Celestial Dragons and Brushing Teeth

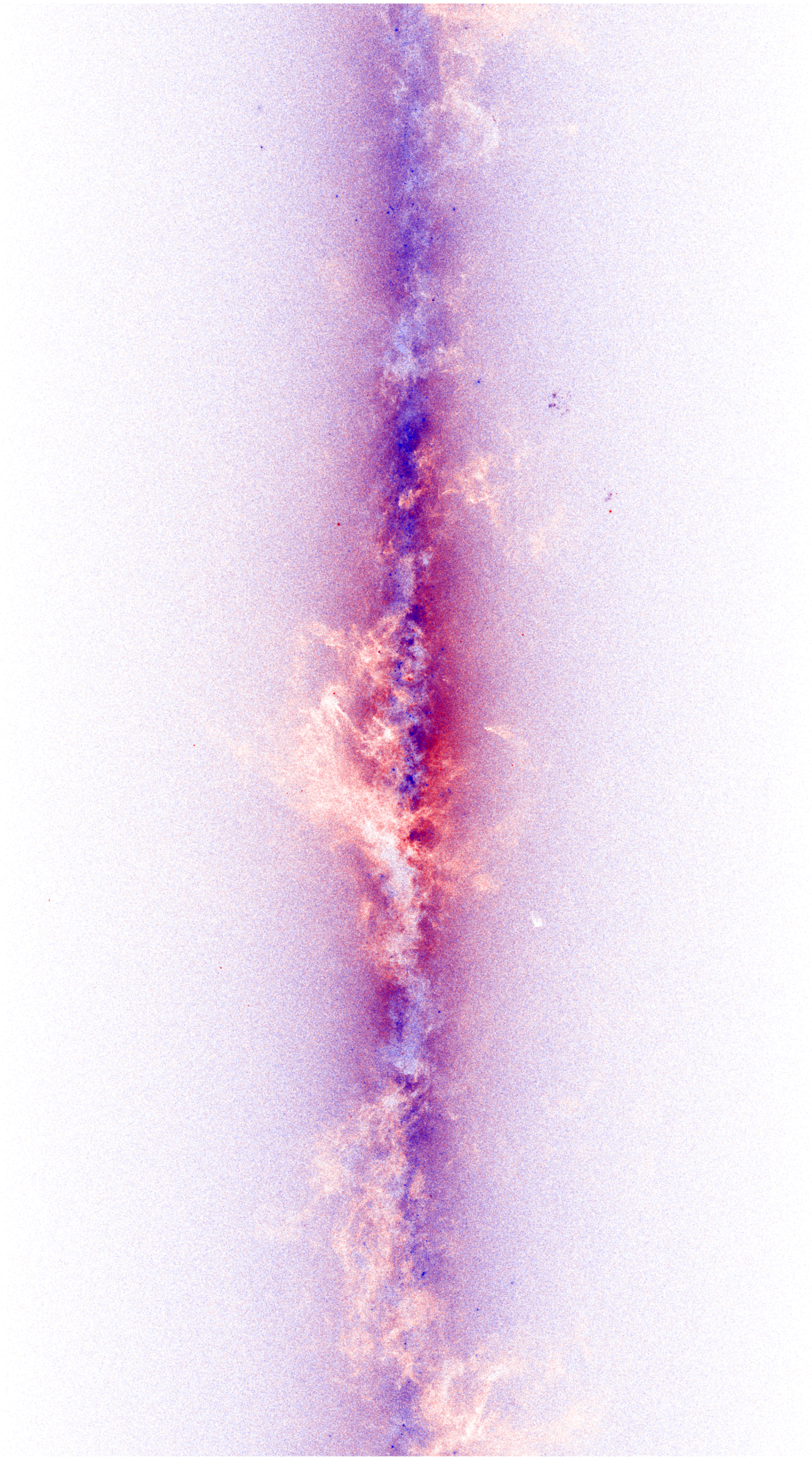
by

William K. Black

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Physics)
in the University of Michigan
2023

Doctoral Committee:

Professor August Evrard, Chair
Doctor Becky Matz
Professor Christopher John Miller
Professor Marcelle Soares-Santos



William K. Black

wkblack@umich.edu

ORCID iD: 0000-0003-4811-7913

DEDICATION

To my wise wife Eden and my stoic son Fletcher.

ACKNOWLEDGEMENTS

I am grateful for my advisor, Professor Gus Evrard. Not only has he helped me expand my horizons to include physics education research, but I've had a wonderful time doing it. Not everyone can say that they get along with their advisor well (and horror stories abound), but I am truly grateful to have met just the right advisor for me. Having a long leash and a wide field in which I could run gave me liberty to explore many avenues heretofore untouched by human hands.

I am grateful for the mentorship and friendship of other faculty at University of Michigan; even if it was only for a single course, Eric Bell, Dragan Huterer, and Jim Liu all helped stoke the flames of curiosity and inspired me to greater heights.

I am grateful for Johnny Esteves for his enthusiasm, friendship, and for the stellar name of "Red Dragon" for my evolving GMM algorithm. I am grateful for Peter Melchior for his open-source algorithm pyGMMis (which served as Red Dragon's backbone) and for his assistance in fitting and interpreting results.

I am grateful for my wife Eden and son Fletcher, for the joy, motivation, and assistance they have given me throughout my time at Michigan.

Above all, I am grateful to my God, for saving me and leading me forward.

PREFACE

When asked what superpower I would like, I consistently answered teleportation or time control. I had always seen myself as an inventor, but I realized this wasn't the most stable of jobs, so at the start of high school I shifted towards engineering. This led me to my first physics course.

I had always loved the puzzle-solving of mathematics and the practical application and discovery of science, but it was first in physics where I really saw and appreciated the union of those two. What really hooked me was the realization that you could predict the future trajectory of a ball using the kinematic equations. From my first physics course and on, I feasted on the topic, hungry to discover as much about the universe as possible at its most fundamental level.

After reading Michio Kaku's *Physics of the Impossible*, I saw that teleportation and time travel were permissible in Einstein's equations (albeit with exotic matter), so I set my focus on the study of gravity. Initially, I took interest in black holes with my undergraduate advisor Dr. David Neilsen. This led to a summer research position at Los Alamos, where I studied primordial black holes in cosmological simulations with my advisor Dr. Joe Smidt. I was amazed that the gravitational theory of cosmology underpinned the entire universe! I saw that Dr. August Evrard was a first-generation computational cosmologist, so I applied to University of Michigan and was accepted. I sat in on a cosmology class of his on one of my first days there, and on that day, he offered a research position.

At first, my task was to look at scaling relations between galaxy clusters' dark matter content and photon emission. As I set out to select the population of red, quiescent galaxies (as compared to blue, star-forming galaxies), I quickly encountered a niche which (in my opinion) needed to be filled: there was no published, redshift-evolving model of both red and blue populations (besides hard cut characterizations, which had several downsides, of chief concern to me that they ignore several photometric bands and uncertainties in measurement). So "see a need, fill a need", and I started working on a Gaussian mixture model of the two populations, which would eventually lead to my Red Dragon algorithm.

Funding was sparse for this effort though, so I often worked as a graduate student instructor (GSI) or was covered by other grants. Over the summers of 2021 and 2022, I got to moonlight as a learning analytics researcher, studying grade gains due to study on the online service *Problem Roulette*. The service was initially designed to serve physics problems from past exams to students in order to help them study. This research directly connected to my drive to be a good teacher, as I could now give quantitative suggestions to students regarding study habits.

This dissertation therefore has two main aspects: the cosmology-inspired astrophysics of galaxy clusters as well as the learning analytics work on quantifying study gains. While seemingly disjoint, I happily found that several tools and perspectives in physics were applicable in the world of learning analytics. I look forward to continually improving myself as a professor, be it in teaching or research.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

LIST OF APPENDICES

LIST OF ACRONYMS

RS Red Sequence

BC Blue Cloud

GV Green Valley

DM Dark Matter

RM redMaPPer

KLLR Kernel-Localized Linear Regression

DBER Discipline-Based Education Research

PER Physics Education Research

EDM Educational Data Mining

PR Problem Roulette

ABSTRACT

This dissertation delves into two domains. The first domain is cosmology and astrophysics, investigating galaxy occupation of dark matter halos as well as characterization of the star-forming and quiescent galaxy populations in photometric color space. The second domain covered is physics education research, investigating student grade gains due to practice study.

Chapter 2 measures how galaxies occupy the dark matter halos of the Buzzard synthetic galaxy catalog, finding several divergences from observations. In characterizing halo richness as a function of mass, we find a significantly shallow slope and a decreasing scatter with increasing mass. We then characterize the fraction of red sequence (RS) galaxies in halos, investigating its dependence on richness, mass, redshift, and radius. Surprisingly, we find that red fraction follows an inverted mass dependence—of *decreasing* with increased halo mass; this leads to high-mass halos lacking RS galaxies. We also find that—again, contrary to expectations—red fraction *increases* with redshift. We also investigate radial density profiles, and find a surprising deficit of galaxies near one-tenth the virial radius.

Chapter 3 formally introduces the Red Dragon algorithm, detailing its selection of galaxies in Buzzard. Red Dragon is a multivariate error-corrected Gaussian Mixture Model (GMM). This results in smooth characterizations of both RS as well as blue cloud (BC) galaxies across time, thereby avoiding discontinuities inherent in swapping RS selection colors. We find a quiescent (low specific star formation rate) selection accuracy of over ninety percent, indicating a successful characterization of the two populations.

Chapter 4 uses Red Dragon to characterize the RS and BC in the COSMOS2015 dataset. This deep field allows for high-precision redshift estimates, which give strong measures of population characteristics (out to $z \sim 1.5$). We again find high selection accuracy ($\gtrsim 90\%$) of the quiescent population. With this characterization, we measure temporal evolution of median galactic specific star formation rate as well as galactic age.

Chapter 5 then turns to the domain of physics education research. We first condition final course grade on mean ACT/SAT math score T, subtracting out the expected grade, in order to account for incoming math proclivity. Using data from the online study service Problem Roulette (PR), we measure grade gains due to practice study, finding that maximum gains were roughly $.77 \pm .12$ grade points, moving from no study to studying 1000 PR questions over the term (roughly ten per day).

These gains persisted at any fixed math score. After modeling expected grade using T and study volume, we then investigate divergences from this expectation for various demographic groups. We find that students whose parents did not earn a college degree earned .27 ± .04 grade points below expectations (6.1σ significant) and find that underrepresented minority students earned .14 ± .04 grade points below expectations (3.6σ significant). All divergences from expectations subceeded maximal grade gains of study. Residual scatter remains comparable to maximal study gains, implying that the model is far from deterministic: individual variation trumps mean trends. Our findings can help motivate student study habits and help teachers identify which students may especially need such encouragement.

CHAPTER 1

Introduction

This dissertation includes two overarching topics: cosmology & astrophysics research (introduced in §1.1; featured in Chapters 2, 3, & 4) and physics education research (introduced in §1.2; featured in Chapter 5). With the first topic, I investigate the clustering and coloring of star-forming and quiescent galaxies; with the second topic, I quantify learning gains due to practice study in introductory physics classes. These two topics align with my passion for research and teaching, as I explore novel astrophysics and implement learning analytics to nurture students' knowledge and passion for the same celestial phenomena.

A full review of cosmology is not necessary for interpreting most of the results in this thesis, so we abstain from a full review. In brief, we have found that the universe recently (astronomically speaking; roughly four billion years ago) stopped decelerating and began accelerating its expansion [Frieman et al., 2008]. As will be discussed shortly in §1.1, we find that the universe is well explained by general relativity (Einstein's model of gravity), where the universe is chiefly composed of some unknown substance dubbed "dark energy" (denoted Λ) which causes the accelerated expansion of the universe, followed by "cold dark matter" (CDM). Only about 5% of the observed matter-energy of our universe is composed of the normal baryonic matter we interact with daily, and a very small remaining fraction of the universe's matter-energy is composed of radiation (light). The leading model of our universe is hence known as the ΛCDM model of general relativity [Costanzi et al., 2019, Aghanim et al., 2020, Huterer, 2023].

1.1 Galaxy Cluster Cosmology

The large-scale structure of the universe bears remarkable resemblance to the human brain [Vazza and Feletti, 2020]. For both structures, $\sim$ 75% of the mass-energy distribution is made of a passive material (dark energy; water) while the remaining $\sim$ 25% makes up the physical structures (matter; cells), which organize their geometry similarly. Both structures resemble webs, with dense nodes (dark matter halos; soma) connected by thinner tendrils (dark matter filaments; axon–dendrite connections). Dark Matter (DM) halos serve as nests for galaxy formation, with the largest DM halos yielding the largest galaxies [Springel et al., 2005]. Though galaxies are occasionally found in the low-density voids between clusters (still having been born within a DM halo), they are more likely found along DM walls or filaments and are most often found in the cores of DM halos [Shandarin and Zeldovich, 1989], forming a cluster of co-orbiting galaxies (like the nuclei of neurons).

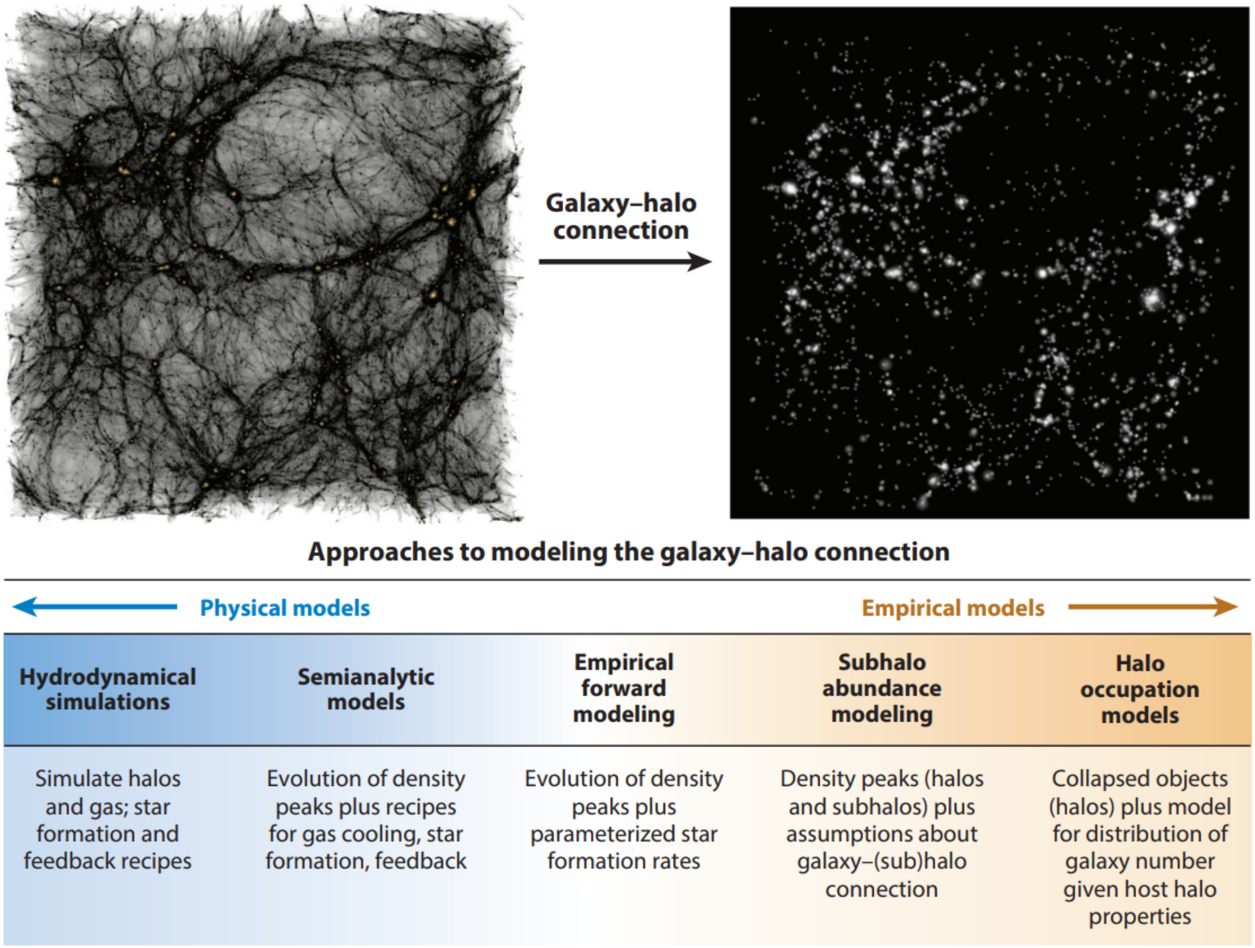
The most massive galaxy clusters form in regions which had larger local densities than average at the universe's beginning [Kaiser, 1984, Springel et al., 2005]. Gravity attracts local objects, pulling dark matter, dust, and galaxies all into walls, then filaments, then halos. These halos serve as nests for galaxy formation, growth, and hierarchical merging. This makes galaxy clusters the most massive gravitationally bound systems in the universe, as both the birth place of massive galaxies and the end destination for gravitational collapse. This "cosmic web" has been characterized both on our actual sky as well as in simulations [Springel et al., 2005, see Figure 1.2].

1.1.1 Simulating galaxies

Two main types of methods simulate galaxies in the cosmic web: physical models and empirical models (see Fig 1.1, reproduced from Wechsler and Tinker [2018]). The core tradeoff between the two is of depth versus breadth: how well resolved and detailed are your galaxies versus how large of a simulation volume can you run?

Analytical models (also called "physical models"), such as the hydrodynamics code Arepo [Springel, 2010, Nelson et al., 2018], simulate an entire universe from scratch, start to end. This includes not only gravitational attraction and expansion of the universe as are included in DM-only codes, but additionally include effects of baryons. In order to make realistic galaxies, many baryonic elements must be accounted for, including modeling gas, stars (and therefore their evolution, creation of metals, and feedback from supernovae explosions), black holes (and their feedback into the environment), galactic winds, and magnetohydrodynamics [Weinberger et al., 2017, Pillepich et al., 2018]. After feeding in these models of astrophysical phenomena, they let the simulation loose on some initial conditions, creating a universe true to those laws and models. Differences between such simulations and reality thus reflect either our lack of understanding of physical laws & processes, or a lack of computational resources. For example, to run cosmological-scale hydrodynamic simulations in a reasonable amount of time—say, in less than a year—we must not only forego modeling the fusion & gravitation & magnetohydrodynamics of *individual* stars, but we must treat stars in aggregate, with e.g. one million stars as a single simulation particle. Such "star particles" must therefore be modeled as a whole, giving net effects of the aggregate.

On the other extreme are **empirical models**, such as *galaxy pasting* methods like AddGals



Approaches to modeling the galaxy–halo connection

← Physical models | Empirical models →

Hydrodynamical simulations	Semianalytic models	Empirical forward modeling	Subhalo abundance modeling	Halo occupation models
Simulate halos and gas; star formation and feedback recipes	Evolution of density peaks plus recipes for gas cooling, star formation, feedback	Evolution of density peaks plus parameterized star formation rates	Density peaks (halos and subhalos) plus assumptions about galaxy–(sub)halo connection	Collapsed objects (halos) plus model for distribution of galaxy number given host halo properties

Figure 1.1: Description of the various models of populating DM halos with galaxies, reproduced from Wechsler and Tinker [2018], Figure 1. Original caption: “Modeling approaches to the galaxy–halo connection. Top panel shows the dark matter distribution in a $90 \times 90 \times 30$ Mpc$/h$ slice of a cosmological simulation (*left*) compared with the galaxy distribution using an abundance matching model (*right*), tuned to match galaxy clustering properties of an observed sample. The grid highlights the key assumptions of various models for the galaxy–halo connection. The models are listed on a continuum from left to right ranging from more physical and predictive (making more assumptions from direct simulation or physical prescriptions) to more empirical (more flexible parameterizations, constrained directly from data).”

