resamples) is consistent with $\langle\beta(z_f)\rangle = 0.366 \pm 0.008$ from the HUDF-corrected catalogues within 2σ. We conclude that the cross-correlation analysis independently supports the results from the HUDF analysis, but note that the spectroscopic samples within the relatively small CANDELS areas are not yet sufficiently large to constrain the redshift distribution with very high precision.

APPENDIX D: DETAILS OF THE ACS+FORS2 COLOUR MEASUREMENTS AND THE ACCOUNTING FOR PHOTOMETRIC SCATTER

D1 ACS+FORS2 colour measurement

To measure colours between the *F*606*W* and FORS2 *I*-band images we convolve each mosaic *F*606*W* image with a Gaussian kernel

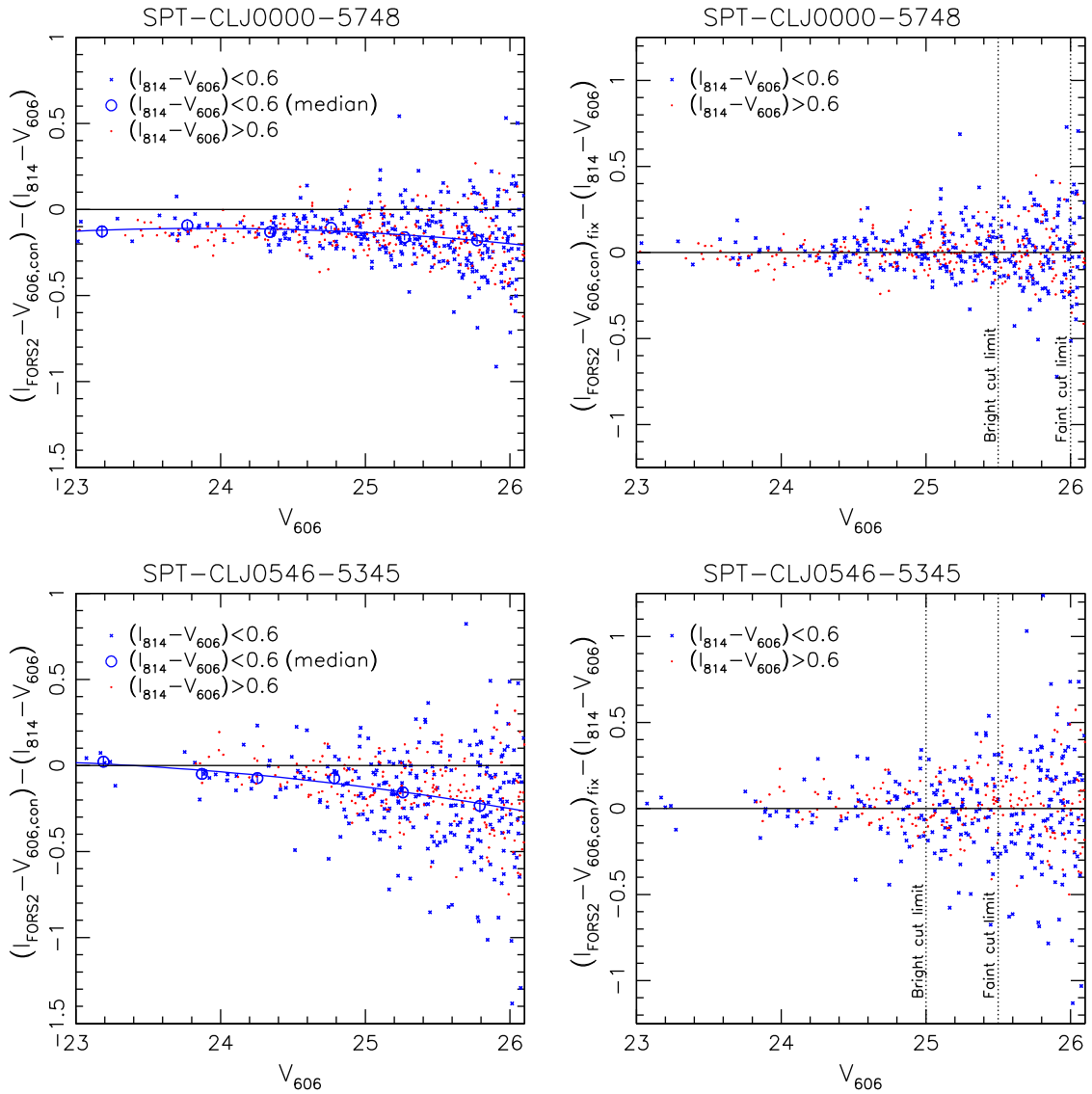

Figure D1. Details on the colour selection for SPT-CL J0000−5748 (*top*) and SPT-CL J0546−5345 (*bottom*). *Left:* Difference between the colour ($I_{FORS2} - V_{606,con}$) measured with the FORS2 *I* band and the convolved *HST*/ACS *F*606*W* images, and the colour ($I_{814} - V_{606}$) measured from the unconvolved *HST*/ACS data in the inner cluster region, as a function of V_{606}. Small blue crosses indicate blue galaxies with ($I_{814} - V_{606}$) < 0.6, while red points show red galaxies with ($I_{814} - V_{606}$) > 0.6. The open circles mark the median values for the blue galaxies within 0.5 mag wide magnitude bins, with error-bars indicating the uncertainty on the mean for a Gaussian distribution, and the curve showing their best-fitting second-order polynomial interpolation. The *right*-hand panels show the same data after subtraction of this function. We sample the photometric scatter distribution for the ACS-FORS2 selection from this distribution of offsets. Because of the lower scatter in the deeper FORS2 data of SPT-CL J0000−5748 we can include fainter galaxies in the ACS-FORS2 selection than for SPT-CL J0546−5345 (see Table 2 and the indicated bright/faint cut limits).

such that the resulting PSF has the same FLUX_RADIUS measured by SOURCE EXTRACTOR as the corresponding FORS2 *I*-band image (we empirically account for the impact of non-Gaussian VLT PSF profiles in Appendix D2). For some of the FORS2 stacks we found small residual systematic offsets of object positions in some image regions with respect to their location in the corresponding ACS mosaic (typically $\lesssim 0.3$ arcsec). To not bias the colour measurement, we therefore fit and subtract a smooth fifth-order 2D-polynomial interpolation of the measured positional offsets to the catalogue positions. We overlayed and visually inspected these corrections on all images to ensure that they are robust. We then measure object fluxes in circular apertures with diameter 1.5 arcsec both in the VLT and the convolved ACS image. We transform them into magnitudes, correct these for galactic extinction and compute the colour estimate $V_{606,\mathrm{con}} - I_{\mathrm{FORS2}}$.