[Busha and Wechsler, 2008, Wechsler et al., 2021], which e.g. take a DM-only simulation ($\gtrsim 100$ times cheaper and faster to run than a hydro sim) and assign galaxies to dark matter 'particles' in such a way that matches observations. In contrast to analytical models, which set initial conditions and evolution according to understood or assumed principles and observations, empirical models have the opposite approach, of constructing the end result such that it matches reality closely.

Both types of simulated universes have their own unique use in analysis: while the former is considered more 'physical' (offering predictive power based on assumed physical prescriptions), the latter costs far less. For example, while state of the art high-resolution hydrodynamics simulations are just now reaching sizes of a 300 Mpc cube (see TNG-Cluster, to be released in 2023), dark matter only simulations (which can have baryons pasted onto dark matter halos to empirically match observations) have already reached high-resolution boxes one thousand times larger [e.g. the Uchuu simulated universe has a width of 3 Gpc; see Ishiyama et al., 2021]! This puts empirical models far closer to cosmological scales than analytic models. Because we lack a perfect understanding of the universe, both analytical and empirical models diverge somewhat from reality; yet both models match reality remarkably well [as illustrated in Figure 1.2, reproduced from Springel et al., 2006], to the point that their spatial distribution of galaxies is visually indistinguishable from reality.

1.1.2 Constraining cosmological parameters with galaxy clusters

Galaxy cluster cosmologists particularly interest themselves in the *mass* of galaxy clusters as well as their *distribution* in space. The comoving number density of halos dN in a given redshift shell dz and mass bin $d\ln M$ is

$$\frac{d^2N}{d\ln M dz} = \frac{\Omega}{4\pi}\left(\frac{dV}{dz}\right)\left(\frac{dn(M,z)}{d\ln M}\right), \tag{1.1}$$

with $\Omega/4\pi$ as the fraction of sky observed; the volume element dV/dz and the halo mass function $dn(M,z)/d\ln M$ both depend on cosmology [Cooray and Sheth, 2002]. Thus, by measuring the left hand side of the equation by counting and estimating mass of galaxy clusters, one can thereby obtain constraints on cosmology. The volume element

$$dV/dz = 4\pi r(z)^2 \frac{c}{H(z)} \tag{1.2}$$

directly depends on the universe's cosmology, vis. through the redshift-evolving Hubble parameter $H(z)$, which measures the expansion of space, and metric distance $r(z)$. While the mass function connects less straightforwardly to cosmology, analytic and emulated cosmological connections exist [Press and Schechter, 1974, Murray et al., 2013]. In brief, it is directly proportional to the fraction of the universe's energy composed of matter Ω_m and dependent on the variance of matter

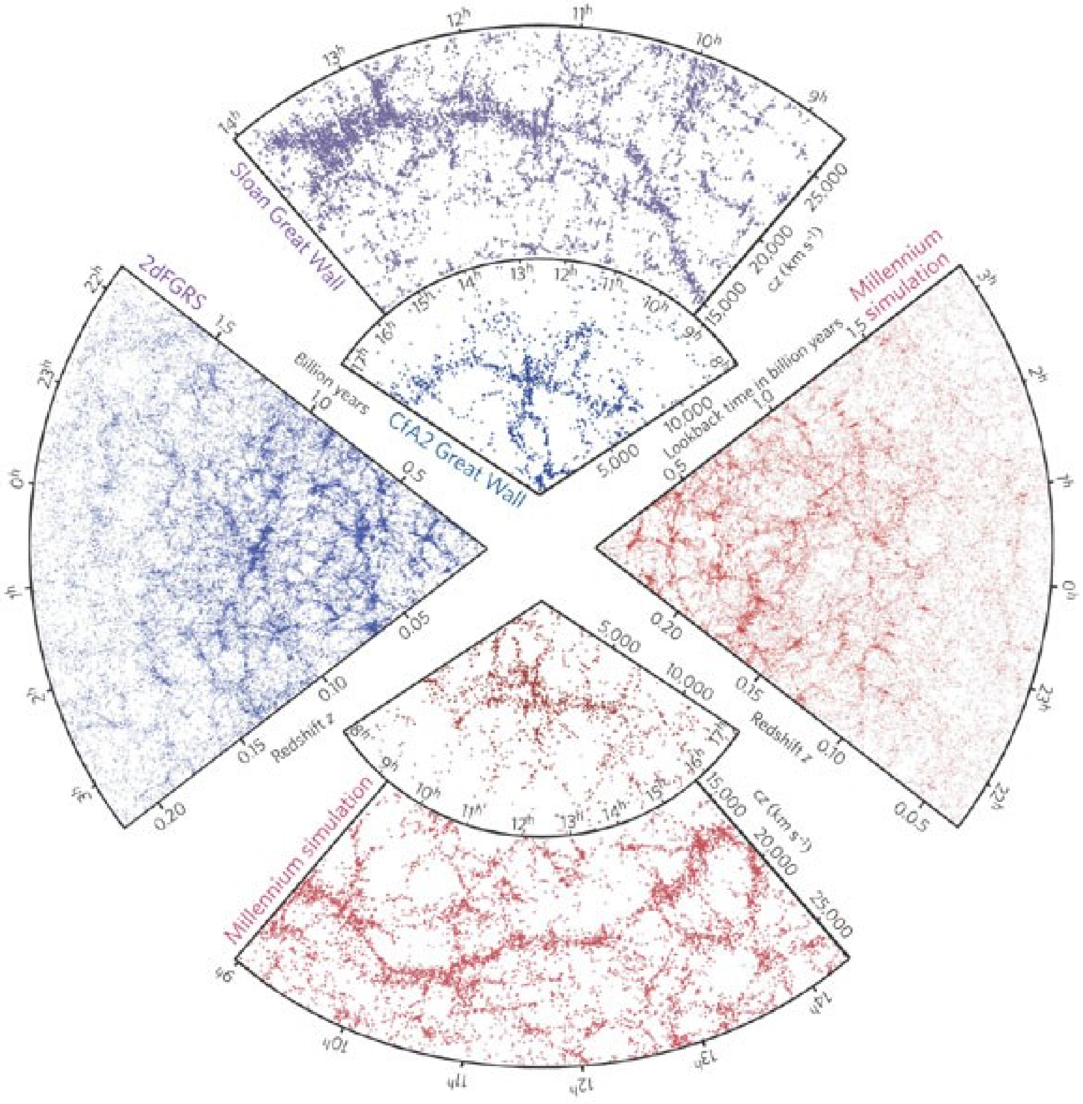


Figure 1.2: A comparison of simulations to observations, with warm colors (lower and right plots) from the Millennium simulation and cool colors (upper and left plots) from the actual sky. Reproduced from Springel et al. [2006], Figure 1. Original caption: “The small slice at the top shows the CfA2 ‘Great Wall,’ with the Coma cluster at the centre. Drawn to the same scale is a small section of the SDSS, in which an even larger ‘Sloan Great Wall’ has been identified. This is one of the largest observed structures in the Universe, containing over 10,000 galaxies and stretching over more than 1.37 billion light years. The cone on the left shows one-half of the 2dFGRS, which determined distances to more than 220,000 galaxies in the southern sky out to a depth of 2 billion light years. The SDSS has a similar depth but a larger solid angle and currently includes over 650,000 observed redshifts in the northern sky. At the bottom and on the right, mock galaxy surveys constructed using semi-analytic techniques to simulate the formation and evolution of galaxies within the evolving dark matter distribution of the ‘Millennium’ simulation are shown, selected with matching survey geometries and magnitude limits.”

density σ_8.[1] Therefore, if we knew the mass and redshift of every DM halo perfectly, we could put tight limits on cosmological parameters, including the Hubble constant H_0, Ω_m, and σ_8. However, **no direct method of measuring a 3D profile of cluster mass exists.** Even gravitational lensing only measures a 2D projection, which suffers from line-of-sight interlopers and only has high signal-to-noise for nearby, high-mass clusters.

To circumvent the inaccessibility of gravitational lensing, we devise **mass proxies**: quantities more readily measured than mass itself which relate closely back to total cluster mass. One such proxy in the optical wavelength regime is **richness** λ, a count of luminous red galaxies (LRGs) within a cluster. Preferentially found towards the hearts of large clusters, LRGs are often in virial equilibrium (roughly speaking, in stable orbits, not expanding or collapsing as a whole; relaxed, close to hydrostatic equilibrium); this implies that analyses of LRGs likely relate closer to virial mass of a cluster than analyses including blue (often actively accreting) galaxies. We thus expect a count of LRGs to scale closer to cluster mass than counts of unvirialized galaxies, still actively falling into the cluster. For example, if a galaxy cluster with 20 LRGs weighs $M = 10^{14}\ M_\odot$, then one might expect a galaxy cluster with 200 LRGs to weigh roughly $M = 10^{15}\ M_\odot$ (following an order of ten increase for both λ and M). Mass proxies thus allow inclusion of clusters without directly measured masses into cosmological analyses.

The expansion of space stretches out light and makes it lose energy as it travels through space, so nearly all galaxies we see are somewhat reddened by this effect. Once this redshifting is accounted for (shifting the spectrum to its original rest frame), we observe two distinct populations: red galaxies and blue galaxies. But how does one delineate between the two groups, and what drives their bimodality?

1.1.3 Rosy Galaxies

Galaxies come in two main flavors: red and blue (as observed in their rest-frames). This distinction between populations is primarily driven by their star formation rates (SFRs): how rapidly they are forming stars (if at all). As the net observed spectrum from a galaxy will tend to be dominated by its brightest members (a supernova can momentarily outshine their entire host galaxy, though they're more often closer to a tenth or hundredth as luminous), we can focus our attention on the most massive stars in a galaxy.

Massive stars burn hot, bright, blue, and quickly. The Wein displacement law of blackbody radiation states that hotter objects shine bluer, since peak radiation wavelength is inversely proportional to absolute temperature. As more massive stars tend to burn hotter and brighter than lower-mass stars, this means that they shine bright blue, then redden as they exhaust fuel, cool down, and

[1] More technically, σ_8 is the amplitude of the linear power spectrum on a scale of 8 Mpc$/h$, but it's sufficient to this dissertation to describe it as simply the clumpiness of space.

eventually die. Therefore, actively star-forming galaxies tend towards a blue hue. However, the more massive a star, the shorter it lives, as it burns more rapidly through its available fuel. The expected lifespan of a typical ("main sequence") star is $\tau_{\mathrm{MS}} \propto M^{-2.5}$ [Hansen et al., 2012], such that if a star is ten times more massive than our sun, it is expected to only live roughly $1/300$th as long as our sun (roughly 30 million years, as cf. our Sun's expected lifespan of 10 *billion* years). This means that if new stars are being formed, the brightest will be blue stars, though they will burn out relatively quickly. If galaxies cease star formation, then no new bright blue stars color the galaxy blue. Without new blue stars to bluen the galaxy, the stars age, cool down, and redden, thereby reddening the entire galaxy as a whole. Galaxies that have stopped forming new stars will therefore redden with age. Regardless of cause, a galaxy is considered **quenched** if it ceases star formation (i.e. reduces its rate by several orders of magnitude). A cessation of star formation occurs when gas cannot cool, condense, & collapse to form new stars, or if the gas is stripped away altogether.

Ram pressure stripping [Steinhauser et al., 2016, Steyrleithner et al., 2020, Kolcu et al., 2022, Vulcani et al., 2022] occurs when a galaxy ploughs through a gas-dense region of space, such as when a spiral galaxy passes through a massive galaxy cluster, near its central region. Pressure from the intergalactic medium—the hot gas between galaxies in clusters—strips away gas from the spiral galaxy, leaving it considerably less gas-rich. What gas remains is heated and disrupted, quenching star-forming regions in the galaxy, vastly diminishing the galaxy's capacity to form new stars. The pressure of ramming into a hot, dense cloud of gas can also cause a shock in the gas, triggering feedback processes which further disrupt star formation.

The two main feedback processes of interest include supernova explosions (dominant in low-mass galaxies) and supermassive black hole jets (dominant in high-mass galaxies); both can heat up and expel gas from a galaxy [Naab and Ostriker, 2017, Li et al., 2018]. Shocks from ram-pressure stripping can create new stars, but it can also cause stars to explode, triggering supernovae. These explosions both heat up the surrounding gas and additionally drive the gas away from the host galaxy, both of which suppress star formation [Ceverino and Klypin, 2009, Hopkins et al., 2020]. Ram-pressure-induced shocks can also initiate activity from the supermassive black hole (SMBH), as they compress gas near the galaxy's core; this then powers the black hole's jets, which—similar to supernovae—both heat and expel gas from a galaxy [Ishibashi and Fabian, 2012]. This effect is particularly pronounced in high-mass galaxies, as SMBH mass is proportional to galaxy stellar mass [Magorrian et al., 1998, Ferrarese and Merritt, 2000]. Both of these feedback processes play a crucial role in ceasing star formation in galaxies, and also occur in galaxies not experiencing ram-pressure stripping.

Each of these factors [and others; see Mamon and Silk, 2012, for a review] lead to a cessation of star-formation in galaxies. These various pathways cause several notable differences between the spectra of star-forming versus quiescent galaxies.

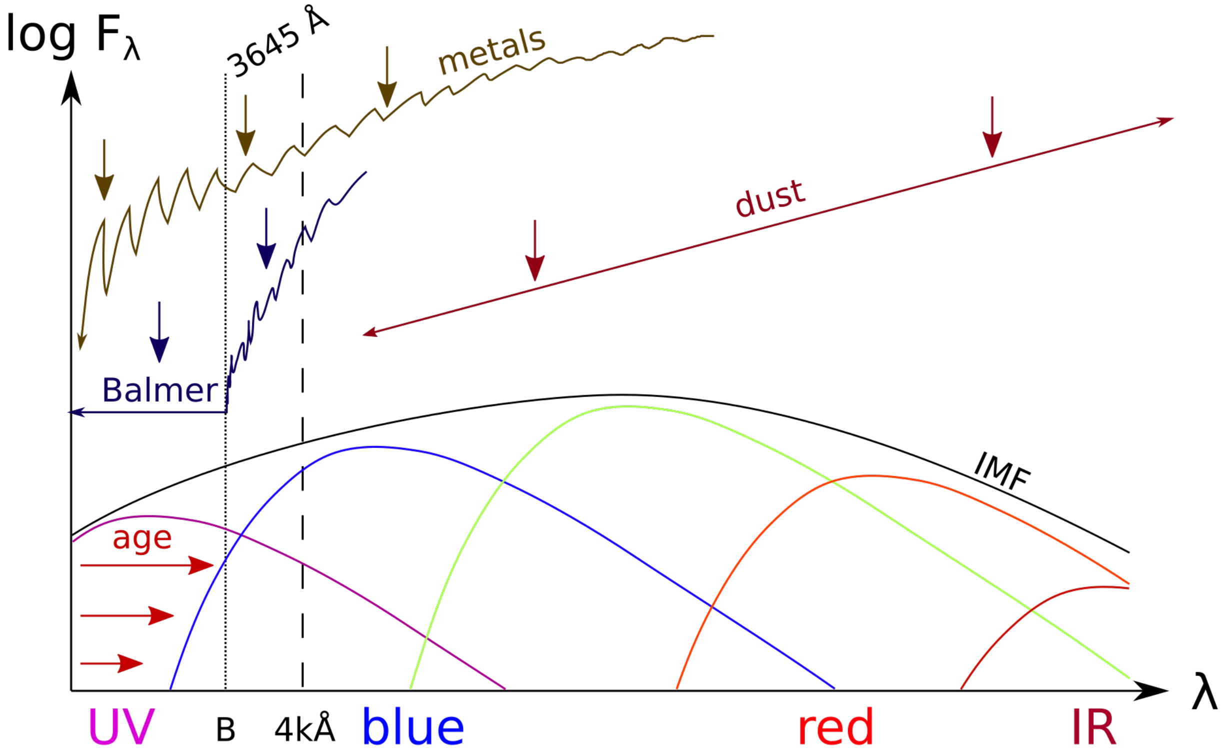


Figure 1.3: Summary cartoon of main astrophysical effects determining galaxy colors. Stars provide the vast majority of optical light. Portrayed here for five temperatures as the five downward-facing curves under the IMF, hotter stars stars emit brighter and bluer light whereas cooler stars emit dimmer and redder light. Because extremely hot and bright stars are both shorter-lived (lifespan $\tau_{\mathrm{MS}} \propto M^{-2.5}$) and less abundant, their relative contribution may be less than more abundant, redder stars. The net contribution of all stars of all temperatures produces the overall shape of the emission spectrum; this depends on both the stellar initial mass function (IMF) as well as the galaxy's star formation history (SFH). As stars age, they produce metals and cool down, reddening from both processes. This makes galactic spectra tend to redden with both age and metallicity, particularly so at the shortest wavelengths. The two effects are degenerate; galaxies with identical $\tau Z^{3/2}$ (where metallicity Z is the fraction of an object's mass which comes from elements heavier than helium) have virtually identical optical colors [Worthey, 1994].
Primary sources of absorption include dust reddening and line absorption. Dust presence causes Rayleigh scattering, the preferential scattering of short-wavelength light (intensity $I \propto \lambda^{-4}$). Such scattering reddens the overall spectrum, which can be clearest observed in the post-break spectrum slope, at wavelengths longer 4kÅ. Various metals lead to a cornucopia of absorption lines, convening around and especially below 4kÅ, resulting in a blanket of absorption. In contrast to the more dispersed and gradual effects of age or metallicity, hydrogen line absorption asymptotes sharply towards 3645 Å from above (with shorter wavelengths both fully ionizing hydrogen atoms and imparting a surplus kinetic energy kick). This convergence of the Balmer series results in a sudden drop in emissions—sharper than that of line blanketing. Thus the IMF & SFH, age, dustiness, and metallicity of galaxies all play a role in determining their overall spectrum.

Figure 1.3 illustrates main causes of spectral differences between galaxies in the optical and infrared range of wavelengths. Besides dark matter, galaxies are chiefly composed of stars, gas, and dust: by mass, our own Milky Way galaxy is roughly 95% DM, 4% stars, and 1% gas [Rodriguez Wimberly et al., 2022]. The Balmer break is caused by the ionization of electrons from the second energy level of hydrogen atoms; any wavelengths shorter than the Balmer wavelength of $\lambda_{\mathrm{B}} \sim 3645$ Å have sufficient energy to completely ionize a hydrogen atom with an electron in the second energy level (or higher). As stars cool down, their hydrogen content moves from an ionized state (infinith energy level) down towards the second energy level, allowing the Balmer break to manifest, resulting in a strong dip in the spectra of these older, cooler stars. Between there and roughly 4000 Å, similar absorption is caused by a synchronicity of metallic[2] ionized absorption lines (as stars make more metals with age, the frequency and strength of these lines increases with galactic age). As mentioned, these effects run with age, which tends to decrease a given star's temperature and increase its metallicity. Dustiness also plays a role, with Rayleigh scattering causing an overall reddening of galactic spectra, depressing shorter wavelengths.

All of the above factors conspire to cause a significant difference in measured intensity around 3800 Å (ranging from the Balmer break at 3645 Å to the 4000 Å break) between galaxies which have largely ceased forming stars and actively star-forming populations. One can measure D_{4000}, the relative intensity on either side of this break in the spectrum,[3] and get a very good estimate as to whether or not this galaxy is actively forming stars. While measuring precise spectroscopy for a large sample of galaxies is prohibitively expensive, the measurement can be made over averaged swaths of the spectrum, requiring only two measurements bookending either side of the break, rather than the precise measurements required for a full spectrum.

1.1.4 Photometry

Rather than measure the full spectrum of a galaxy, one can measure averaged portions of a spectrum using **bandpasses**. Photometric bandpasses act roughly like box functions, integrating the energy of a measured spectrum across the filter's width. (However: it's important to note that measured light diverges significantly from a perfect box function of the observed spectrum. Most significantly, ground-based telescopes suffer from atmospheric absorption; our air absorbs the majority of wavelengths shorter than ultraviolet and absorbs many patches in the infrared range.) Astronomers then convert the integrated flux into what's known as a photometric "**magnitude**",

[2] When astronomers speak of 'metals', they refer to elements heavier than Helium, rather than the chemical definition.

[3] Particularly, it is the ratio of the average spectral intensity in 4050–4250 Å over the average spectral intensity in 3750–3950 Å; see Bruzual A. [1983].

proportional to the negative log flux of the galaxy.[4] The x-band magnitude is calculated as:

$$m_x \equiv -5 \log_{100}(F_x / F_{x,0}), \tag{1.3}$$

where $F_{x,0}$ is some reference point like the flux of Vega or a fixed flux density (as with the AB system of photometry DES uses for its u, g, r, i, z, and Y bands). With an array of bandpasses spanning wavelengths of interest, one can then get a far cheaper pseudo-spectrum for an object (and, indeed, using stellar population synthesis modeling, one can probabilistically impute an entire spectrum from a set of photometric measurements). This can measure the relative spectral intensity on either side of the 4kÅ break, i.e. D_{4000}.

An ideal photometric measurement of D_{4000} would use the difference of two magnitudes from thin bands on either side of the feature of interest, with one near the Balmer wavelength and the other just above 4000 Å. (This isn't strictly required; it turns out that many neighboring bands register significant spectral differences between star-forming and quiescent galaxies—at least in their distribution, if not in their mean value.) Because observed wavelengths become longer with redshift, with observed wavelength following

$$\lambda_{\text{observed}} = \lambda_{\text{rest}}\,(1 + z) \tag{1.4}$$

(where z is redshift and λ_{rest} is the emitted wavelength, as observed in the object's rest frame), the choice of which two bands optimally measure D_{4000} changes across redshift. As the incoming light is increasingly redshifted, longer-wavelength bandpasses are required to capture the 4kÅ break.

To compare the ratio of fluxes with magnitudes, as when measuring D_{4000}, one must subtract the two magnitudes. For two bandpasses, e.g. DECam g and r-bands,

$$m_g - m_r = -2.5 \log_{10}\left(\frac{F_g}{F_{g,0}}\right) + 2.5 \log_{10}\left(\frac{F_r}{F_{r,0}}\right) = -2.5 \log_{10}\left(\frac{F_g}{F_r}\right) \tag{1.5}$$

(the flux baselines cancel out, as DES uses the AB magnitude system, as mentioned earlier, with baselines having a fixed flux density as reference). This difference in magnitudes, written as "$g - r$" for simplicity, is then proportional to the log of the relative fluxes of each band. It is known as a photometric "color", since colors, as humans perceive them, are relative intensities of various lights;[5] although an object may give off many wavelengths of light, we tend to name them by their

[4] This sadistic system was created to roughly match the ancient system of Hipparchus from the second century BC; eventually, in 1856 AD Norman Pogson codified this standard, of five magnitudes corresponding to a factor of 100 change in brightness.

[5] Exceptions to this include over-saturation, which causes colors to appear white, as well as other distortions due to the nature of the human eye, such as red and green appearing yellow, or the sky appearing blue rather than a more violet hue.

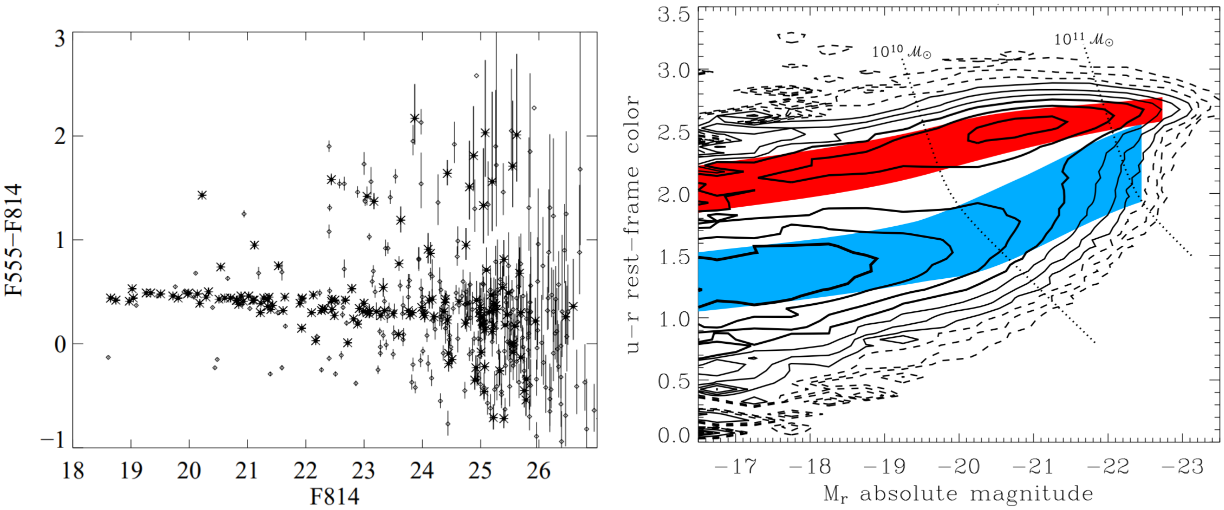


Figure 1.4: Examples of color–magnitude diagrams of galaxies. Color on vertical and magnitude on horizontal for each. Note that magnitudes of the right plot are absolute (i.e. dimming due to distance is accounted for, such that it measures *intrinsic* luminosity, rather than *apparent*) and are plotted in reverse order to those of the left plot. **Left:** Galaxies observed when viewing the core of galaxy cluster Abell 2390 with the *Hubble Space Telescope* [Gladders et al., 1998]. Asterisks indicate elliptical (generally quiescent) galaxies; diamonds indicate other morphologies. Reproduced from Gladders and Yee [2000]. **Right:** Galaxies from low-z SDSS wide field observations; contours show number count density, corrected for selection effects. Colored population fits calculated by Baldry et al. [2004]. Reproduced from Mamon and Silk [2012, Fig. 3].

dominant wavelength.

1.1.5 Galaxy distribution in photometric space

When galaxy clusters were first viewed in color–magnitude photometric space (i.e. plotting spectral slope versus spectral intensity), astronomers were quite surprised to discover a narrow band of quiescent galaxies tightly clustering around the same line [Bower et al., 1992]. Figure 1.4 depicts the galaxy distribution in color–magnitude space for both the galaxy cluster Abell 2390 (left) as well as for the wide-field galaxy survey SDSS (right). Galaxies in this red and tightly clustered color region were called Red Sequence (RS) galaxies, whereas galaxies in the blue and relatively loosely grouped region were called Blue Cloud (BC) galaxies. When the photometric color spans the 4000 Å break (such as rest-frame $u-r$ color, shown above), quiescent galaxies cluster together tightly at redder colors, whereas star-forming galaxies have a wider dispersion centered around bluer colors [Strateva et al., 2001, Bell et al., 2004].

At a fixed stellar mass (or magnitude), these distributions are well-modeled by a dual Gaussian mixture model (GMM), the combination of two normal distributions (see §1.1.6). Astronomers call the dip between RS and BC the Green Valley (GV). Because star formation rates are skew-lognormal, rather than bimodal, there is some disputation as to whether the GV represents a significantly distinct class of galaxies [Schawinski et al., 2014, Eales et al., 2017, Feldmann, 2017, Eales et al., 2018, Leja et al., 2022]. Regardless of classification, the GV tends to include galaxies actively transitioning from BC to RS [Schawinski et al., 2007, Schawinski, 2012].

As seen in the left panel of Figure 1.4, the RS is particularly strong in galaxy clusters, so much so that the RS could be used to identify clusters on the sky! Gladders and Yee [2000] proposed an algorithm for galaxy cluster detection which used the narrowness of the RS to find galaxy overdensities on the sky. Cluster RS members are all typically significantly redder than not only BC members, but also *all* galaxies at lower redshifts. This means that when trying to select galaxy cluster members in a field of view, removing all galaxies bluer than the cluster's RS virtually *eliminates* foreground contamination (galaxies between us and the cluster).

Koester et al. [2007] introduced MaxBCG, which, rather than using a simple color cut to find RS members, uses a redshift-evolving, error-cognizant, two-color model of the RS. It models $g - r$ and $r - i$ colors as Gaussian functions, with their means evolving with redshift and their widths constant. The two Gaussians are multiplied together to calculate likelihoods, enforcing null correlations between colors. In the following years, as data grew increasingly precise and abundant, additional photometric bands could then be taken into account and scatter of each color could be allowed to evolve, rather than remain constant.

The redMaPPer (RM) algorithm modeled a redshift-evolving multivariate (multi-color, accounting for correlations) Gaussian fit to bright members of the RS [Rykoff et al., 2014]. This model used each of the DES main photometric bands ($griz$, yielding primary colors $g - r$, $r - i$, and $i - z$) to measure a redshift-evolving mean color; the mean color evolved linearly with magnitude (which slope is visible in both panels of Figure 1.4) with a redshift-evolving slope in each color. While scatters were measured and allowed to evolve with redshift, inter-color correlations were frozen to $\rho = 90\%$, rather than being fit [Rykoff, 2021].[6] The fit focused exclusively on bright RS cluster members, as BC galaxies are less often as bright and have worse redshift estimates. The resulting characterization of RS mean color as a function of redshift and magnitude improved previous efforts in using the RS to estimate a galaxy's redshift from photometry.

Because the RS is constantly evolving in multi-color space, if photometry for a RS galaxy is known, its redshift can be estimated with fairly good precision. The redMaGiC algorithm [Rozo et al., 2016] does just this, using the RM fit to the RS to estimate redshifts for LRGs (bright RS members). This resulted in a median difference of redMaGiC-estimated photometric redshifts z_{photo} ("photo-z") from known spectroscopic redshifts z_{spec} ("spec-z") of $z_{\text{spec}} - z_{\text{photo}} = .005$ and

[6] This increased stability of the code, as compared to allowing correlations to freely evolve, and aligned fit results better with simulated and spectroscopic data. It tended to reduce outliers and also computed richness.

a scatter of $\sigma_z/(1+z) = .017$, with only 1.4% of galaxies as 5σ outliers. Combining the photo-z estimates of all cluster members gives an accurate and precise estimate of a galaxy cluster's redshift, as compared to estimating cosmological redshift (i.e. distance) from lone galaxies. Alternatively, this relation between RS mean color and redshift can be turned on its head and used to identify clusters. If you assume all galaxies are RS members and estimate photo-z values for each, then certain regions of the sky will have galaxies which are not only clustered on the (2D) sky, but also clustered in redshift space as well. For example, if on a small patch of the sky—subtending $\lesssim .1^\circ$ (the moon subtends just over half a degree; this is roughly the size of a small pea held at arms' length), you find 200 galaxies with very similar photo-z estimates, you can have good confidence that you've found a galaxy cluster.

A significant downside of the RM algorithm is its ignorance of the BC. Projection effects—where clusters lie along the same line of sight—significantly limit the power of galaxy cluster cosmology in constraining cosmological parameters [Allen et al., 2011]. Identifying whether an observed cluster is truly one or two halos is difficult; one can easily misinterpret two halos as a single cluster, or misinterpret a single halo as two clusters. This corrupts analyses, weakening their potential to constrain cosmological parameters. Background subtraction—removing all galaxies not belonging to a cluster—is also challenging, but having a model of the BC would allow for more informed cluster memberships; photo-z estimates for the BC would often be less precise than for the RS, but they would vastly help in determining cluster membership. (For example, a galaxy which looks like a RS member at some redshift could potentially be a BC member at a higher redshift, or what looks like a background cluster could actually be actual BC members of a foreground cluster.) In order to model both RS as well as BC, we use Gaussian mixture models (GMMs), modeling the two populations as separate Gaussian distributions.

1.1.6 GMM modeling of galaxy populations

A Gaussian distribution (also known as a "normal distribution" or "bell curve") is characterized by its mean μ and variance σ^2:

$$\mathcal{N}(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]. \tag{1.6}$$

For a multidimensional distribution, one must not only consider the variance σ^2, but additionally the correlation ρ between dimensions. This is symmetrically encoded in the covariance matrix:

$$\Sigma = \begin{bmatrix} {\sigma_1}^2 & \rho_{1,2}\sigma_1\sigma_2 & \cdots \\ \rho_{2,1}\sigma_2\sigma_1 & {\sigma_2}^2 & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} \tag{1.7}$$

Combining this with the vector of mean colors $\vec{\mu}$ gives a multidimensional Gaussian distribution:

$$\mathcal{N}(\vec{x};\vec{\mu},\Sigma) = \frac{1}{\sqrt{(2\pi)^N\,|\Sigma|}} \exp\left[-\frac{1}{2}(\vec{x}-\vec{\mu})^{\mathrm{T}}\Sigma^{-1}(\vec{x}-\vec{\mu})\right], \tag{1.8}$$

where N is the dimensionality of the system (equal to the length of the vector μ) and the vertical bars about Σ indicate its determinant. Uncertainties in measurements can be encoded in a noise covariance matrix Δ (see equation (3.4)), which is added to Σ such that $\Sigma \longrightarrow (\Sigma+\Delta)$ in each instance above. Because colors far from the 4kÅ break often still exhibit significant differences in mean or scatter of RS and BC, these added colors can greatly aid in identifying whether a galaxy belongs to the RS or BC. This gives a multi-color model more power than using a single photometric color alone (e.g. as a measurement of D_{4000}), as had been common in selecting quiescent galaxies.

As the dual photometric populations of RS and BC diverge from unimodality, mixtures of Gaussians may then be used. In particular, we can write the likelihood of a given vector of galaxy colors $\vec{x}$ given a parameter set $\boldsymbol{\theta} = \{f_R, \vec{\mu}_{\mathrm{RS}}, \Sigma_{\mathrm{RS}}, \vec{\mu}_{\mathrm{BC}}, \Sigma_{\mathrm{BC}}\}$:

$$\mathcal{L}(\vec{x}\,\big|\,\boldsymbol{\theta}) = f_R\,\mathcal{N}(\vec{x};\vec{\mu}_{\mathrm{RS}},\Sigma_{\mathrm{RS}}) + (1-f_R)\,\mathcal{N}(\vec{x};\vec{\mu}_{\mathrm{BC}},\Sigma_{\mathrm{BC}}) \tag{1.9}$$

where red fraction f_R then gives the relative abundance of each population. By optimizing likelihood $\mathcal{L}$, we can thus find a best fit to the populations. Note that the model is readily expanded to include additional components; this may be necessary if the populations are not fit well by a two-component model or if the RS population is extremely weak (e.g. $f_R << 10\%$).

When characterizing galaxy populations in photometric color space, the parameter set $\boldsymbol{\theta}$ evolves significantly with several factors. If redshift or stellar mass are not accounted for, the color distribution of galaxies may distort from a bimodal distribution.

Each element of $\boldsymbol{\theta}$ depends strongly on redshift. As a galaxy's incoming light is increasingly redshifted, the 4kÅ break drifts from being observed in one photometric band to another. This causes a spike in photometric color about this transition redshift as well as shifts in other parameters [Nishizawa et al., 2018, DeRose et al., 2019]. To a lesser extent, we observe a trend of reddening with increased age of the universe known as the Butcher–Oemler effect [Butcher and Oemler, 1978], as the fraction of quenched galaxies increases monotonically from cosmic dawn (at redshift

$z \sim 2$) to the present ($z \sim 0$).

The parameter set $\boldsymbol{\theta}$ is also known to depend on stellar mass [Baldry et al., 2004, Balogh et al., 2004]. Not only are galaxies more likely red at high stellar mass, but the redness of that mean color also increases, for both RS and BC (see Figure 1.4). Scatter of the RS decreases with increased stellar mass whereas BC scatter has a non-monotonic dependence on stellar mass.

The red fraction f_R is also known to drift significantly with environment [Balogh et al., 2004]. High-density environments like galaxy clusters tend to have far higher red fractions than low-density environments like voids, or even filaments. It remains unclear whether mean colors of RS and BC significantly drift with local density, or whether the small drifts in mean colors observed are due to differences in stellar mass (since stellar mass correlates with environmental density).

Because all parameters of $\boldsymbol{\theta}$ drift with galactic stellar mass and strongly depend on redshift, we thus desire a model of the galaxy populations which can account for these dependencies. This is the backbone of the Red Dragon algorithm, described in Chapters 3 & 4.

In a sidestep from my core cosmology and astrophysics research, Chapter 5 discusses my foray into the world of learning analytics, investigating grade gains due to practice study in introductory physics classes.

1.2 Learning Analytics

As mentioned in the preface, I researched student learning of physics in the summers of 2021 and 2022, investigating grade gains due to practice study. The field of learning analytics (LA) uses data to gain insights into education; when focused on a particular discipline, it is referred to as Discipline-Based Education Research (DBER); in the context of *physics* education, it takes the appellation of Physics Education Research (PER) [Fraser et al., 2014, Henderson et al., 2017]. For my particular project in PER, I mined data from a service called Problem Roulette (see §1.2.2); this data draw and subsequent analysis falls under the realm of Educational Data Mining (EDM), using data mining and machine learning to discover patterns in student learning and performance [Romero and Ventura, 2010].

The nascent field of learning analytics, at the intersection of education and data science, holds immense power in shaping the future of education. Analysis of raw data yields insights into trends, allowing educators to implement evidence-based interventions to improve academic outcomes. In particular, we quantified the extent to which practice study improves student grade outcomes.

1.2.1 How did you get that A?

What, exactly, causes students to do well in courses? This question has been posed by students, teachers, and educational researchers alike. Student success is often seen as a mixture of incoming skills and knowledge along with effort put towards learning (e.g. study volume), but no consensus among researchers exists as to the absolute hierarchy of factors causing awarded grades. Short of performing human experimentation, most analyses perform only post-hoc analyses on outgoing data, focusing on *correlations*. Many factors correlate with academic outcomes.

A meta-analysis by Richardson et al. [2012] showed that, after academic self-efficacy (essentially asking students: "Do you expect to do well?"), measures of intelligence correlated highly with grade outcomes, followed by grade goal and personality traits. No demographic nor other psycho-social contextual factors were significantly correlated, after accounting for the former factors. In the sections that follow, we discuss the effect of effort, intelligence, personality, and demography on grade outcomes.

1.2.1.1 Intelligence

If a student is a faster or more efficient thinker—if they tend to learn from examples quicker than the average student—then practice study will impact their grade differently. Though the exact definition of *intelligence* is somewhat disputed, psychometric researchers elide this issue by focusing on objective, quantitatively defined indicators. Best known in popular science is the intelligence quotient (IQ) test (originating in 1905 by Alfred Binet and Théodore Simon) [Williams, 1915], which provides a quantitative number relating to mental capacity. Among modern psychometric researchers, "the g factor" is primarily used as a measure of general mental ability [Jensen, 1998]. The accuracy wherewith an individual is able to perform certain tasks correlates with ability to perform other certain tasks. Increased strength of correlation (with other tests, and therefore with g) means the test is a better measure of a global factor which underlies each process. Tests such as vocabulary comprehension, mathematics, spatial visualization, and pattern completion have particularly high correlations (IQ tests correlate $\sim 77\%$ with g), indicating that they better measure g than tests like musical ability (which has a relatively low correlation, meaning that it is a poor predictor of g).