D2 Tying the ACS+FORS2 colours to the ACS-only colours

We have ACS-based $V_{606} - I_{814}$ and ACS+FORS2-based $V_{606,\mathrm{con}} - I_{\mathrm{FORS2}}$ colour estimates for the galaxies in the inner cluster regions. We use these galaxies to refine the calibration of the $V_{606,\mathrm{con}} - I_{\mathrm{FORS2}}$ colours for all galaxies and tie them to the $V_{606} - I_{814}$ colour selection available in the 3D-HST CANDELS catalogues. The left-hand panels of Fig. D1 show the difference of these colour estimates as a function of V_{606} for two example clusters. The top row corresponds to SPT-CL J0000−5748, which has one of the deepest and best-seeing FORS2 *I*-band stacks in our sample, resulting in relatively moderate photometric scatter. Here the analysis reveals a ∼0.11 mag colour offset for bright galaxies. We expect that this offset is in part caused by the offset in equation (16). Further contributions might come from uncertainties in the I_{FORS2} zero-point calibration due to the small number of stars available for its determination, or inaccuracies in the PSF homogenization. In comparison, the bottom row reveals a larger photometric scatter for SPT-CL J0546−5345, which has a shallower magnitude limit and worse image quality (see Table 2). For such VLT data we typically detect a shift of the median colour difference (indicated through the open circles) at faint magnitudes towards negative values. In part this is caused by the asymmetric and biased scatter in logarithmic magnitude space. However, further effects could lead to a magnitude-dependent colour offset: for example, we acknowledge that our PSF homogenization only ensures equal flux radii between the bands. However, residual differences in the actual PSF shapes might lead to slightly different fractions in the total PSF flux lost outside the aperture. This would lead to a magnitude-dependent colour offset given that fainter objects are typically less resolved. Understanding the exact combination of these effects for each cluster is not necessary given that we directly tie the $V_{606,\mathrm{con}} - I_{\mathrm{FORS2}}$ colours to the $V_{606} - I_{814}$ colours empirically: To do so, we fit the median values of the colour offsets determined in 0.5 mag-wide bins between $23 < V_{606} < 26$ with a second-order polynomial in V_{606} and subtract this model from all $V_{606,\mathrm{con}} - I_{\mathrm{FORS2}}$ colour estimates in the cluster field to obtain $(V_{606,\mathrm{con}} - I_{\mathrm{FORS2}})_{\mathrm{fix}}$ (see Fig. D1). We only use relatively blue galaxies with $V_{606} - I_{814} < 0.6$ to derive this fit. This is motivated by small differences in the effective filter curves of I_{FORS2} and I_{814}. In particular, I_{FORS2} cuts off transmission red-wards of ∼870 nm, while I_{814} has a transmission tail out to ∼960 nm. Thus, we expect non-negligible colour differences for very red objects. Given that we generally apply fairly blue cuts in colour this is not a problem for our analysis. However, we exclude red galaxies when deriving the fit as they are overrepresented compared to CANDELS in the cluster fields.

Table D1. Overview of $V - I$ colour cut limits applied in our analysis.

z_{l}	$V_{606} - I_{814}$		$(V_{606,\mathrm{con}} - I_{\mathrm{FORS2}})_{\mathrm{fix}}$	
	$24 < V_{606} < 26$	$26 < V_{606} < 26.5$	Bright	Faint
<1.01	0.3	0.2	0.2	0.0
>1.01	0.2	0.1	0.1	−0.1

Note. Colour cut limits applied in our analysis. *Column 1:* Cluster redshift range. *Column 2:* Colour-cut in ACS-only colour $V_{606} - I_{814}$ for galaxies with $24 < V_{606} < 26$. *Column 3:* Colour-cut in ACS-only colour $V_{606} - I_{814}$ for galaxies with $26 < V_{606} < 26.5$. *Column 4:* Colour cut in the ACS+FORS2 colour $(V_{606,\mathrm{con}} - I_{\mathrm{FORS2}})_{\mathrm{fix}}$ after tying it to the $V_{606} - I_{814}$ colour (see Appendix D2), as employed for 'bright cut' magnitude bins with low photometric scatter $\sigma_{\Delta(V-I)} < 0.2$. *Column 5:* As column 4, but for the 'faint cut' magnitude bins with increased photometric scatter $0.2 < \sigma_{\Delta(V-I)} < 0.3$.

D3 Accounting for photometric scatter

D3.1 ACS-only colour selection

In the inner cluster regions covered by the *F*606*W* and *F*814*W* ACS images we include galaxies in the magnitude range $24 < V_{606} < 26.5$. The brighter magnitude limit has been chosen as galaxies passing our colour selection at even brighter magnitudes are dominated by foreground galaxies. The fainter magnitude limit approximately matches the S/N cut applied in the weak lensing shape analysis (see Section 5). Our ACS images have typical 5σ limits for the adopted 0.7 arcsec apertures of $V_{606,\mathrm{lim}} = 27.15$ and $I_{814,\mathrm{lim}} = 26.60$. Therefore, the faintest galaxies included at the colour cut ($V_{606} = 26.5$, $V_{606} - I_{814} = 0.3$) still have fairly high photometric signal-to-noise $(\mathrm{S/N})_{606} = 9.1$ and $(\mathrm{S/N})_{814} = 7.2$. Accordingly, photometric noise has only minor impact on the colour selection for these galaxies. None the less, we account for it by adding random Gaussian scatter to the S14 catalogues, which are typically based on deeper ACS mosaic stacks compared to the ones used for our shape analysis, prior to the colour selection, such that they have the same limiting magnitudes in V_{606} and I_{814} as our cluster field observations. Also, we apply a slightly bluer colour selection for the galaxies in the faintest magnitude bin (see Table D1 and Section 6.2).