Psychometric measures like g have strong correlates with various outcomes. Of particular relevance to this discussion, they bear strong, positive correlation to favorable academic outcomes [even on accounting for differences in education, occupation, and socio-economic background; see Jensen, 1980, 1998, Herrnstein and Murray, 1994, Soares et al., 2015, Singh et al., 2022]. Because g also correlates with several social and economic indicators [Jensen, 1980, 1998], we expect to see significant second-hand correlations between these socio-economic factors and grades.

Tests such as the ACT and SAT are effectively psychometric tests, taking proxy estimates of intelligence [roughly 85% correlated with *g*; see Frey and Detterman, 2004, Koenig et al., 2008, Coyle and Pillow, 2008]; the better individuals can digest, retain, infer, and apply information, the higher they tend to score. These tests largely measure innate ability, as increased study or tutoring have marginal effects on score at best [ACT scores shift by an insignificant 0.6% ± 2.6% while SAT scores shift by 1.75% ± 0.25%; see Briggs, 2001, 2009].[7] The tests show no cumulative measurement biases between genders and races (vis.: while men, women, whites, blacks, and hispanics each tended to do slightly better and worse on different individual test questions, in aggregate the test has no bias) [Drasgow, 1987, Jensen, 1980, Brown et al., 1999]. The format of the ACT and SAT tests also matches that of many college courses (especially mathematics-based courses), where students answer high-stakes multiple-choice questions in a timed setting. Despite their shortcomings, the ACT and SAT remain widely available pre-college predictors of academic performance and retain significant predictive power in estimating college GPA [Noble and Sawyer, 2002, Coyle and Pillow, 2008, Koester et al., 2016, Rodríguez-Hernández et al., 2020].

However, measures of cognitive ability alone are insufficient to fully explain student grades. For example, many successful graduates score low on the ACT & SAT tests, yet still succeed regardless. Quantitative measures of *personality* additionally have significant yet perpendicular (i.e. not explained by intelligence) power in explaining student grade outcomes.

1.2.1.2 Personality

Köseoglu [2016] measured several significant correlations between grade outcomes and the Big Five personality traits [Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism; discovered by the same factor analysis that discovered *g* factor; see Roccas et al., 2002, Ridgell and Lounsbury, 2004, Higgins et al., 2007]. The most significant correlation was in conscientiousness (40% correlated with GPA, explaining 17% of the variance); Agreeableness and Openness are also found to correlate with academic performance to a lesser extent [see also Poropat, 2009].

As women tend to score higher in each of the Big Five personality traits on average, these correlates could explain a portion of why the average male student earns lower grades than the

[7] Briggs [2009] addresses the question of why > $4M annually is spent on test prep; several factors contribute. First, scores tend to increase on retesting, so without a comparison group, companies can boast correct yet deceitful messages of significant improvement of their patrons; companies that sell test prep materials have little incentive to provide rigorous statistics when misleading yet factual claims can lure in cash. Second, marginal improvements of score (e.g. 20 points on the SAT-Math, a $\sim$ 3% increase) are still perceived by both students and a minority of admission counsellors (even among selective colleges) as increasing likelihood of admission, despite the College Board explicitly warning against such interpretation. This perception increases with score, so students who already scored highly on the tests may feel more pressure to earn a (marginally) better score, regardless of actual realized benefit. With test preparation materials available freely online, and now that a large majority of common app colleges require neither test, hopefully less resources (both monetary cost as well as opportunity cost) will be wasted in such endeavors.

average female student [Aluja and Blanch, 2004, Al-Shawwa et al., 2014, Charles-Ogan, 2015, Alzahrani et al., 2018, Unwalla, 2020]. However, neuroticism has been shown to correlate with test anxiety over and above the effects of gender [Kaplan et al., 2015]. Because women on average score higher in neuroticism than men, this could explain why women on average perform worse in a stressful testing environment than men, despite on average performing better on lower-stress assignments.

Conscientiousness correlates highly with methodical study [Aluja and Blanch, 2004, Komarraju et al., 2011]. A notable and strong correlate (~ 75%) with conscientiousness is "grit", a measure of *perseverance of effort* (and incidentally, *consistency of interest*) towards long-term goals [Duckworth et al., 2007, Duckworth and Quinn, 2009, Duckworth et al., 2011, Credé et al., 2017]. Grit is largely perpendicular to intelligence, having an entirely separable and significant effect on students' grades [Borghans et al., 2016]. In particular, higher levels of grit explain higher levels of deliberate practice and self-regulation of study, which explain improved academic outcomes [Christopoulou et al., 2018, Zentner et al., 2018].

1.2.1.3 Study habits

Time and effort spent studying correlates highly with academic performance [Zimmerman and Kitsantas, 2005, Stinebrickner and Stinebrickner, 2008, Metcalfe et al., 2011, Lindo et al., 2012, Grodner and Rupp, 2013, Guidry, 2017, Pope et al., 2018, Spitzer, 2022]. For example, Stinebrickner and Stinebrickner [2008] shows that an additional hour of study per day yielded a gain equivalent to a five-point increase in ACT score. Richardson et al. [2012] found that self-regulatory learning strategies like time/study management and effort regulation were correlated with academic success. However, not all forms of study are equally beneficial.

Of particular relevance to our study of student grade gains from practice study are results of a meta-analysis by Dunlosky et al. [2013]. They showed that study methods such as summarizing, highlighting, and rereading were far less effective than methods such as **practice testing** ("self-testing or taking practice tests over to-be-learned material") and **distributed practice** ("implementing a schedule of practice that spreads out study activities over time"). Both of these effective study methods benefited learners of different ages and abilities, boosting scores across a variety of situations. A more recent study showed again that spaced out quizzing on material predicted GPA superior to time spent reading the same material [Hartwig and Malain, 2022]. This supports use of regular practice quizzing, such as will be discussed in §1.2.2.

1.2.1.4 Demographics

A particularly contentious topic in grade causality is the differences of grade between various demographic groups. For example, in two introductory physics courses (140 and 240) at the University of Michigan, the average grade of males is $.15 \pm .02$ points higher than the average grade of females—a 7σ significant difference! More strikingly, international students on average score $.74 \pm .05$ points higher than underrepresented minority students—a 14σ significant difference! What causes these grade inequalities between demographic groups? Why doesn't every demographic group have a statistically identical grade distribution? Are gender, race, or other demographic indicators *causal* factors, or merely *incidental*?

These questions may be answered in part through analyses, but experimentation such as twin studies would really be needed to determine absolute causality; instead, we must relegate ourselves to investigate which factors best explain grades—if nationality is a better predictor of grades than intelligence, personality, and study combined, then something is seriously wrong. However, such analyses are fraught with confounding factors such as the selection effect of admissions, only admitting particular populations of students. To account for such selection effects, one must normalize such findings against secondary indicators (such as looking not only at grade alone, but course grade relative to other courses, or at a fixed high school GPA).

Each of the previous sections have correlates with various demographic distinctions. Measures of **cognitive ability** differ significantly by socio-economic status on average. Measures of **personality** and of **study habits** differ significantly between the sexes and between cultures on average [Costa et al., 2001, McCrae et al., 2005, Chapman et al., 2007, Schmitt et al., 2008, Weisberg et al., 2011, Hsin and Xie, 2014, Weaverdyck et al., 2020]. Because each of these factors bear predictive power on learning outcomes, it is unsurprising that we see differences in grade outcomes between these demographic categories. However, studies which account for each of intelligence, personality, and study habits can then measure deviations *beyond* expectations from the former three factors for various demographic sub-groups, determining whether significant residuals remain.

1.2.2 Problem Roulette

To quantify grade gains due to study, we use data from an online study service hosted at University of Michigan known as Problem Roulette (PR) [Evrard et al., 2015]. In short, after selecting a particular course and topic, PR serves individual problems from past exams to students, allowing them access to a powerful study tool. The service is optional and free to all students. In its original form, in 2011, it pulled questions from a Google server; it has since been streamlined, with many teacher analytics available, such as fraction of students answering a particular question correctly. Since 2017, PR has served over six million problems to over thirty thousand students for over a dozen courses (the exact number of courses fluctuates year to year). Local instructors supply multiple-choice questions from past exams to the service, which reads them in and presents them to future students.

Using data from this service (and connecting to student grades), we can forge valuable insights into student learning. From PR (through the UM Center for Academic Innovation), we have data such as how many problem-answering sessions a student engaged in, what time of day (or of the week, or of the year) the session was initiated, how long the session lasted; how many questions the student attempted per session, whether they skipped the question or answered it correctly or re-tried the problem; whether the students used exam mode, topical mode, or timed mode; and which courses the students interacted with, allowing cross-course comparison of study behaviors. This allows for creation of secondary fields like the clusteredness or consistency of study sessions (as measured e.g. by entropy or standard deviation of start times).

From the UM Learning Analytics Data Architecture (LARC) database [Lonn and Koester, 2019], we can then connect these measures of student study volume to scholastic and demographic indicators. We focus on course grade (as tied to the course studied for) and GPAO (cumulative GPA, omitting the course in question) as outcomes. We pull ACT and SAT math sub-scores as measures of mathematical ability. We additionally retrieve various demographic indicators, including parental education, nationality, ethnicity, high school zip code (from which we pull data about population and median income), and other fields (discussed in §5.2.4). These all then provide us with a measure of academic outcomes, proxy measures of cognitive ability, and demographic distinctions.

Combining information from PR and LARC, we now have a dataset ripe for analysis. We focus our analysis on two main questions. First: How much does study benefit students? Second: Are any demographic groups struggling to succeed, all else equal? Though we lack direct information on student personalities, we are well-poised to make precise measurements and provide quantitative answers to these questions. Rather than perform simple correlations or linear regressions, we turn to a local linear regression tool, capable of detecting non-linearities in trends (which we find many).

1.2.3 KLLR

Life is largely non-linear. On small enough scales, a single value may suffice, but as scope widens, linear trends appear, followed by quadratic, and so forth (in accordance with Taylor series). This is a fundamental issue with only looking at correlations or doing linear fits: as your domain of investigation widens (as you obtain more data), trends often move beyond linear. A standard linear regression (or correlation, or other scalar summary statistic) may fail to identify significant details in the data [see the Datasaurus Dozen for a beautiful example of how summary statistics miss out on important trends; explained in Matejka and Fitzmaurice, 2017]. We therefore often make use of a non-linear modeling tool to discern trends with more precision and clarity than traditional scalar or linear analyses.

Kernel-Localized Linear Regression (KLLR) measures a local intercept, slope, and scatter (and optionally skew & kurtosis) continuously along any trend [Farahi et al., 2018, Farahi et al., 2022]. KLLR measures these values with a Gaussian-weighted kernel, giving more consideration in the localized linear regression to nearby points when calculating a local slope. This single-parameter fit then only depends on the kernel width. See the KLLR GitHub page for a good illustration of varying kernel width. In short, if kernel width is too large, trends are distorted or lost; if kernel width is too small, slopes (and other fit values) become quite noisy. While some knowledge of the system at hand is useful for setting kernel width, results are often quite robust to kernel width variation. By starting at a large kernel width and gradually narrowing, one can then gradually reveal nonlinearities in trends.

This tool is used in each content chapter of this dissertation. Among other purposes, KLLR models galaxy occupation in DM halos, fits evolving GMM parameters, and measures student gains of study. In each case, KLLR reveals trends which otherwise would be undetectable or warped using simpler methods (or alternatively, would require the use of unwieldy high-order polynomial fits). This allows for more precise characterization of these trends, entering the era of precision cosmology (and nurturing the growth of physics education research).

1.3 Publication Status of Chapters

The following chapters of this dissertation have all either been published already or will soon be sent for publication.

- Chapter 2 was published February 2021 in the DES Document Database, #15737
- Chapter 3 was published October 2022 in Monthly Notices of the Royal Astronomical Society (MNRAS) Volume 516, Issue 1
- Chapter 4 will be submitted to MNRAS
- Chapter 5 was published June 2023 in Physical Review Physics Education Research (PRPER) Volume 19, Issue 1

CHAPTER 2

Buzzard Halo Report

Chapter Summary

In order to use galaxy clusters to constrain cosmological parameters, analyses require tight priors on confounding effects such as projection, selection, and miscentering. The Buzzard Flock (a set of synthetic sky catalogs) supports the Dark Energy Survey by providing a means to model and quantify these and other effects. Buzzard uses the AddGals algorithm to map galaxies onto a base DM-only simulation, matching the observed luminosity function and luminosity-limited two-point correlation functions by design, but the method is largely decoupled from dark matter halo information. In this report, the resultant halo occupation distribution (HOD) of Buzzard 2.0.0 halos is critically examined, with a particular focus on the fraction of red galaxies (targeted by the redMaPPer cluster finder). A new six-color RS finder, called Red Dragon, is developed and its results compared with existing RS definitions.

Buzzard halo HODs show distortions as compared to existing observations and simulation expectations. Key divergences are: i) the RS in high-mass halos is underpopulated, with mean red fractions below 60% in halos with masses above $10^{14.5}$ $\mathrm{M}_\odot$; ii) the mass dependence of the mean red fraction is inverted, with low mass halos having slightly higher red fractions than high mass halos, and; iii) in high mass halos the radial galaxy number density profile deviates significantly from NFW expectations, with a deficit of galaxies around one-tenth of the virial radius.

2.1 Introduction

2.1.1 Richness–mass scaling relation

Counting the number of galaxy clusters of a given richness at different redshifts, one can infer cosmological parameters. Equation 2.1 gives the number of halos in log mass bin μ_a, redshift bin

z_i, and solid angle Ω on the sky.

$$\bar{N}(\mu_a, z_i) \equiv \bar{N}_{ai} = \frac{\Omega}{4\pi} \int_{z_i}^{z_{i+1}} dz \left(\frac{dV}{dz}\right) \int_{\mu_a}^{\mu_{a+1}} d\mu \left(\frac{dn}{d\mu}\right) \tag{2.1}$$

The volume element $dV/dz = 4\pi r^2(z)/H(z)$ depends solely on cosmology, since comoving distance $r(z)$ is a function of the Hubble parameter $H(z)$. Analytic and emulated models connect the mass function too back to cosmology. Thus integrating the RHS only allows certain combinations of cosmological parameters to yield certain halo number counts (the LHS). Therefore accurate halo counts constrain cosmological parameters.

Though the volume element (first integral) is well constrained by SNe and BAO measurements, determining the true mass of a halo (to which mass bin μ_a it belongs; second integral) takes considerable effort. Besides lensing measurements, which give highly accurate masses for large nearby clusters, we only have *proxies* for mass, such as galaxy count, temperature, or luminosity. We need a kernel, then, to connect an observed signal, such as richness λ, to the underlying halo mass.

$$N_{ij} = \frac{\Omega}{4\pi} \int_{z_i}^{z_{i+1}} dz \int_{\lambda_j}^{\lambda_{j+1}} d\lambda \int_{-\infty}^{\infty} d\mu \, \left(\frac{dV}{dz}\right) \left(\frac{dn}{d\mu}\right) \, \mathrm{Pr}(\lambda|\mu, z) \tag{2.2}$$

The mass-conditioned richness kernel $\mathrm{Pr}(\lambda|\mu, z)$, commonly known as the richness–mass scaling relation, gives the probabilistic count of bright red galaxies (i.e. richness $\lambda \equiv \sum P_{\mathrm{red}}$) in a halo of a given mass and redshift.

Though the halo model approximates well the observed dark matter distribution, the LHS number counts N_{ij} are sometimes ambiguous. Even with perfect information about 3D distribution of galaxies, no perfect distinction exists between calling an extended system two nearby halos or calling it a single, peanut-shaped halo. In observations, projection complicates things even more: two halos may lie in the same line of sight and be mistaken for a single cluster. For more potential issues with the cluster–halo connection, see Allen et al. [2011]. Effects of systematic issues like these can be estimated via simulation to account for their impact in cosmological parameter estimation.

2.1.2 Richness in the HOD

The largest uncertainty in equation 2.2 comes from the kernel $\Pr(\lambda|\mu, z)$, connecting the halo mass function $dn/d\mu$ to observables like λ. What we need then is a link between observables (such as galaxy count) and intrinsic qualities (such as mass), known as the halo occupation distribution (HOD). Broadly speaking, the HOD determines how galaxies behave within halos (of various masses), entailing three main concepts:

1. λ–M scaling, i.e. $\Pr(\lambda|\mu, z)$ or its inverse $\Pr(\mu|\lambda, z)$;
2. radial distribution of galaxies in the halo; and
3. expected velocity distribution of galaxies in the halo.

To complete the halo model, we must also consider how halos relate to each other—how they cluster in position and velocity space (observed as clustering in angle and redshift).

Optical richness λ is a probabilistic count of red galaxies above some brightness threshold (here $0.2\ L_*$) in a galaxy cluster. These red galaxies generally correspond to the more central, relaxed parts of a halo, closest to hydrostatic equilibrium. Thus, richness relates to virial mass of a halo better than a raw galaxy count, which would include more infalling galaxies, not in equilibrium.

There are several upsides to using galaxy cluster richness. First, Optical light is relatively cheap, since it can be collected with simple mirrors or lenses based on the ground (as compared to X-rays, which must be observed without Earth's atmosphere, as orbiting telescopes). Second, new surveys from the Euclid and James Webb telescopes will detect many new clusters, allowing for much more complete views of galaxy richness. Third, getting an accurate HOD can inform astrophysics as well, by putting constraints on quenching, galaxy formation, galaxy evolution, &c. Fourth, the more data the better. As detailed in Evrard et al. [2014], one can use all observables at once (e.g. signals from weak lensing, SZ, optical, X-ray) to best estimate actual cluster masses. Therefore, it's in our interest to get the best possible priors on the λ–M relation.

DES uses the cluster detection algorithm redMaPPer [RM; Rykoff et al., 2014], which uses the RS to locate cluster centers and assign membership probabilities to nearby galaxies. See Euclid Collaboration [2019] for a comparison of other detection algorithms, as well as some details on some of the issues inherent in cluster detection—as mentioned in section 2.1.1, many confounding effects (such as projection, percolation, & miscentering) weaken the constraining power of richness [Allen et al., 2011]. Using simulations, one can model these effects and account for them as best as possible.