D3.2 ACS+FORS2 colour selection

The colour estimates $(V_{606,\mathrm{con}} - I_{\mathrm{FORS2}})_{\mathrm{fix}}$ that include the FORS2 data are more strongly affected by photometric scatter than the $V_{606} - I_{814}$ colours obtained from the high-resolution ACS data only (see Fig. D1). To ensure that we can still apply a consistent colour selection to the S14 catalogues we do the following.

First, we limit the analysis to relatively bright V_{606} magnitudes, to ensure that the scatter is small enough to not compromise the exclusion of galaxies at the cluster redshift considerably. For this we compute the r.m.s. scatter $\sigma_{\Delta(V-I)}$ in the colour difference $\Delta(V-I) \equiv (V_{606,\mathrm{con}} - I_{\mathrm{FORS2}})_{\mathrm{fix}} - (V_{606} - I_{814})$ of blue galaxies ($V_{606} - I_{814} < 0.6$) in 0.5 mag-wide bins in V_{606}. For the ACS+FORS2 colour selection we only include magnitudes bins with scatter $\sigma_{\Delta(V-I)} < 0.3$. Here we employ our standard ('bright') colour cut for the magnitude bins with low scatter $\sigma_{\Delta(V-I)} < 0.2$, and a more conservative ('faint') colour cut for magnitude bins with slightly larger scatter $0.2 < \sigma_{\Delta(V-I)} < 0.3$, see columns 5 and 6 in Table 2 for the corresponding magnitude bins in each cluster and Table D1 for the colour cuts as a function of cluster redshift.

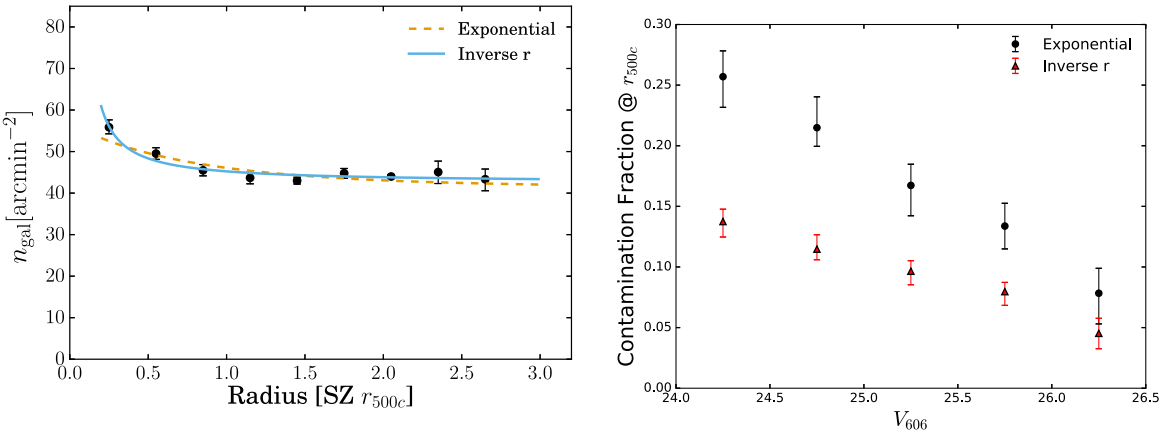

Figure E1. Cluster member contamination for the analysis centred around the X-ray centroid when no colour selection is applied. *Left:* Number density profile combining all magnitude bins, where the curves indicate the best-fitting exponential and $1/r$ (SIS) model for the cluster member contamination. *Right:* Comparison of the estimated contamination fraction f_{500} for the two models as a function of magnitude.

Secondly, we add noise to the $V_{606} - I_{814}$ colour estimates in the CANDELS catalogue prior to the colour cut, similarly to our approach for the ACS-only selection. However, in contrast to Appendix D3.1 we do not assume a Gaussian noise distribution here, but randomly sample the noise from the actual distribution of the colour differences $(V_{606,\rm con} - I_{\rm FORS2})_{\rm fix} - (V_{606} - I_{814})$ shown in the right-hand panels of Fig. D1. The motivation for not using a Gaussian approximation is given by the skewness in the distribution and presence of outliers. In practice, we again divide the galaxies into 0.5 mag-wide bins in V_{606}. We further subdivide these galaxies into sub-bins according to their $V_{606} - I_{814}$ colour if sufficiently many galaxies are available to provide sub-bins containing at least 30 galaxies each. For each galaxy in the CANDELS catalogue we then identify the corresponding bin/sub-bin and randomly assign a colour difference drawn from this bin/sub-bin. Note that we introduce the further colour subdivision as red galaxies (which are later removed by the colour cut) show a lower scatter at a given V_{606} magnitude.[17]

APPENDIX E: LIMITATIONS OF A STATISTICAL CORRECTION FOR CLUSTER MEMBER CONTAMINATION

Weak lensing studies that use wide-field imaging data and do not have sufficient colour information for a robust removal of cluster galaxies can attempt to statistically correct their shear profiles for the dilution effect of cluster members in the source samples (see e.g. Hoekstra et al. 2015). For this, they need to estimate the relative excess counts as a function of cluster-centric distance, ideally accounting for the impact of masks, obscuration by cluster members and magnification, and fit it with a model, typically in the form

$$n_{\rm measure}(r) = \frac{n_{\rm bg}}{1 - f(r)}, \tag{E1}$$

and scale the shear profile as

$$\langle g_{\rm t}\rangle^{\rm boosted}(r) = \langle g_{\rm t}\rangle(r)\frac{1}{1 - f(r)}. \tag{E2}$$

[17] This is expected since $(V_{606,\rm con} - I_{\rm FORS2})$ receives roughly comparable scatter contributions from $V_{606,\rm con}$ and $I_{\rm FORS2}$, with a reduced scatter in $I_{\rm FORS2}$ for red objects.

Here we consider two previously employed models for the projected density profiles of cluster galaxies, namely the projected singular isothermal sphere (SIS) model

$$f(r) = f_{500}\frac{r_{500c}}{r} \tag{E3}$$

(e.g. Hoekstra 2007) and an exponential model

$$f(r) = f_{500}{\rm e}^{1-r/r_{500c}} \tag{E4}$$

(e.g. Applegate et al. 2014), where f_{500} corresponds to the contamination at r_{500c}.

We do not use this approach for our *HST* analysis as the 2×2 ACS mosaics are too small to derive a robust estimate of the background source density directly. To test this, we use our source catalogues without colour selection, estimate the mask- and obscuration-corrected source density profiles in magnitude bins, and fit them with both $f(r)$ profiles. Combining the analysis from all clusters we find that both profiles provide acceptable fits for most of the radial range covered by the ACS data. For example, when using only a single broad magnitude bin, the SIS model returns $\chi^2 = 6.0$ for 7 degrees of freedom, whereas the exponential model returns $\chi^2 = 14.3$. The SIS model is clearly a better fit at small radii (see the *left*-hand panel of Fig. E1), but the exponential profile is not ruled out at high significance. Yet, the two models yield uncomfortably different contamination fractions (shown in the *right*-hand panel of Fig. E1 as a function of V_{606}). As a test for the impact of these differences we artificially apply the two different boost correction schemes (taking their magnitude dependencies into account) to our colour-selected shear profiles and compare the resulting mass estimates. Here we find that the exponential model leads to mass estimates which are higher compared to those from the isothermal model by ∼14 per cent. As it is currently not clear what the correct functional form would be, we conclude that the application of such a contamination correction would introduce substantial systematic uncertainty.

One could consider to reduce this uncertainty by using external blank fields to constrain the background source density. Using our colour-selected catalogues we have demonstrated that the careful matching of source selection criteria and noise properties between the cluster and reference fields, which would be required for such an approach, is in principle possible. However, instead of providing an important validation test as in our study, the information in the number density would then be used to correct the signal, assuming

that all other related analyses steps were done correctly. Large-scale structure variations also introduce significant variations in the source densities between the five CANDELS fields. Without colour selection we find that they lead to uncertainties in the estimated mean background density of ∼6 per cent at $V \sim 24$ to ∼3 per cent at $V \sim 26$.

In addition to the increased systematic uncertainty, the use of a contamination correction also increases the statistical uncertainty compared to a robust colour selection that adequately removes cluster galaxies. First, the cluster members dilute the small-scale signal, which is the regime providing the highest signal-to-noise contribution for our analysis. Secondly, source density profiles are typically too noisy to measure the contamination for individual clusters. On the other hand, if an average contamination model is applied, extra scatter in the mass constraints is introduced.

APPENDIX F: IMPACT OF CONTAMINATION BY VERY BLUE CLUSTER MEMBERS

The tests presented in Section 6.8 show no indication for a significant residual contamination by cluster members. However, our estimates from Section 6.4 suggest that, in the presence of noise and averaged over our cluster sample, our ACS-only (ACS+FORS2) colour selection should leave a residual contamination of ∼1.9 per cent (∼1.1 per cent) of very blue field galaxies at the corresponding cluster redshifts. Whether or not this can introduce a residual excess contamination by cluster members depends on the relative properties of the galaxy distributions in the field and cluster environment.

Luminous Compact Blue Galaxies (LCBGs, e.g. Koo et al. 1994) represent an extreme star-bursting population of galaxies with very blue colours and compact sizes. Such galaxies were also identified in cluster environments (Koo et al. 1997), making them the most relevant potential contaminant for our colour-selected weak lensing source sample. Crawford et al. (2011), Crawford, Wirth & Bershady (2014) and Crawford et al. (2016) identify and study LCBGs in five massive clusters at $0.5 < z < 0.9$ using a photometric preselection, Keck/DEIMOS spectroscopy, and *HST* morphological measurements. For the $z > 0.6$ clusters in their sample Crawford et al. (2011) find that the number density enhancement of the cluster LCBG population compared to the LCBG field density is comparable to or lower than the corresponding enhancement of the total cluster population compared to the total field population. In addition, Crawford et al. (2016) find that the relevant properties of the cluster LCBGs (star formation rate, dynamical mass, size, luminosity and metallicity) are indistinguishable from the properties of field LCBGs at the same redshift.

Accordingly, we can make the conservative assumption that the relative fraction of cluster members that pass our colour selection is equal to or lower than the fraction of field galaxies passing the selection $f_{\rm pass,field} \sim 1.9$ per cent (1.1 per cent) for the ACS-only (ACS+FORS2) selection, accounting for noise (see Section 6.4). We then estimate the approximately expected average fraction of cluster galaxies in our colour-selected source sample at r_{500c} as

$$f_{500,\rm expected} = f_{\rm pass,field} f_{500,\rm no-cc} \left(\frac{n_{\rm gal,cc}}{n_{\rm gal,no-cc}}\right)^{-1} = 0.009\,(0.008), \quad \text{(F1)}$$

where $f_{500,\rm no-cc} \simeq 0.15$ indicates an estimate for the average contamination at r_{500c} based on a number density profile analysis when the colour selection is *not* applied (see the right-hand panel of Fig. E1, averaging the values for the more conservative exponential model according to the relative weight of the corresponding magnitude bin in the reduced shear profile fits), and $n_{\rm gal,cc}/n_{\rm gal,no-cc} = 0.33$ (0.22) corresponds to the fraction of galaxies in the cluster fields passing the colour selection within the magnitude range of the ACS-only (ACS+FORS2) analysis. We do not attempt to model the radial distribution of the expected contaminating cluster galaxies, as LCBGs appear to follow a rather shell-like distribution with a depletion in the cluster core (Crawford et al. 2006). Instead, we assume that $f_{500,\rm expected}$ provides a reasonable approximation for the typical contamination, which is likely conservative given that the average $\langle r_{500c}\rangle = 770$ kpc of our cluster sample (based on the lensing analysis and assuming the concentration–mass relation from Diemer & Kravtsov 2015, see Section 7) is more representative for the inner (500 kpc) than the outer (1.5 Mpc) limit of the fit range for our default analysis. With these conservative assumptions, the relative bias for the average lensing efficiency caused by the expected cluster contamination is $\Delta\langle\beta\rangle/\langle\beta\rangle = -f_{500,\rm expected} = -0.009$ (−0.008). Given that this is even smaller than the uncertainty on $\langle\beta\rangle$ from line-of-sight variations between the CANDELS fields (Section 6.6), this bias could well be ignored (we still include it in the systematic error budget in Table 8). But we note that future studies could attempt to model the contamination more accurately and apply a correction.