We expect the HOD of simulations to match observations (such as that of Simet et al. [2018], shown as the cyan band in Figure 2.1), but simulations can vary widely from reality, especially when user-defined baryon pasting (adding galaxies onto a dark matter only simulation) occurs

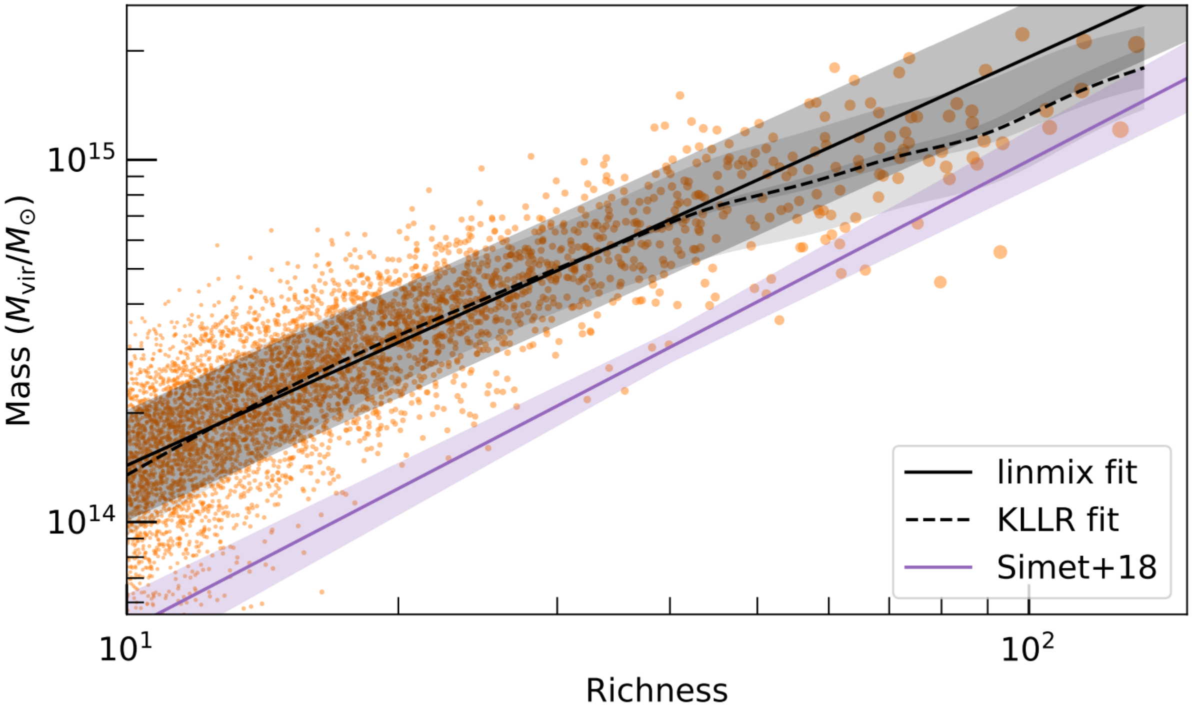


Figure 2.1: Buzzard HOD. Shown for all 4590 $z|[.1, .33)$ halos with $\lambda > 10$, showing variation in mass with richness. Following notation of equation C.1, the black linmix fit yields normalization $\alpha = 13.03 \pm 0.02$, slope $\beta = 1.13 \pm 0.01$, spread $\sigma = 0.153 \pm 0.002$. The dashed line shows a KLLR continuous fit for the data, revealing a lower slope and normalization at higher richness.

Cosmology	$\Omega_\mathrm{m} = .286$	$H_0 = 70$ (km/s)/Mpc
(ΛCDM)	$\Omega_\mathrm{b} = 0.046$	$n_s = 0.96$
	$\sigma_8 = 0.82$	$N_\mathrm{eff} = 3.046$
Conventions	Δ_vir, r_vir, M_vir	$\mu \equiv \log_{10} M_\mathrm{vir}$
Vetting	$\Omega_\mathrm{sky} \doteq 5000^\square$	$\mu \geq 13.75$
	$m_i < m_*(z) + 1.75$	$z\|(.1, .7)$

Table 2.1: Buzzard cosmology and conventions used. This table gives information about Buzzard Flock cosmology (ΛCDM) as well as the conventions and vetting bounds used in this analysis. The luminosity vetting was completed in i-band, using the characteristic magnitude of the luminosity function: $m_i < m_*(z) + 1.75$, where $m_*(z)$ is approximated in equation 9 of Rykoff et al. [2014], valid in the above redshift range.

rather than solving full magnetohydrodynamic equations. Nevertheless, simulations match reality well enough to put useful constraints on priors such as the richness–mass scaling relation.

2.1.3 The Buzzard Flock v2.0.0

A suite of cosmological-sized simulated galaxy surveys known as the Buzzard Flock provide simulation support for analyzing DES data [DeRose et al., 2019]. The AddGals algorithm [Busha and Wechsler, 2008] pastes galaxies onto a dark-matter only simulation, weighted by local DM densities, matching the observed luminosity function and luminosity-dependent two-point function.

AddGals uses the luminosity function $\phi(M_r, z)$ to assign galaxies of a certain r-band magnitude M_r to DM densities in a given DMO sim. This density is chosen such that it matches the luminosity-dependent two-point function: $P(R_\delta|M_r, z)$, which is modeled as a log-normal distribution for central galaxies + a normal distribution for field galaxies (calibrated by Millennium). (BCGs had to be inserted manually in order to match observations, using $P(M_{r,\mathrm{cen}}|M_\mathrm{vir}, z)$, the distribution of central galaxy magnitude for a given halo.) Thus, by construction, AddGals matches the luminosity function and luminosity-dependent 2-pt function.

After passively evolving these magnitudes with redshift, AddGals assigns galaxies with SEDs pulled from SDSS, matching between ADDGALS and SDSS via local galaxy overdensity. Thus, the magnitude bands inherent in the Buzzard flock follow reality closely, statistically speaking.

Since AddGals is ignorant of halos (besides $M_{r,\mathrm{cen}}$), several effects manifest, differing from observations of high-mass halos, including a shallower mass–richness slope, distorted radial profiles, and significantly lower red fractions than observed. Figure 12 of DeRose et al. [2019] shows the depressed richness in Buzzard clusters as compared to observations, lower by a factor of $\lesssim 1.6$ for all measured masses.

As Table 2.1 details, this paper analyzes only halos of mass $M > 10^{13.75}$ $\mathrm{M}_\odot$ in redshift range

$z|[.1, .7)$, using only galaxies with luminosity $L > 0.2\ L_*$ in i-band (calculated following Rykoff et al. [2014]).

2.2 Methods

2.2.1 Modeling the Red Sequence

Galaxies largely fit into two color regions: 1) the lower wavelength RS with primarily older 'red and dead' elliptical galaxies and 2) the higher wavelength BC with primarily younger star-forming spiral galaxies. Intermediate galaxies are classified as GV. This isn't a binary (nor ternary) selection—since the RS and BC overlap [see e.g. Hao et al., 2009], one instead can ascribe to each galaxy a probability P_{red} of belonging to the RS. Cluster richness is then defined as $\lambda \equiv \sum P_{\text{red}}$.

Many studies use a binary, boolean selection for P_{red}, with a hard line drawn in either color–magnitude (CM) or color–color (CC) space to distinguish the RS from the BC [see e.g. Whitaker et al., 2012, Carretero et al., 2015, Adhikari et al., 2020]. Seen in these two different spaces at different redshifts and in different color bands, the RS & BC become more or less distinct. In color space, these two populations appear as Gaussian mixtures, with a tight, densely populated RS component, and a broad BC component.

2.2.1.1 GMM+: tilted Gaussian Mixture in CM space

Since the redness of RS galaxies increases with luminosity (see 2.2, right panel), a fair fit to the RS and BC is a tilted two-Gaussian mixture in CM space. Whereas past studies [e.g. Hao et al., 2009] first performed the mixture in color space and *then* accounted for the RS tilt, I designed a likelihood model that took both into account at once: a Gaussian mixture model (GMM) plus a tilt (thus dubbed GMM+). While this model works well for $z|[.1, .6]$, as I investigated color-magnitude spaces more, I grew unsettled with GMM+ mischaracterization rates (e.g. when viewed in CC space), the discontinuous nature due to shifting color bands, and the non-Gaussianity of galaxy color distribution, especially at higher redshifts.

Neither CM nor CC space present perfect cuts between RS and BC, as shown qualitatively in Figure 2.2. A hard cut in CC space (e.g. a $-45°$ line cutting between the two peaks) couldn't account for the narrowing of the RS with decreasing magnitude, leading to either characterizing bright blue galaxies as part of the RS or a deficit of RS galaxies, with faint red galaxies characterized as BC. A hard cut in CM space doesn't account for the wide spread of the BC beyond the RS lower boundary, with several members even extending to 'redder than red' regions. Additionally, most models for RS characterization have sharp transitions to new color bands at certain redshifts (as

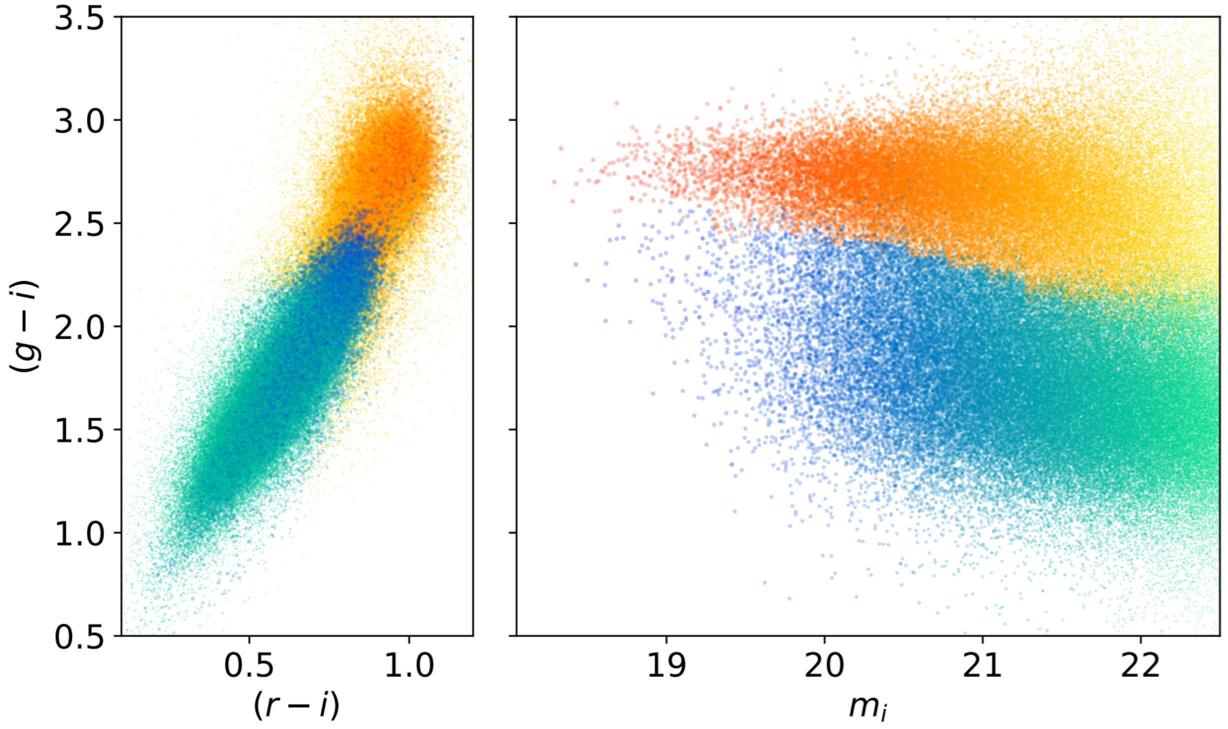


Figure 2.2: Qualitative example of how a simple hard cut in either CC or CM space mischaracterizes galaxies. This fitting used a two-Gaussian mixture model sliced by i-band magnitude to give a continual two-component model across all colors (see equation 2.3) and magnitude. Autumn colors show RS galaxies; winter colors BC galaxies.

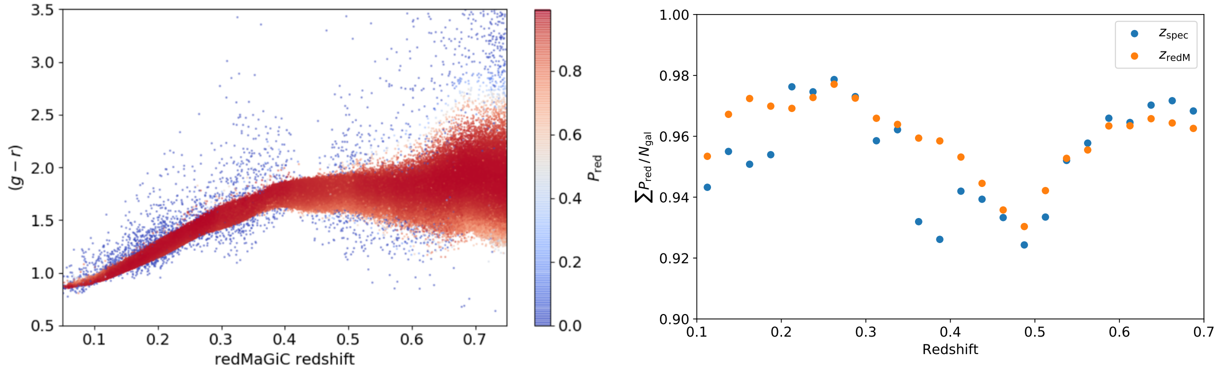


Figure 2.3: Red Dragon operating on the redMaGiC-selected red sequence galaxies. Taken from the redMaGiC Y3A2 High Dens sample. **Left:** $g - r$ color of RS galaxies versus redMaGiC-estimated redshift $z_{\rm redM}$, colored by RD's estimate of $P_{\rm red}$ (note the near-binary selection, i.e. there are very few grey points here). Note the increased scatter at $z \sim .4$, where the 4kÅ break leaves m_g band, requiring comparison with bands $\geq m_i$ or redder to capture this feature. Were this plotted with $g - i$ on the vertical axis, there would be significantly less scatter for $z|[.4, .7]$. **Right:** Total richness $\lambda \equiv \sum P_{\rm red}$ (measured by RD) of redMaGiC-selected RS galaxies, compared to the total number of RS galaxies identified by redMaGiC in a given redshift bin (if RD were identical to redMaGiC, this fraction would be 100%—that all redMaGiC RS galaxies were equally characterized as red by RD). Redshift bins are estimated by redMaGiC $z_{\rm redM}$ and through spectroscopy $z_{\rm spec}$ (with $\Delta z = .025$). Both binning methods show that Red Dragon estimates that 95% of the galaxies considered red by redMaGiC to be true RS members, suggesting that it's more selective about which galaxies to consider red (though still in a similar neighborhood of characterization).

galaxy features such as the 4 kÅ break[1] evolve with redshift, transitioning to different color bands), causing discontinuities in red fraction. To mitigate these issues, one must go beyond CC and CM space.

2.2.1.2 Red Dragon: three-Gaussian mixture in full color space

To circumvent the above mentioned issues, the Red Dragon algorithm interpolates a series of GMMs across redshift bins, following RS, GV,[2] & BC components. This fitting used the full DES color space, derived from its *griz* filters:

$$\vec{c} = \{g - r,\ g - i,\ g - z,\ r - i,\ r - z,\ i - z\}, \tag{2.3}$$

[1] A coincidence of metallic spectral lines around 4 kÅ (implying an older stellar population) along with the Balmer break around 3646 Å (implying Hydrogen ionization) work together to create the most striking feature in galaxy spectra: the 4 kÅ break. This increased opacity below 4 kÅ separates RS galaxies photometrically from BC galaxies [see Dunlop, 2012].

[2] To quote the MICE collaboration: “For our purposes, it is indifferent whether this green sequence corresponds to a physically distinct type of galaxies or just to the inadequacy of the Gaussian distribution to represent the red and blue populations.” [Carretero et al., 2015]

making it a fit in six color (C^6) space. By interpolating between redshift bins, a continuously evolving definition of RS (along with GV and BC) assigns membership probabilities $P_{\rm red}$, $P_{\rm green}$, $P_{\rm blue}$ to galaxies. Thus, richness here is defined as $\lambda = \sum P_{\rm red}(z, \vec{c})$.

Running on the redMaGiC-selected red sequence galaxies (the redMaGiC Y3A2 High Density sample), RD gives them in total $\sim 95\%$ red characterization (see Figure 2.3), implying that RD is more selective than RM, but in a similar neighborhood of characterization.

See figure 2.4 for comparisons between the GMM+ and RD classifications of galaxies. RD is slightly less rich, namely that $\lambda_{\rm RD} \sim 98\% \, \lambda_{\rm GMM+}$ for $\mu \gtrsim 14$. For $\mu \gtrsim 14.2$, the richness–mass relation is significantly tighter for RD than for GMM+, indicating its greater constraining power. It has the additional upside of a continuous definition of red fraction across redshift, unmarred by a transition in bands. This gives a more logical definition of the RS, allowing it to be better compared across a wider spread of redshifts.

The continuous definition across redshifts and the tighter correlation to halo mass with larger halos makes Red Dragon an optimum red sequence selector for this work.

2.2.2 KLLR: Kernel localized linear regression

KLLR [Farahi et al., 2018, Anbajagane et al., 2020] uses a Gaussian kernel to weight points, allowing a continuous definition of mean relations (e.g. Figure 2.8), along with local slope, intercept, and scatter for this analysis.

Using the same Gaussian kernel, one can get a pseudo-count of datapoints used at any plotted spot by summing the weights. This pseudocount $\tilde{N}$ then can be used to calculate a continuous (pseudo) error of the mean $\sigma / \tilde{N}^{1/2}$, as used in the plots to follow. Once the pseudocount falls below one, KLLR is either interpolating between sparse data or extrapolating. Since scatter can't be calculated for $N < 2$, wherever $\tilde{N} < 2$ residual data should be shown as points or treated in aggregate, as a single, wider bin.

2.3 Results

Since these are results for *halos* in the Buzzard Flock (existing in 3D position space: $\{q_x, q_y, q_z\}$), they don't have a bijective comparison to *clusters* observed on the sky (marked in sky coordinates + redshift space: $\{\alpha, \delta\} \oplus \{z\}$). So take these comparisons with a grain of salt.

Note: These results are for Buzzard version 2.0.0, but the results are nearly identical to the previous iteration (version 1.9.8).

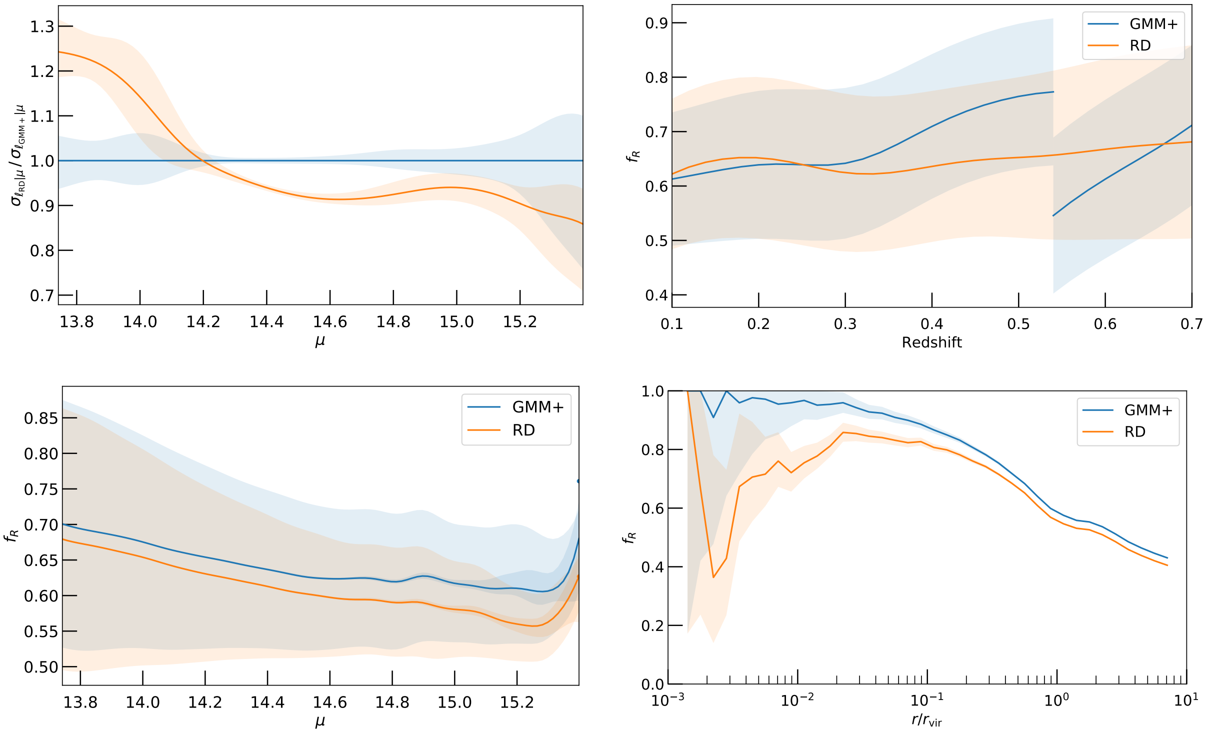


Figure 2.4: Comparison plots between GMM+ and Red Dragon. These are shown as blue and orange lines respectively. **Upper Left**: Scatter in decimal log richness ℓ at fixed mass shown for RD relative to GMM+, with bootstrap error bars. This implies that for high mass halos ($\mu \gtrsim 14.2$), Red Dragon's characterization of the red sequence serves as a better mass proxy than the GMM+ model. **Upper Right**: Evolution of the red fraction with richness, including scatter. The upward trend becomes even *more* apparent at high masses, contradicting Butcher–Oemler Hypothesis for the Buzzard Flock. This showcases the continuous definition of red fraction, as compared to the common mixture model approach. **Lower Left**: Red fraction as a function of mass, showing intrinsic scatter as well as bootstrap errors on the mean. Besides for the few most massive halos ($\mu \gtrsim 15.3$), red fraction *decreases* with increased halo mass, contrary to expectations from e.g. Yang et al. [2008] figure 5, where $f_R(\mu)$ monotonically increases. **Lower Right**: Red fraction as a function of relative halo-centric distance $x \equiv r/r_{\rm vir}$, including scatter. Note that at $r \sim r_{\rm vir}$ both methods unexpectedly dip.