APPENDIX G: ADDITIONAL FIGURES

Figs G1 to G12 complement Figs 14 and 16, showing the corresponding results for the other clusters. In particular, the left-hand

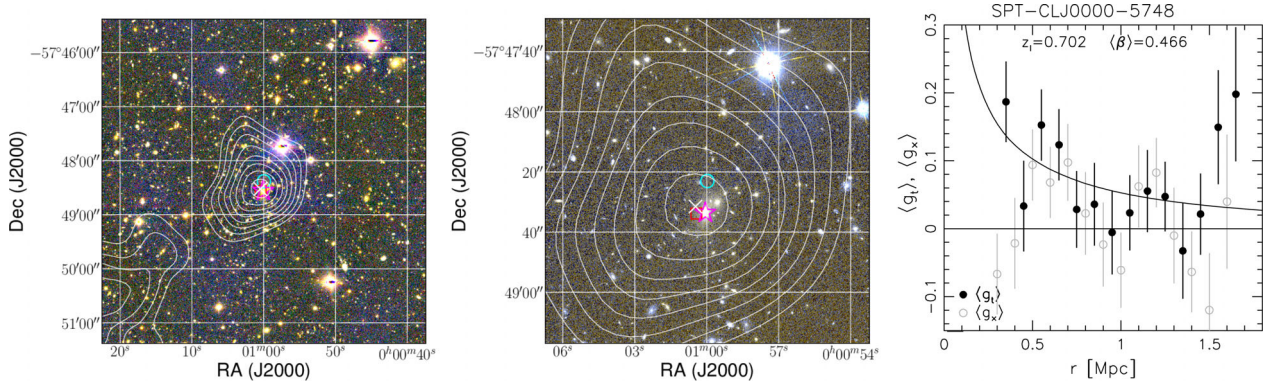

Figure G1. Weak lensing results for SPT-CL J0000−5748. See the descriptions in the captions of Figs 14 and 16 for details.

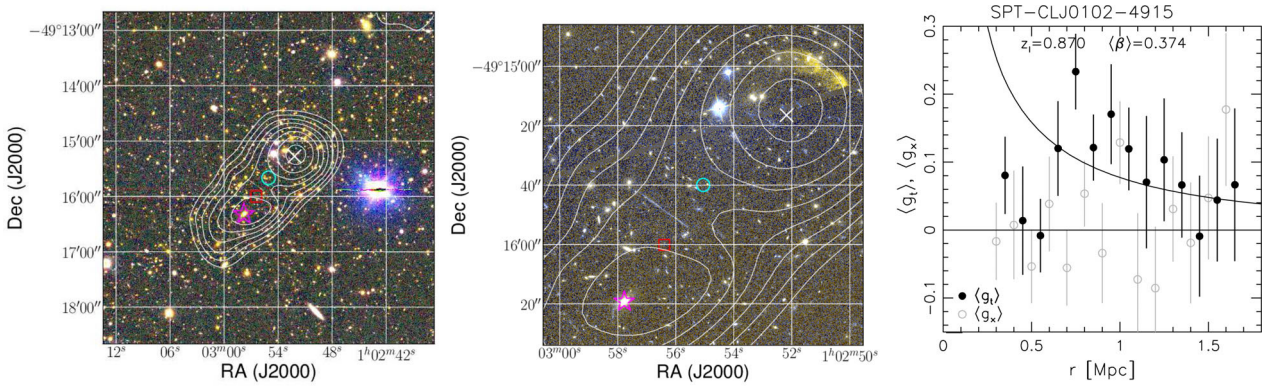

Figure G2. Weak lensing results for SPT-CL J0102−4915. See the descriptions in the captions of Figs 14 and 16 for details.

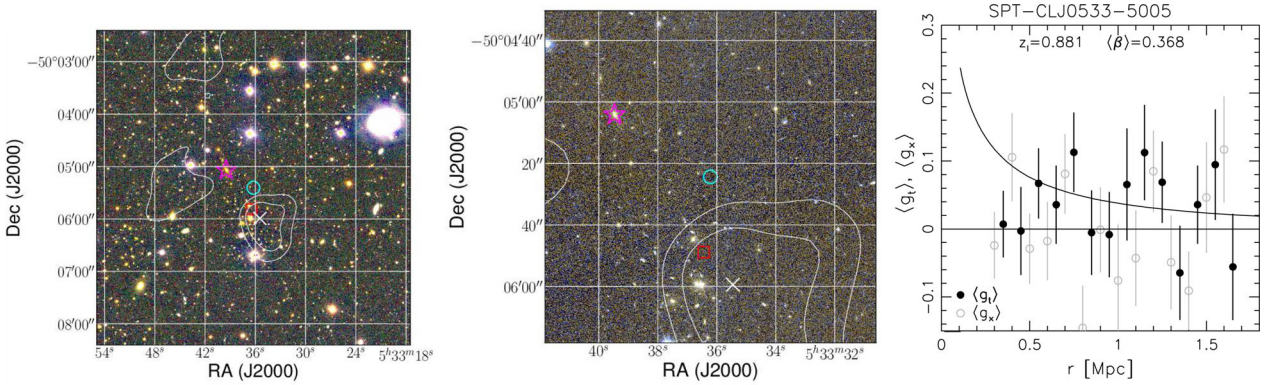

Figure G3. Weak lensing results for SPT-CL J0533−5005. See the descriptions in the captions of Figs 14 and 16 for details.