	Weighted mean	Offset ($z = 0.4$)	Slope ($d/d\zeta$)
α	2.2110 ± 0.0062	2.206 ± 0.013	-0.02 ± 0.11
β	1.0155 ± 0.0055	1.0108 ± 0.0071	0.299 ± 0.066
σ^2	0.01963 ± 0.00019	0.02009 ± 0.00065	0.0102 ± 0.0054

Table 2.2: HOD temporal evolution parameters, corresponding to Figure 2.5. Weighted mean found by summing across redshift with Gaussian errors. Parameterized using $\zeta \equiv \log\left[(1+z)/(1+0.4)\right]$.

2.3.1 Mass–richness relation

The kernel connecting mass and richness (see eqn 2.2) serves as a linchpin in optical cluster cosmology. This probability $P(\lambda|M)$ or its inverse $P(M|\lambda)$ are often fit with power laws of form:

$$\langle \log_{10} \lambda | \mu \rangle = \alpha + \beta \log_{10}(M/M_p) \pm \sigma \tag{2.4}$$

(using pivot mass $M_p = 10^{15.5}$ $\mathrm{M}_\odot$ to uncorrelate offset α from evolution of slope β). Temporal evolution of each parameter generally scales with some power of scale factor $a \equiv 1/(1+z)$ or dimensionless Hubble parameter $E(z) \equiv H(z)/H_0$ (relative to some pivot). Parameters here use $\zeta \equiv \ln \frac{1+z}{1+z_p}$, with $z_p = 0.4$ as pivot redshift.

Figure 2.5 shows the HOD's redshift evolution, looking at halos with $\mu \geq 14.2$. This uses a linear regression (in log log space) run by linmix [Kelly, 2007]. Though α and σ show no strong evidence of redshift evolution, β increases significantly with redshift (see Table 2.2 for full evolution). Thus we have log richness normalization $\alpha = 2.211 \pm 0.006$, slope $\beta = (1.016 \pm 0.006) - (0.30 \pm 0.07)\,\zeta$, and intrinsic scatter $\sigma = 0.1403 \pm 0.0007$.

More generally, one can measure how slope and scatter vary with mass, rather than fitting them as constant (as in equation 2.4). Figure 2.6 shows a KLLR fitting of Buzzard halos compared to observations from Bleem et al. [2020]. Since the count of halos decreases dramatically with increased mass,[3] the linmix fit is influenced more by lower-richness halos; hence changing the lower bound on which masses are modeled (as with Figure 2.5, where $\mu \geq 14.2$)) can change the slope significantly. This makes KLLR analysis particularly useful, since modelling the HOD continuously disjoints slope measurements from the lower mass bound.

Of note here in Figure 2.6, at low richness, the sharp edges of Red Dragon emerges: most galaxies are given relatively binary classifications as either RS / GV / BC members, with relatively few galaxies classified in-between. This shows up as the horizontal striations in the low-richness regime.

We see here significant deviation in Buzzard's mean richness across all mass scales, i.e. Buzzard halos not only have too few red galaxies, as notes in DeRose et al. [2019, see Figure 12], but they

[3] E.g.: there are ~ 100 times more $\mu = 13.75$ halos than $\mu = 14.75$ halos.

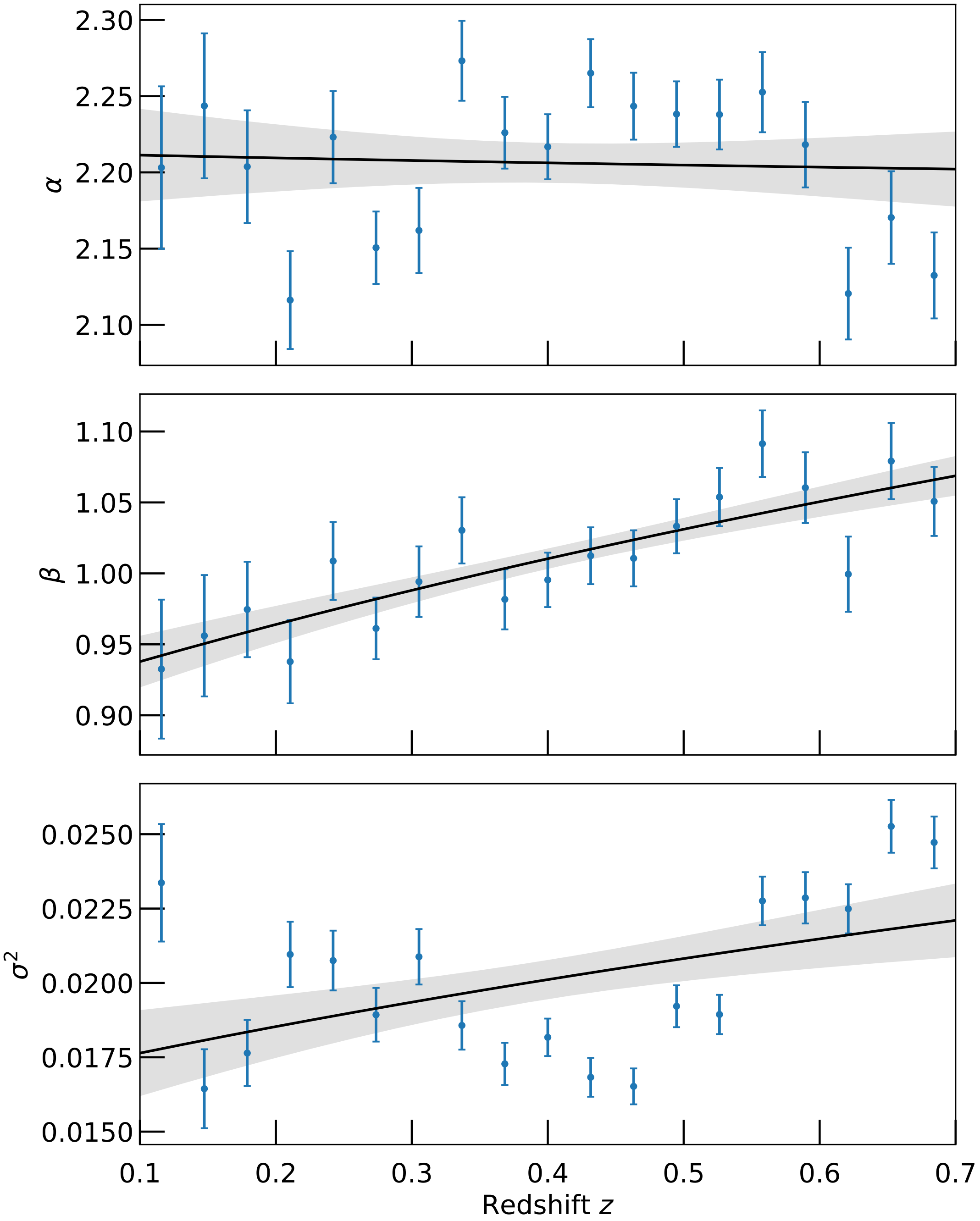


Figure 2.5: Redshift evolution of mass–richness relation. Shown for halos mass $\mu \geq 14.2$, as parameterized by equation 2.4. Constants given in Table 2.2. Fit lines show mean and scatter in the mean as measured by linmix, fit in $\log(1+z)$ space. The offset and scatter show insignificant evolution with redshift, whereas the slope clearly increases with redshift.

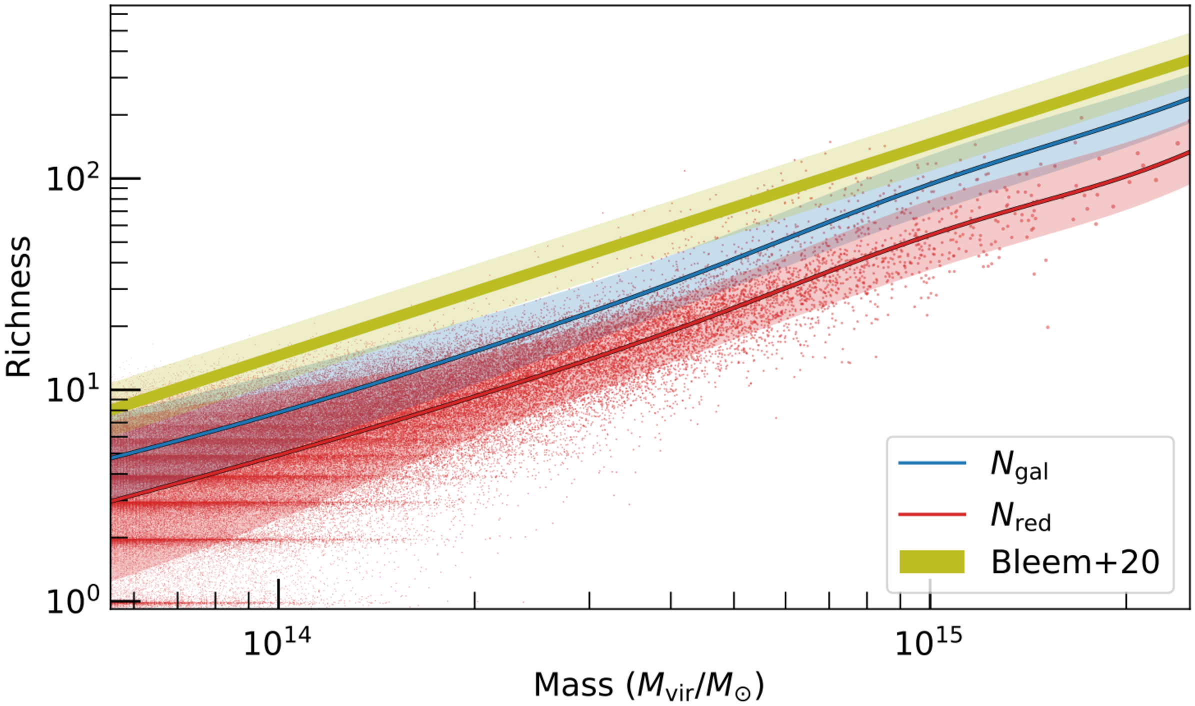


Figure 2.6: HOD for Buzzard, spanning the full range defined in Table 2.1. The yellow fit shows the spread of expectations across redshifts from Bleem et al. [2020]. The red and blue fit show mean and scatter for richness and total galaxy count respectively. Red points indicate richness of individual halos.

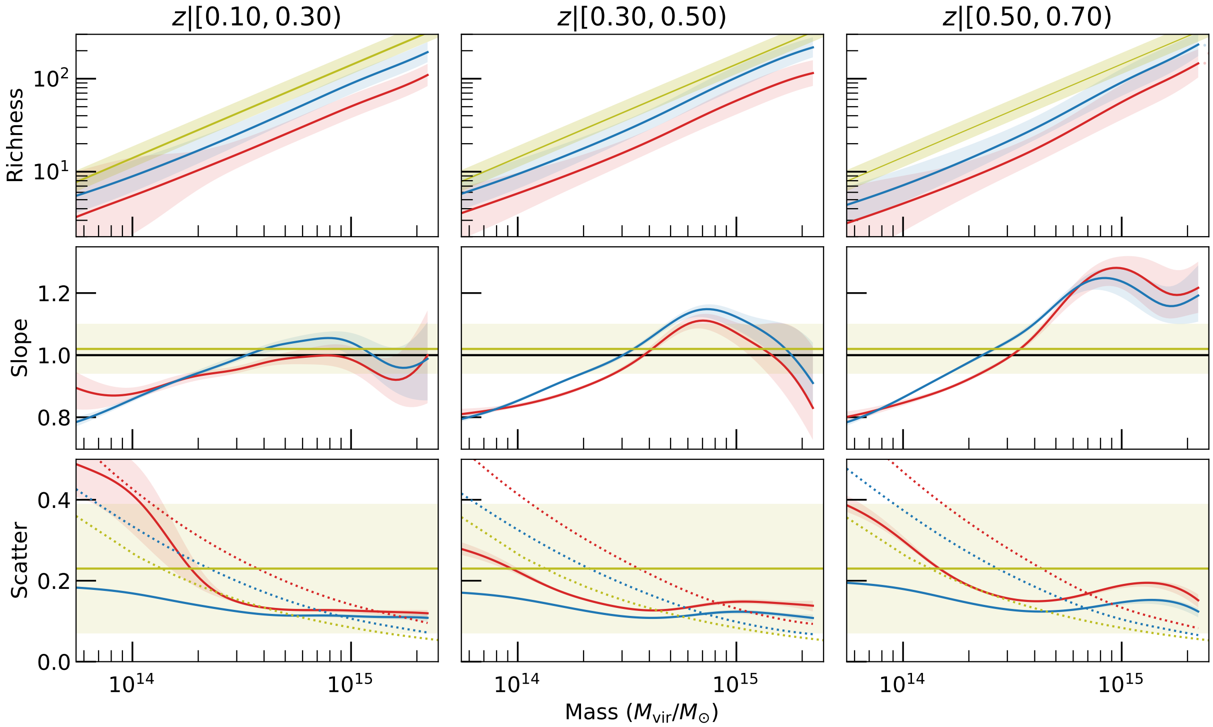


Figure 2.7: KLLR continuous fits in several redshift bins. Color scheme identical to Figure 2.6. Errors in slope and scatter from bootstrap resampling. Dotted lines in the scatter indicate expected Poisson scatter at mean richness or galaxy count: $\langle\lambda\rangle^{-1/2}$ and $\langle N_{\rm gal}\rangle^{-1/2}$ respectively.

also contain too few galaxies in general. Even were every galaxy red in the Buzzard flock, richness values would pale in comparison to observed richnesses.

Though the slope and scatter in Figure 2.6 appear visually similar to expectations from Bleem et al. [2020], a more detailed analysis in thinner redshift bins shows off some of the peculiarities of the Buzzard catalogue.

Whereas observations fail to support significant running of slope with mass, Figure 2.7 indicates otherwise. Buzzard shows a significant increase in mass–richness slope up to $M \sim 7 \times 10^{14}$ $\rm M_\odot$. Beyond that mass, slope behaves differently with redshift: high-mass halos have higher mass–richness slopes at high redshifts than at lower redshifts. DeRose et al. [2019] and Costanzi et al. [2019] find increased scatter in richness with increased mass for simulated and observed clusters, yet in Buzzard halos, we see a significant *decrease* in richness scatter $\sigma_{\ln\lambda|M}$ with mass up to the same mass of $M \sim 7 \times 10^{14}$ $\rm M_\odot$ (somewhat decreasing as Poisson scatter), after which the scatter levels out.

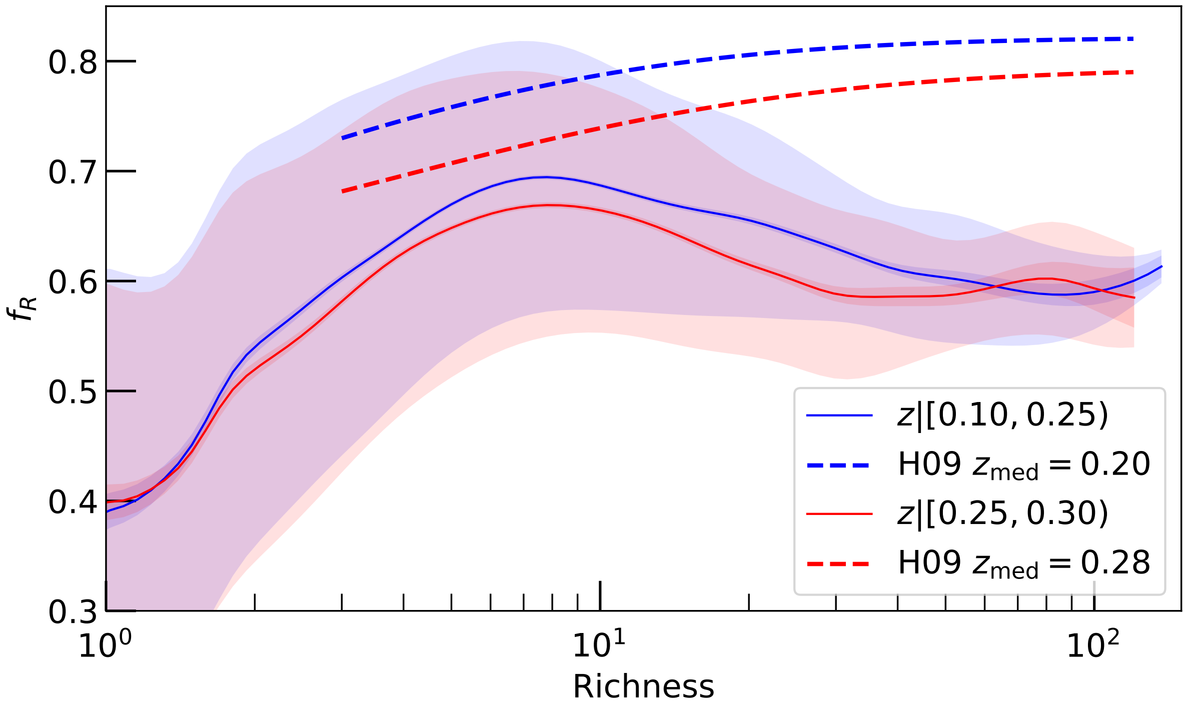


Figure 2.8: Red fraction as a function of richness (solid lines; error of mean and scatter shown as transparent regions), as compared to Hansen et al. [2009] (dashed lines; lower limits). Buzzard's red fraction is significantly low across all richnesses, as compared to observations. (Right-side points indicate residual data too sparse for KLLR analysis.)

2.3.2 Red fraction

Since interloping galaxies are more likely blue field galaxies, projection effects usually lower red fraction. Thus, observed f_R serve as *lower* limits for halo expectations (see Figures 2.8 and 2.11).

2.3.2.1 Relation to halo size

Figure 2.8 plots the same data ranges as Hansen et al. [2009], Figure 11. Including scatter, their data never fall below $f_R = .67$ (and $f_R > .73$ for low redshift), yet Buzzard's mean f_R almost *always* falls below that mark, when f_R ought to consistently be *above* the dashed lines.

Figure 2.9 shows the decreasing trend of red fraction with halo mass, that (to the contrary of observations and simulations—see Donnari et al. [2020]) larger halos in the Buzzard Flock have *lower* red fractions.

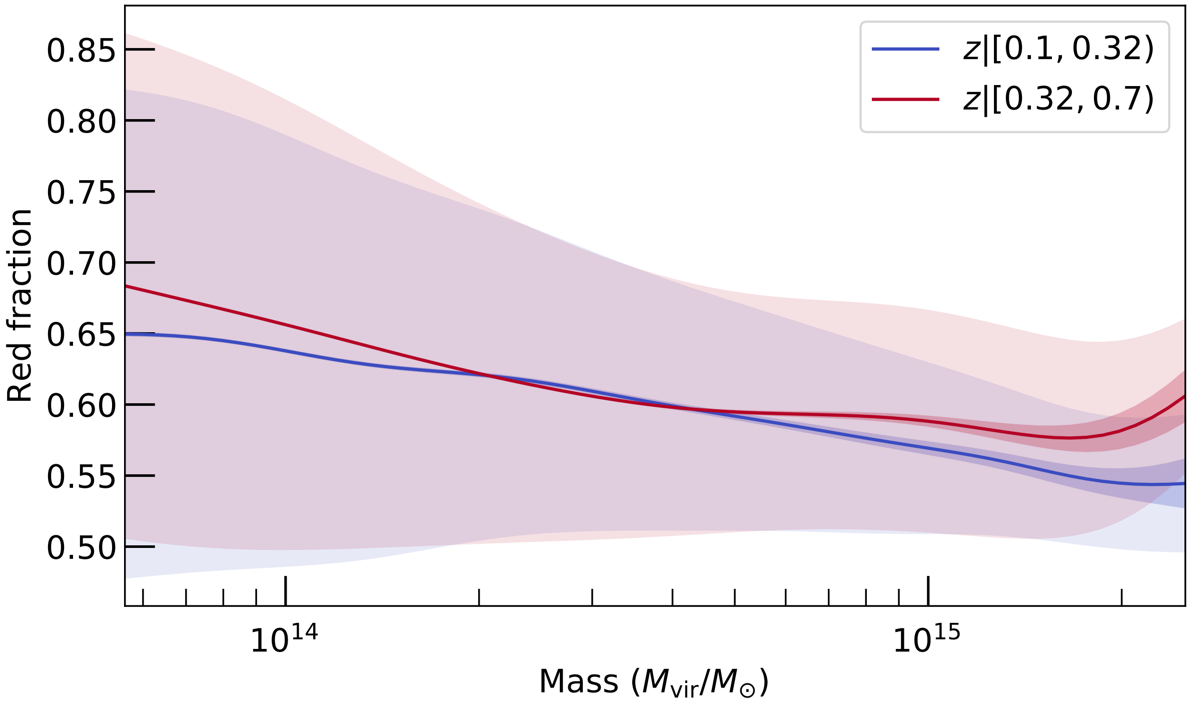


Figure 2.9: Red fraction as a function of mass, binned by redshift. Error of the mean shown in thin transparent regions with scatter shown as larger transparent regions.

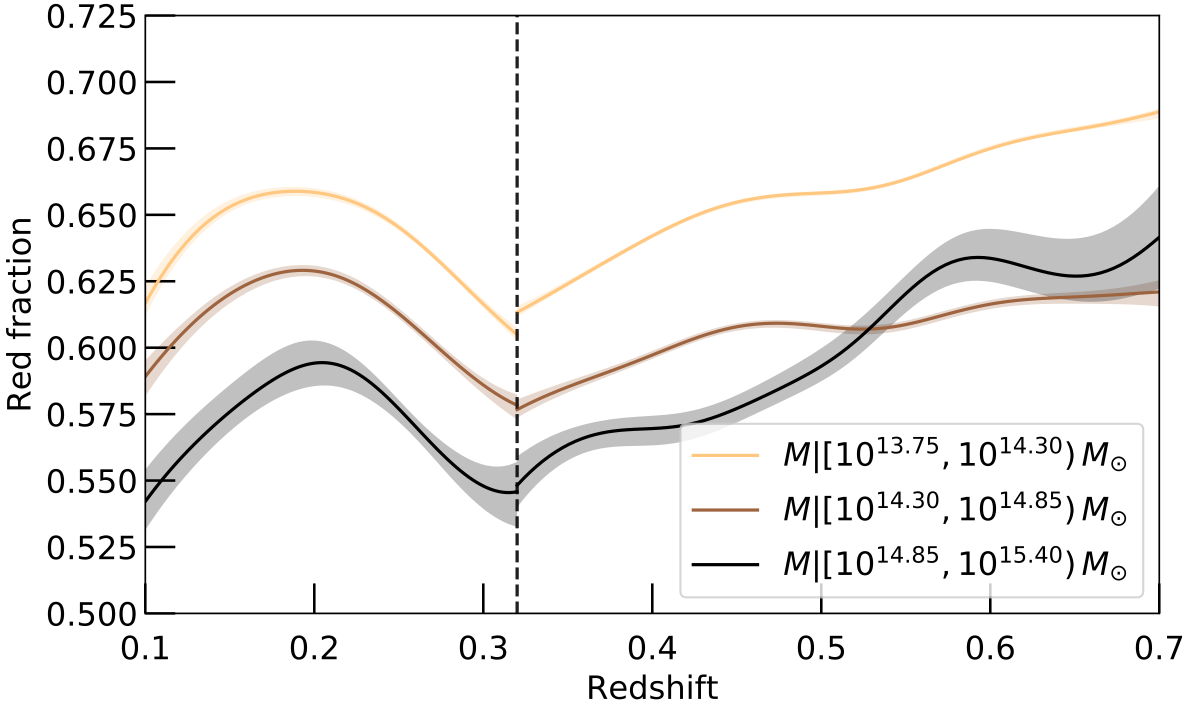


Figure 2.10: Red fraction as a function of redshift, binned by mass. Transparencies show bootstrap error of the mean. The dashed line indicates the break between simulations composing Buzzard, where the larger simulation has poorer resolution.

2.3.2.2 Relation to redshift

The Butcher–Oemler effect [Butcher and Oemler, 1978] suggests that the cores of galaxy clusters redden with time (f_R decreases with z). Though weak, this trend has been observationally confirmed [Nishizawa et al., 2018] and confirmed via simulations [Donnari et al., 2020].

This is expected, in part due to stellar evolution, and in part due to interactions within halos. Stellar evolution will turn stars off the main sequence and onto the red giant branch with time, making the net color redder over time. Violent relaxation and other interactions between galaxies will tend to make galaxies more elliptical and redder. Thus, as redshift increases, red fraction in clusters is expected to decrease.

Figure 2.10 (and similarly Figure 2.9) reveals the opposite: red fraction *increases* with redshift in the Buzzard Flock, especially in the most massive halos. This runs contrary to observations and expectations.

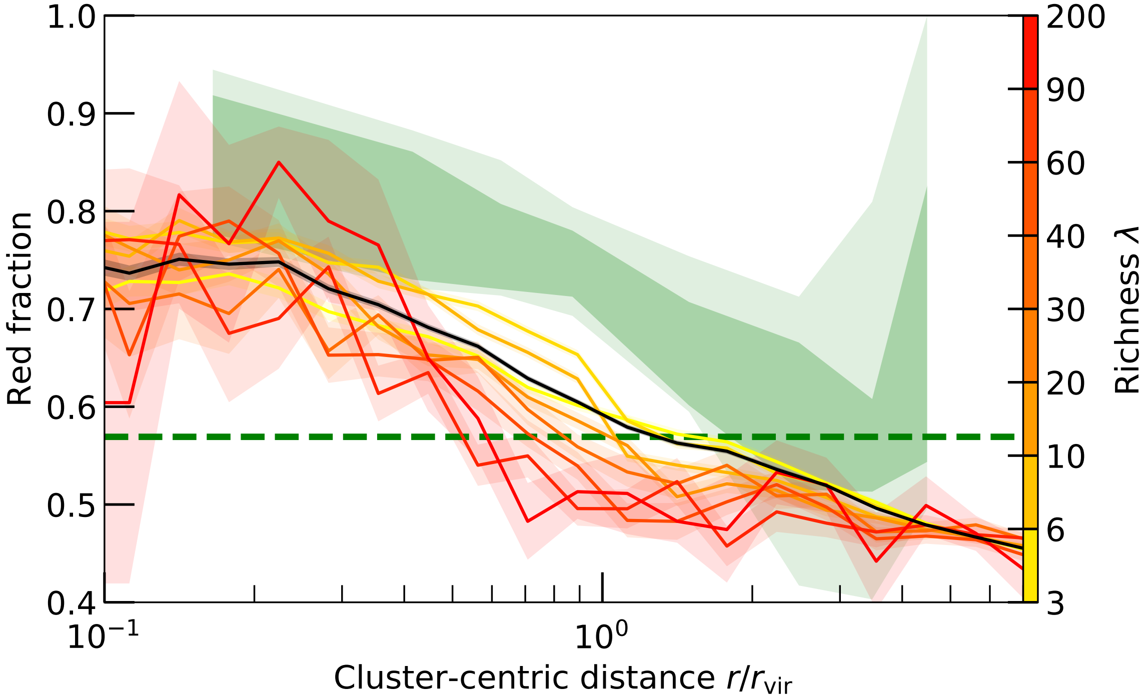


Figure 2.11: Red fraction as a function of cluster-centric radius, measured in similar redshift range and richness bins. Autumn-colored transparent regions indicate error of the mean. Green transparent regions indicate range of means and scatters from Figure 12 of Hansen et al. [2009] while the green dashed line shows the expected mean global f_R ibid. Black line indicates mean across all richness bins, falling about 15% bluer than observations.

2.3.2.3 Relation to radius

The cores of galaxy clusters are known to be redder than their outskirts [Hansen et al., 2009, Nishizawa et al., 2018].

Figure 2.11 shows measured $f_R(r/r_{\mathrm{vir}})$ across several richness bins, matching Hansen et al. [2009]. Most of these fall far below not only the observed mean (darker green) but also the observed scatter (lighter green), indicating a severe deficit of red galaxies. Again we see the opposite trend of red fraction with richness, in that the richest clusters have significantly lower red fractions than the poorer clusters (at least around $r = r_{\mathrm{vir}}$).

Both Figures 2.11 and 2.12 show an unexpectedly sharp decrease in red fraction around $r|(.4, 1)\, r_{\mathrm{vir}}$, especially in the more rich / massive clusters.

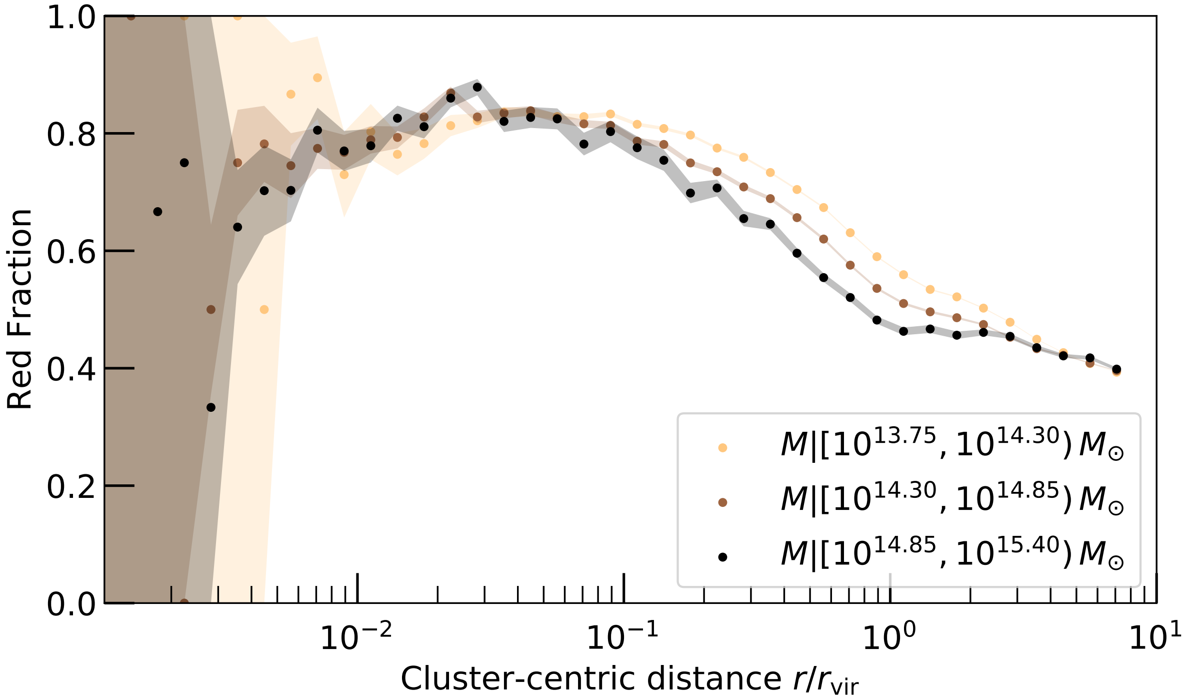


Figure 2.12: Red fraction as a function of cluster-centric radius, measured in mass bins, with error of the mean transparent. Note the dip in f_R near virial radius, increasing with mass.

2.3.3 Radial profiles

Galaxies follow an NFW distribution in clusters [Lin et al., 2004, Hansen et al., 2005, Lin et al., 2017, Nishizawa et al., 2018].

In radial bins relative to each virial radius ($x \equiv r/r_{\rm vir}$), I measured galactic number count and richness ($\lambda \equiv \sum P_{\rm red}$) for each galaxy cluster. This dataset allows for radial number density profiles (for all galaxies, or only for red galaxies), as well as profiles of red fraction ($f_R \equiv N_{\rm gal}/\lambda$).

For each galaxy cluster, I found an expected number of galaxies to be found in each radial bin

$$N_{{\rm exp},i} = \bar{\rho}(z)\, {r_{\rm vir}}^3\, ({x_{i+1}}^3 - {x_i}^3) \tag{2.5}$$

in order to calculate the overdensity

$$\delta = N_i/N_{{\rm exp},i} - 1 \tag{2.6}$$

(where $\bar{\rho}(z)$ is the mean galaxy density; see Figure A.1). Figure 2.13 shows this overdensity as compared to the expected overdensity from an NFW profile $\delta_{\rm NFW}$, giving relative overdensities.

I compare the overdensity of galaxies (as compared to their universal mean density at a given redshift) to the overdensity of dark matter, as compared to the mean density of dark matter. Thus we have

$$\delta_{\rm NFW} \equiv \frac{\rho_{\rm NFW}}{\rho_m} = \frac{1}{\Omega_m(z)} \frac{\Delta_c}{3A} \frac{1}{x(c^{-1}+x)^2} = \frac{\Delta_v}{3A} \frac{1}{x(c^{-1}+x)^2} \tag{2.7}$$

as a baseline overdensity, with concentration-dependent prefactor $A \equiv \left[\ln(1+c) - \frac{c}{1+c}\right]$; virial overdensity relative to *critical* density $\Delta_c \equiv 18\pi^2 + 82x - 39x^2$ (where $x \equiv \Omega_m(z) - 1$ here), along with virial overdensity relative to *mean matter* density $\Delta_v \equiv \Delta_c/\Omega_m(z)$, following the conventions of Hu and Kravtsov [2003]. Concentration is calculated by COLOSSUS [Diemer, 2018], using the $c(z, M_{\rm vir})$ relation from Bhattacharya et al. [2013]. In practice, the different masses and redshifts make a difference in $\delta_{\rm NFW}$ of less than a factor of four across all halos.

In this relative space of $\delta/\delta_{\rm NFW}$, I found the mean and scatter in mass and richness bins, then brought them back into overdensity space using the expected $\delta_{\rm NFW}$ from the mean mass and redshift of a given subsample. (This corrects for some drift in mass and redshift, decreasing scatter in the final relation.) This is the overdensity plotted in Figure 2.13, along with the relative space below.

All masses had a significant underdensity (as compared to NFW) near $r = \frac{1}{10} r_{\rm vir}$ (most papers don't go below this radius). As mass increases, the central overdensity increases from about 100 times below NFW at $\mu \sim 13.75$ to about 100 times above at the highest masses. The combination of these two effects is that at low masses, the profile would be neatly fit by an NFW curve, but at high masses, there's a significant elbow near $r = \frac{1}{10} r_{\rm vir}$, interior to which the higher masses uptick dramatically.

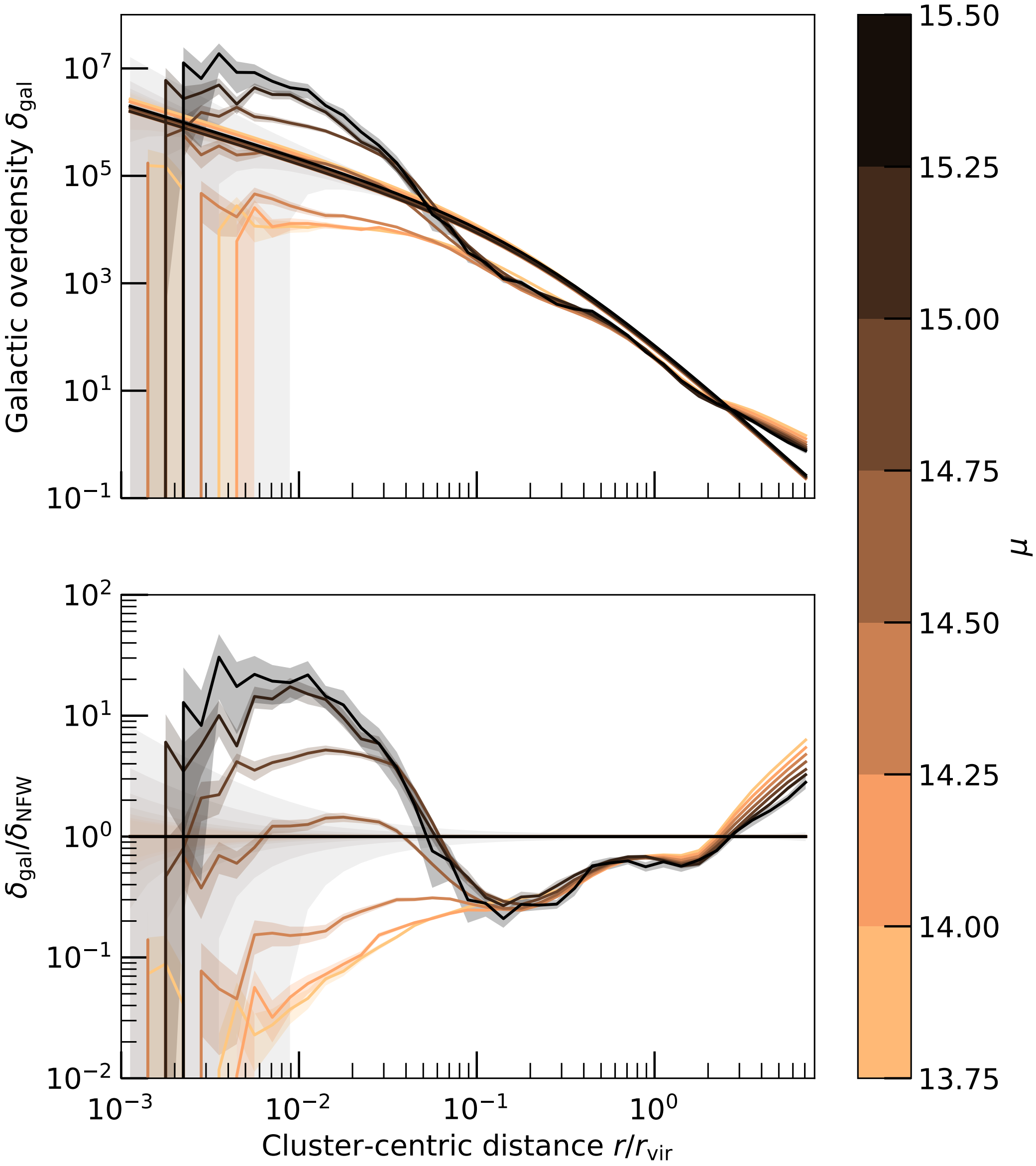


Figure 2.13: Buzzard radial profiles. **Above:** Radial profile of galactic number count overdensities (as compared to mean galactic number density). Plotted transparencies show expected Poisson error shown for NFW profiles and error of the mean for each mass bin. The lifting tail at the right (caused by the two-halo term) matches observations [Hansen et al., 2005, Diemer and Kravtsov, 2014, Nishizawa et al., 2018]. **Below:** Same as above, relative to expected $\delta_{\rm NFW}$, as calculated from the mean (z, μ) of a given bin.