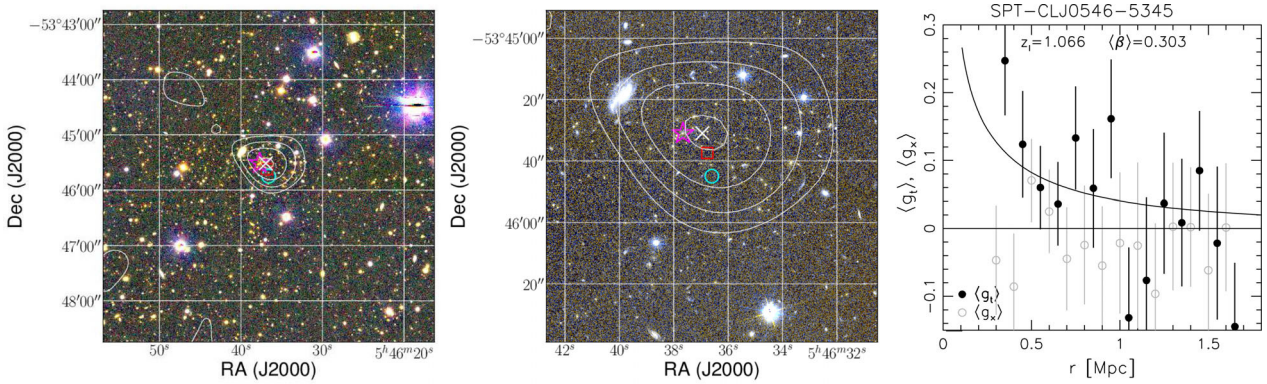

Figure G4. Weak lensing results for SPT-CL J0546−5345. See the descriptions in the captions of Figs 14 and 16 for details.

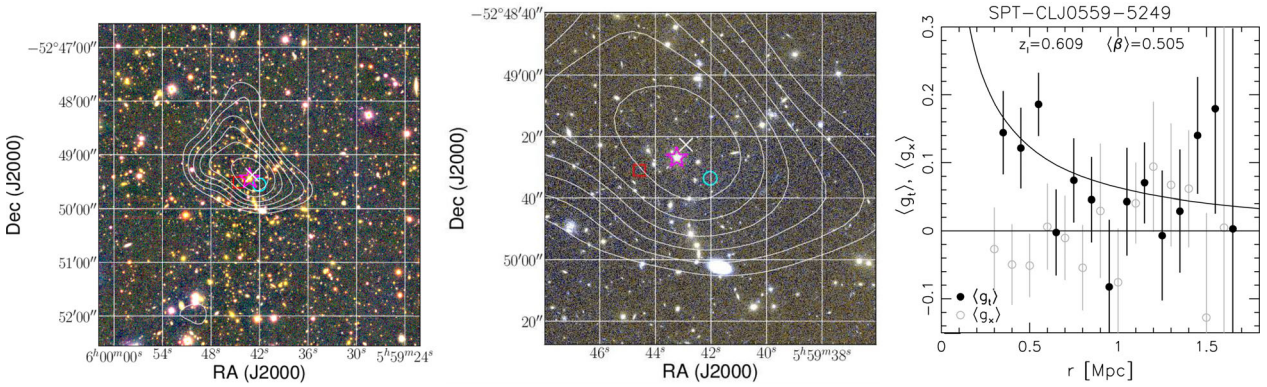

Figure G5. Weak lensing results for SPT-CL J0559−5249. See the descriptions in the captions of Figs 14 and 16 for details.

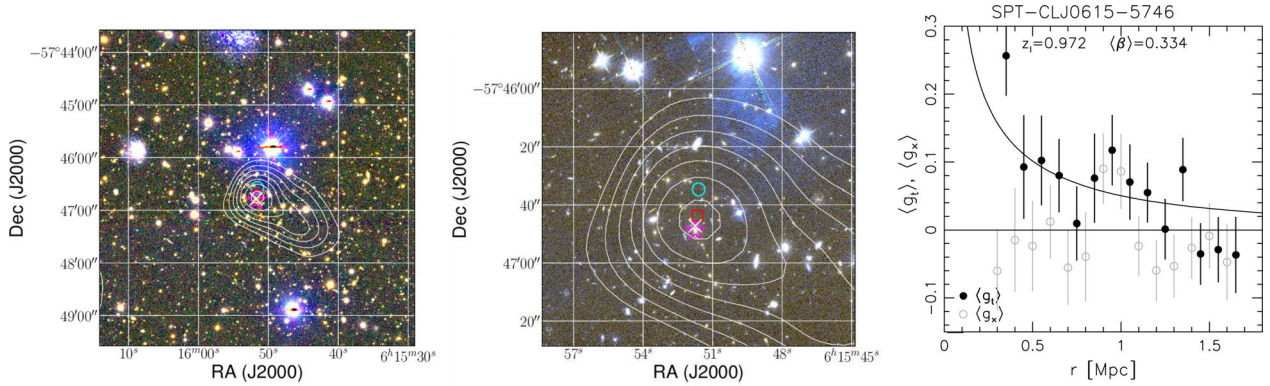

Figure G6. Weak lensing results for SPT-CL J0615−5746. See the descriptions in the captions of Figs 14 and 16 for details.

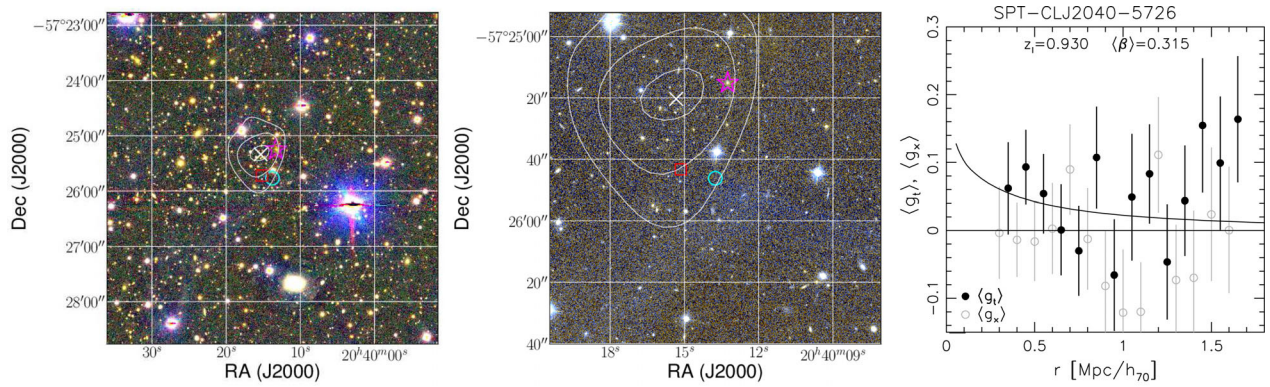

Figure G7. Weak lensing results for SPT-CL J2040−5725. See the descriptions in the captions of Figs 14 and 16 for details.

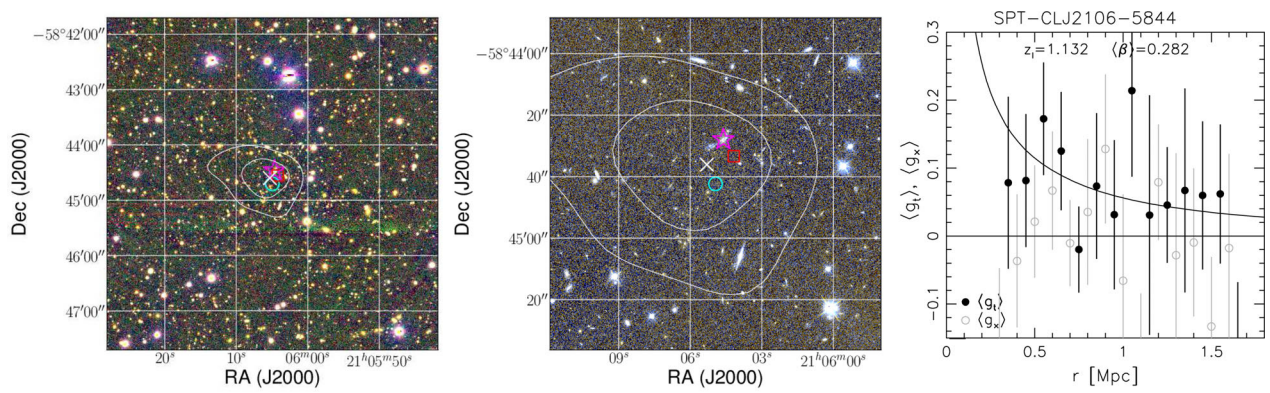

Figure G8. Weak lensing results for SPT-CL J2106−5844. See the descriptions in the captions of Figs 14 and 16 for details.

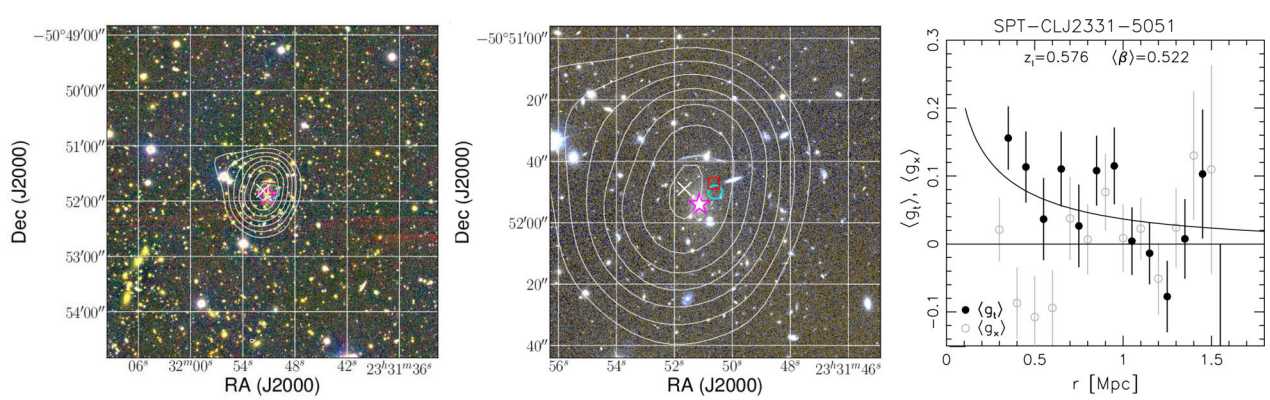

Figure G9. Weak lensing results for SPT-CL J2331−5051. See the descriptions in the captions of Figs 14 and 16 for details.

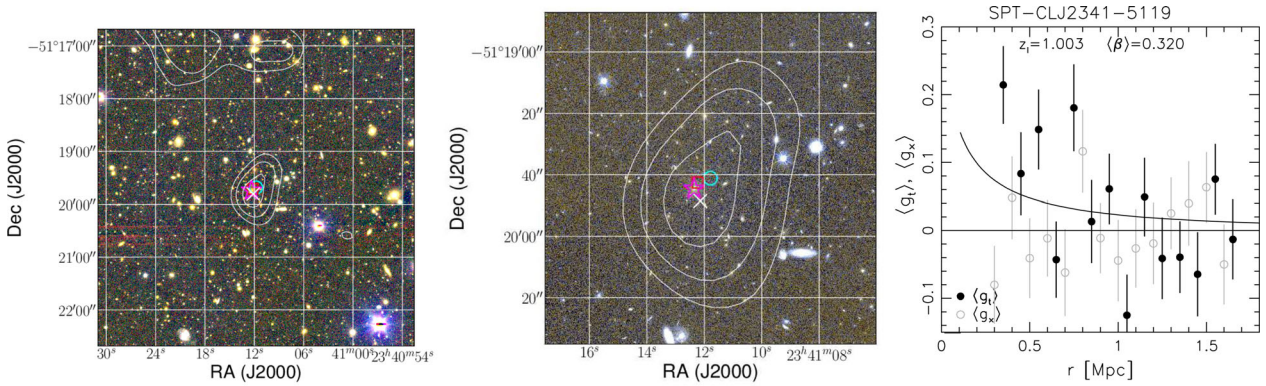

Figure G10. Weak lensing results for SPT-CL J2341−5119. See the descriptions in the captions of Figs 14 and 16 for details.

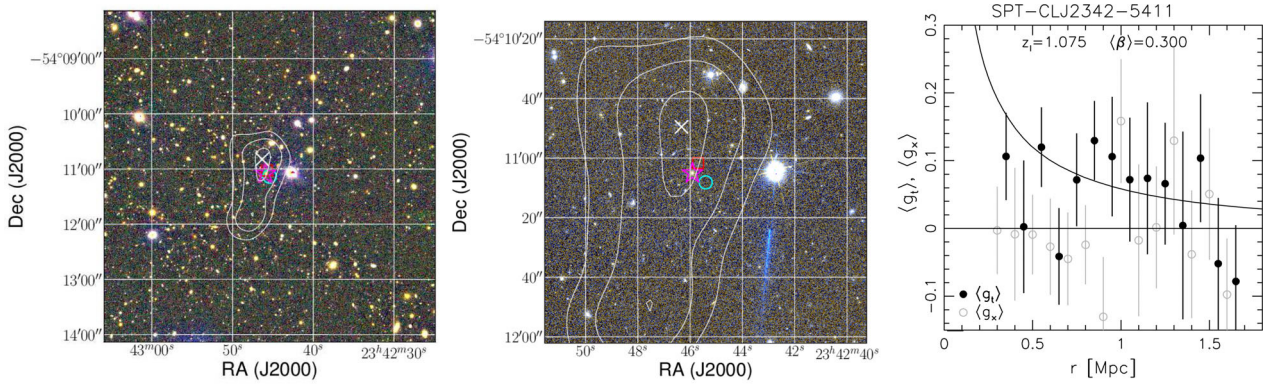

Figure G11. Weak lensing results for SPT-CL J2342−5411. See the descriptions in the captions of Figs 14 and 16 for details.

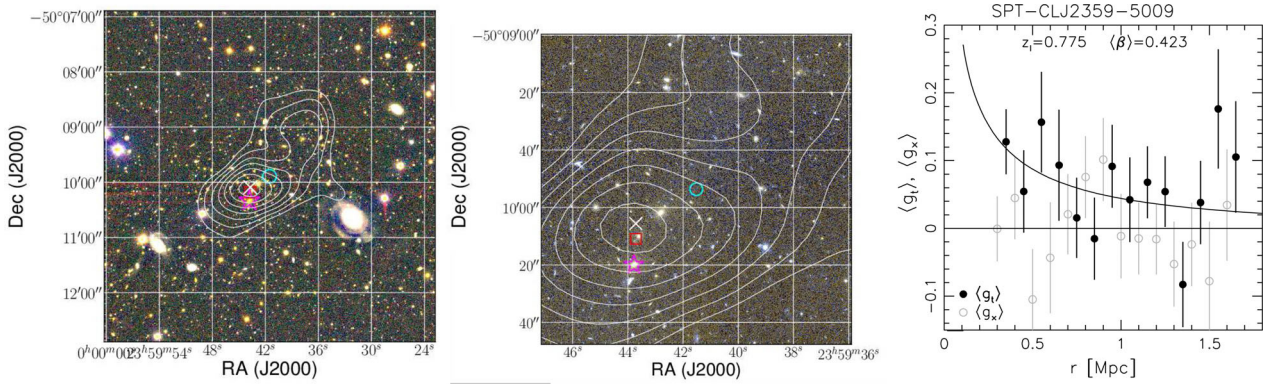

Figure G12. Weak lensing results for SPT-CL J2359−5009. See the descriptions in the captions of Figs 14 and 16 for details.

and middle panels show the weak lensing S/N mass reconstructions overlaid on to the corresponding VLT/FORS2 *BIz* and central ACS colour images, as well as the locations of the different cluster centres used in our analysis. In the corresponding right-hand panels we show the weak lensing shear profiles centred on to the X-ray centroids.

[1]*Argelander-Institut für Astronomie, Universität Bonn, Auf dem Hügel 71, D-53121 Bonn, Germany*
[2]*Kavli Institute for Particle Astrophysics and Cosmology, Stanford University, 382 Via Pueblo Mall, Stanford, CA 94305-4060, USA*
[3]*Department of Physics, Stanford University, 382 Via Pueblo Mall, Stanford, CA 94305-4060, USA*
[4]*Kavli Institute for Cosmological Physics, University of Chicago, 5640 South Ellis Avenue, Chicago, IL 60637, USA*
[5]*Faculty of Physics, Ludwig-Maximilians University, Scheinerstr 1, D-81679 München, Germany*
[6]*Excellence Cluster Universe, Boltzmannstr 2, D-85748 Garching, Germany*
[7]*Leiden Observatory, Leiden University, Niels Bohrweg 2, NL-2300 CA Leiden, the Netherlands*
[8]*Argonne National Laboratory, 9700 S. Cass Avenue, Argonne, IL 60439, USA*
[9]*Department of Astronomy, University of Florida, Gainesville, FL 3261, USA*
[10]*Dark Cosmology Centre, Niels Bohr Institute, University of Copenhagen, Juliane Maries Vej 30, DK-2100 Copenhagen, Denmark*

[11]*Department of Physics and Astronomy, Stony Brook University, Stony Brook, NY 11794, USA*
[12]*MIT Kavli Institute for Astrophysics and Space Research, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA*
[13]*Department of Astronomy, University of Washington, Box 351580, Seattle, WA 98195, USA*
[14]*SLAC National Accelerator Laboratory, 2575 Sand Hill Road, Menlo Park, CA 94025, USA*
[15]*Department of Physics, Harvard University, 17 Oxford Street, Cambridge, MA 02138, USA*
[16]*Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138, USA*
[17]*Department of Physics & Astronomy, Colby College, 5800 Mayflower Hill, Waterville, ME 04901, USA*
[18]*Fermi National Accelerator Laboratory, Batavia, IL 60510-0500, USA*
[19]*Department of Astronomy and Astrophysics, University of Chicago, 5640 South Ellis Avenue, Chicago, IL 60637, USA*
[20]*Department of Physics, University of Chicago, 5640 South Ellis Avenue, Chicago, IL 60637, USA*
[21]*Academia Sinica Institute of Astronomy and Astrophysics (ASIAA), 11F of AS/NTU Astronomy-Mathematics Building, No. 1, Section 4, Roosevelt Rd, Taipei 10617, Taiwan*
[22]*Department of Physics, IIT Hyderabad, Kandi, Telangana 502285, India*
[23]*Department of Astronomy and Astrophysics, University of California, Santa Cruz, CA 95064, USA*
[24]*Department of Physics, McGill University, 3600 Rue University, Montreal, Quebec H3A 2T8, Canada*
[25]*Department of Physics, University of California, Berkeley, CA 94720, USA*
[26]*Institute for Computational Cosmology, Durham University, South Road, Durham DH1 3LE, UK*
[27]*Max Planck Institute for Extraterrestrial Physics, Giessenbachstrasse 1, D-85748 Garching, Germany*
[28]*School of Physics, University of Melbourne, Parkville, VIC 3010, Australia*
[29]*Cerro Tololo Inter-American Observatory, Casilla 603, La Serena, Chile*

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