2.4 Chapter Conclusions

In addition to introducing the Red Dragon algorithm for RS selection, this paper gives a detailed analysis of halos in the Buzzard Flock v2.0.0 synthetic galaxy catalog. Several anomalies stood out:

1. In addition to low richness normalization, the HOD exhibited a significantly shallow slope at low masses as well as a high-end slope which increased with mass. Additionally, the HOD's scatter significantly *fell* with increasing mass. (See Figure 2.7.)

2. Contrary to observations, red fraction *decreased* with respect to halo size (see Figures 2.8 & 2.9) and age (see Figure 2.10).

3. Red fraction exhibits a dip in radial distribution near the virial radius, with an intensity scaling with halo size (see Figures 2.11 & 2.12).

4. Radial density profiles exhibit a significant deficit at a tenth virial radius as compared to NFW expectations. Halos also exhibit a mass-scaling increase in core density, varying over nearly four orders of magnitude in overdensity.

As the aphorism goes, all models are wrong, but some are useful. Though these deviations from observations are significant, statistically speaking, they don't make the catalogue unusable. However, these deviations ought to be kept in mind as the catalogs are used going forwards.

Moving forward, we will create reddened versions of the Buzzard catalogs, with $f_R(\mu, z, r/r_{\rm vir})$ better matching observations (as well as a lifted HOD, with fewer high-mass + low-richness systems). These rozen catalogs (dubbed Rose Garden) will provide better statistics for halos, and will additionally allow investigation into the effect of red galaxy content (in halos) on cluster selection and scaling relations.

Future synthetic galaxy catalogs ought to check not only cluster composition (as measured from a projected lightcone), but also the three-space halo composition, as detailed in this work. This will make galaxy catalogs more useful for both large-scale cosmology, but also small-scale cosmology (as well as cluster astrophysics).

CHAPTER 3

Red Dragon: A Redshift-Evolving Gaussian Mixture Model for Galaxies

Chapter Summary

Precision-era optical cluster cosmology calls for a precise definition of the red sequence (RS), consistent across redshift. To this end, we present the Red Dragon algorithm: an error-corrected multivariate Gaussian mixture model (GMM). Simultaneous use of multiple colors and smooth evolution of GMM parameters result in a continuous RS and blue cloud (BC) characterization across redshift, avoiding the discontinuities of red fraction inherent in swapping RS selection colors. Based on a mid-redshift spectroscopic sample of SDSS galaxies, a RS defined by Red Dragon selects quiescent galaxies (low specific star formation rate) with a balanced accuracy of over 90%. This approach to galaxy population assignment gives more natural separations between RS and BC galaxies than hard cuts in color–magnitude or color–color spaces. The Red Dragon algorithm is publicly available at bitbucket.org/wkblack/red-dragon-gamma.

3.1 Introduction

Galaxies cluster not only in physical space, but in color space as well [Strateva et al., 2001, Bell et al., 2004]. The advent of CCD technology revealed a strong dichotomy in galaxy colors: a tightly-packed red sequence (RS; predominantly quiescent, passively evolving ellipticals) and a broader blue cloud (BC; predominantly active, star-forming spiral galaxies) [Bower et al., 1992, Schawinski et al., 2014]. Galaxies that fall between the RS and BC populate the 'green valley' (GV).

Observations of distant protoclusters are now unveiling some of the earliest examples of RS galaxies. MAGAZ3NE J095924+022537 [McConachie et al., 2022], a spectroscopically confirmed protocluster at $z = 3.37$, contains a UVJ-quiescent ultramassive galaxy and the majority of its 38 protocluster members also appear to be quiescent.

Astrophysically, the RS serves as an imperfect proxy for selecting galaxies with low specific star formation rate (sSFR). The distribution of sSFR is skew-lognormal, with a peak of blue active star-forming galaxies at sSFR $\sim 10^{-10}$ yr^{-1} at low redshift (the galactic main sequence) and a tail towards lower sSFR [Wetzel et al., 2012, Eales et al., 2018, Leja et al., 2022]. This form suggests that the sSFR frequency distribution could be modeled as a dual Gaussian mixture. Photometric colors converge at lower sSFRs, such that galaxies with sSFR $\lesssim 10^{-11.3}$ yr^{-1} share approximately the same red color [Eales et al., 2017]. This decreased scatter strengthens the duality between RS and BC, producing a dual Gaussian in photometric color space [see e.g. Baldry et al., 2004, Hao et al., 2009, Krywult and Pollo, 2018]: a narrow component for the low-sSFR RS and a bluer and broader component for the high-sSFR BC.

Galaxy clusters are natural nests for red galaxies. Due to the Gaussian random nature of ΛCDM initial conditions, peaks in density on different scales are coupled, so the earliest galaxies to form preferentially reside in regions that are destined to host clusters of galaxies in the late universe [Springel et al., 2005]. Clusters at redshifts $z \lesssim 1$ thus naturally contain an older galaxy population than that of the low-density field. Stellar populations older than roughly 1 Gyr become almost uniformly red [Conroy and Gunn, 2010], so cluster galaxies on average tend to be redder than those in the field. Gas cooling times increase as protocluster temperatures rise, which reduces cold-phase gas, and supermassive black holes inject turbulent entropy [Voit, 2005, Croton et al., 2006]. These and other plasma processes combine to rapidly shut down star formation in massive galaxies [Donahue and Voit, 2022]. Other dynamical processes that can shut down star formation, or *quench*, a galaxy are more commonly found in proto-cluster environments, including major mergers between galaxies, rapid flyby encounters, and ram-pressure stripping [Moore et al., 1996, Boselli et al., 2022].

The strong red sequence in clusters greatly aids in galaxy cluster selection. Identification of clusters by their RS was first proposed by Gladders and Yee [2000], using the narrow width of the RS to estimate cluster membership with far greater accuracy than that of BC members. The maxBCG algorithm [Koester et al., 2007] further improved cluster selection, using a hard $\pm 2\,\sigma_{\mathrm{RS}}$ cut in photometric color to select clusters. The algorithm defines richness N_{200} as the count of red galaxies within an estimated virial radius R_{200}. This count of virialized galaxies within the cluster serves as a halo mass proxy [Rozo et al., 2009a]. Rozo et al. [2009b] developed an improved richness estimate λ—the sum of RS background-corrected membership probabilities within a given cluster—which also employed a different cluster radius and a Gaussian color filter.

More recent algorithms and surveys have extended the RS's use for cluster cosmology. To further improve the red/blue galaxy distinction, Hao et al. [2009] developed a single-color error-corrected Gaussian Mixture Model (ECGMM) in color-magnitude space. As compared to a typical

GMM, their ECGMM accounted for photometric errors contributing to the scatter. As a rule of thumb, while the photometric error remains smaller or of similar order of magnitude as the intrinsic scatter, one can still accurately and precisely tease out the intrinsic scatter from a noisy distribution. Around this time, the first results from the SpARCS survey [The *Spitzer* Adaptation of the Red-sequence Cluster Survey; Wilson et al., 2009, Muzzin et al., 2009] produced hundreds of $z > 1$ cluster candidates using a selection method similar to that of Gladders and Yee [2000]. To analyze DES photometry, Rykoff et al. [2014] refined the redMaPPer algorithm, which selects RS galaxies using an error-corrected multi-color × magnitude space Gaussian fitting of the RS. This algorithm iteratively selects the RS, measuring a slope of its color as a function of z band, giving a redshift-continuous, multi-color update to richness. Building on similar methodology, Rozo et al. [2016] introduced the redMaGiC algorithm to select luminous red galaxies, estimating galactic redshifts with high accuracy. These methods serve as a basis for DES cluster finding in cosmological analyses [Rykoff et al., 2016, Abbott et al., 2020], with the richness indicator λ serving as a mass proxy.

Other groups have also modeled the RS as a function of redshift and magnitude for multiple colors. Klein et al. [2018, 2019] created a methodology similar to redMaPPer, though with additionally allowing for evolution in scatter with magnitude. Another cluster finder, CAMIRA, uses stellar population synthesis (SPS) models to estimate RS mean color at a given redshift and magnitude, accounting for the RS color-magnitude (CM) slope via a mass–metallicity relation [Oguri, 2014]. They then calibrate their color model by measuring a fine tuning off of that SPS model. Both of these methods lack color-correlation terms and only characterize the RS—ignoring the BC. Multi-color selection of the RS is crucial to a continuous definition of the population across wide redshift ranges.

Quenching a galaxy induces a feature in its spectrum known as "the 4000Å break": a sharp drop in spectral intensity at short wavelengths, reflecting the absence of young blue massive stars. To measure the break strength (D_{4000}) well a given redshift, one needs measurements from bands on either side of the break's wavelength at that redshift. As the D_{4000} location shifts to redder bands with increasing redshift, redder colors are necessary for identifying more distant RS galaxies.

If a RS selector uses only one color at a time, with discrete swaps between colors at transition redshifts, these hard transitions can result in an $O(10\%)$ discontinuity in red fraction, $f_R(z)$ [up to $\sim 16\%$; see e.g. Nishizawa et al., 2018]. This jolt in red fraction would echo in single-color richness estimates $\lambda_{\rm col}$: Two identical clusters on either side of a redshift transition could then have significantly different $\lambda_{\rm col}$ values, introducing non-trivial systematic errors in halo mass–richness scaling relations. Evolving a multi-color Gaussian mixture across redshift smoothly defines the RS, obviating the discontinuities caused by color swapping.

To better model RS and BC, we present **Red Dragon**: a multivariate Gaussian mixture model,

Table 3.1: Characteristics of the datasets used in this analysis. The galaxy pasting algorithm ADDGALS created Buzzard's synthetic galaxy catalog, assigning SDSS-like galaxies to underlying N-body lightcone outputs of dark matter structure [Wechsler et al., 2021].

Dataset	Type	Redshift	sSFR	$N_{\rm gal}$
SDSS/low-z	Observation	0.1 ± 0.005	Yes	44 452
SDSS/mid-z	Observation	$(.3, .5)$	Yes	90 609
Buzzard (DES)	Synthetic	$[0.05, 0.84]$	No	91 004 552

Table 3.2: Approximate redshift range over which each photometric band contains the 4kÅ rest wavelength, for SDSS [Doi et al., 2010] and DES [Abbott et al., 2018] filters.

band	$z_{\rm break,SDSS}$	$z_{\rm break,DES}$
g	$[0.0, 0.36)$	$[0.0, 0.38)$
r	$[0.36, 0.71)$	$[0.38, 0.78)$
i	$[0.71, 1.06)$	$[0.78, 1.13)$
z	$[1.06, 1.37)$	$[1.13, 1.50)$
Y	N/A	$[1.38, 1.55)$

smoothly evolving population characterization across redshift. Using multicolor information, Red Dragon gives continuous and consistent population definitions (and therefore red fractions) across redshift, characterizing well the underlying photometric distribution of galaxies.

We begin in Section 3.2 by introducing the datasets used in this analysis. Section 3.3 then details our algorithm. Then we demonstrate results of its application to real and synthetic galaxy catalogs in Section 3.4, and discuss several considerations for applying this algorithm in Section 3.5. We summarize main takeaways from the paper in Section 3.6.

3.2 Data

We analyze galaxies from the two sources listed in Table 3.1: local observed galaxies from SDSS [at low and mid redshifts; Szalay et al., 2002] as well as a wide-redshift sample from the Buzzard Flock synthetic galaxy catalog, designed to support science analysis of the multi-band photometric Dark Energy Survey (DES) [DeRose et al., 2019, 2021, Wechsler et al., 2021]. In all galaxy samples, we consider only galaxies with $L_i(z) > 0.2\, L_{*,i}(z)$, using the i-band characteristic luminosity as a function of redshift $L_{*,i}(z)$ as defined in Rykoff et al. [2014]. Note that neither SDSS sample is complete to this limit; while the low-z sample is nearly complete (only $\lesssim 0.5$ mag short), the mid-z sample is strongly biased towards bright red galaxies. These samples offer complementary tests of Red Dragon's ability to identify the RS.

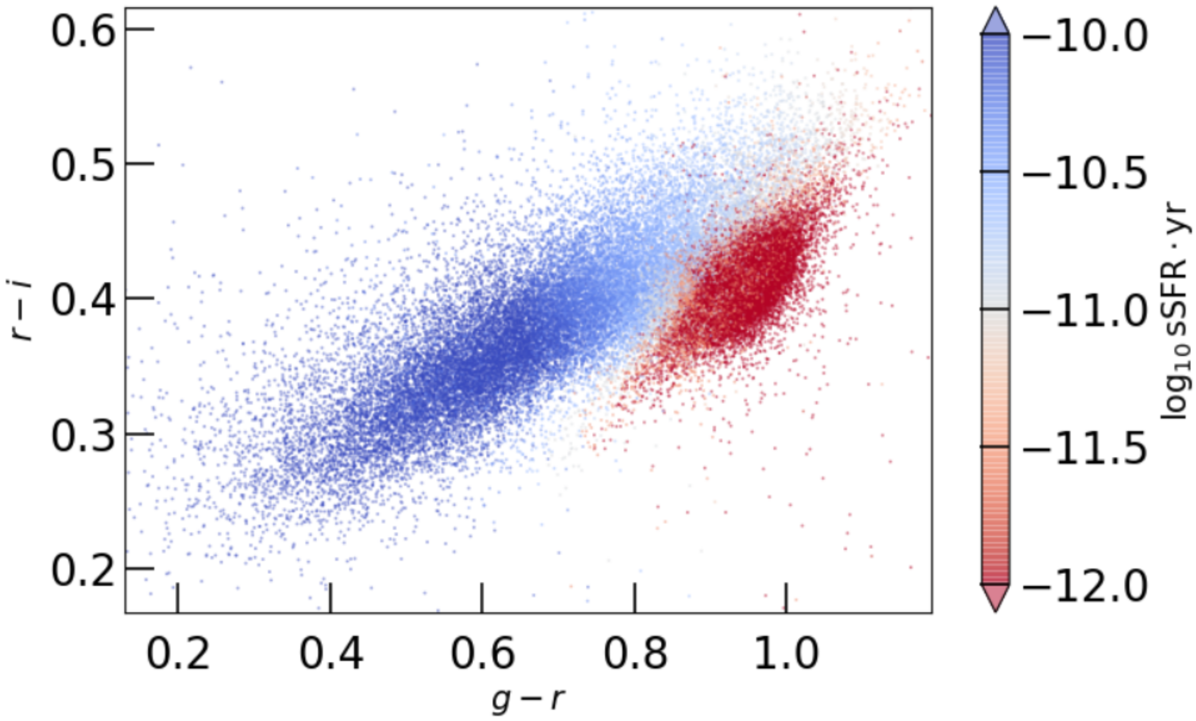


Figure 3.1: SDSS low-redshift galaxy sample ($z = 0.1 \pm 0.005$) shown in the space of two "primary" colors. The $g - r$ color measures the strength of the 4000 Å break (D_{4000}), sensitive to current SFR. The $r - i$ color measures the post-break spectrum slope, an indicator of dust content (degenerate with age and metallicity). Points are colored by log(sSFR/yr^{-1}) (see legend), with values below roughly -11 corresponding to the quenched population.

3.2.1 SDSS

SDSS galaxies were selected from the spectroscopic sample in order to obtain precise redshift estimates and information on quenched status. We choose two complementary samples: a nearby, narrow redshift sample with wide magnitude coverage and a more distant, bright galaxy sample spanning the first D_{4000} filter transition. The thin slice at $z = 0.1 \pm 0.005$ reduces the effect of redshift drift on population colors, and lies far from the $g \to r$ transition of the 4000 Å break's observed wavelength. The wider range of mid-redshift galaxies, $z|(0.3, 0.5)$, spans the $g \to r$ transition, allowing evaluation of RS selection continuity. The two datasets contain similarly-sized galaxy populations: roughly 45,000 and 90,000 galaxies respectively.

For the low-redshift $z = 0.1 \pm .005$ sample[1], redshift errors were typically $\lesssim 10^{-4}$. We limit summed photometric error to be below 0.3 mag to exclude galaxies with poor photometry. Figure 3.1 displays the low-z galaxy sample in $g - r$ vs $r - i$ space, colored by sSFR. One can readily distinguish the two populations, corresponding to star-forming and quiescent galaxies. This

[1] SDSS/low-z sample extracted from SDSS SkyServer with this SQL script.

sample contains about 45% red fraction.

The other SDSS galaxy sample[2], spanning redshifts 0.3 to 0.5, has typical redshift error $\lesssim 10^{-3}$. Spectroscopic selection requirements produce a much higher red fraction of $\sim$ 90% in this sample, but our focus on redshift continuity is insensitive to this enhanced RS selection.

Specific star formation rates for each SDSS sample were calculated using methods from Conroy et al. [2009] and are employed as a truth label to test against Red Dragon's selection of quenched galaxies. These values are based on SED fits to $ugriz$ photometry. Their model assumes a smooth, parametric star formation history for their estimates. Because it does not include spectral lines or indices, nor any NUV or NIR photometry, there is significant systematic uncertainty in the derived sSFR values.

3.2.2 Buzzard

We also investigate the Buzzard synthetic galaxy catalog. Buzzard is a statistical replica of a deep-wide galaxy survey built from galaxy color distributions measured as a function of local cosmic overdensity [Hogg et al., 2004]. It is empirically tuned, mimicking the DES photometric pipeline, with similar photometry, number counts, and photometric uncertainties out to high redshift. We use its two highest-resolution sections, extending our redshift range out to $z = 0.84$ with a magnitude-complete sample. This selection uses distance-based (cosmological) redshift. This then crosses both the $g \to r$ and $r \to i$ transitions of Table 3.2, testing Red Dragon's ability to continuously characterize populations across multiple transitions. Magnitude errors are assigned to each galaxy at a level consistent with DES photometric errors, and photometric redshifts derived from these noisy value.

To create the Buzzard galaxy catalog, the ADDGALS algorithm [Busha and Wechsler, 2008, Wechsler et al., 2021] populates lightcone outputs of N-body simulations with galaxies. The empirical method introduces galaxy bias using a local dark matter density measure. Colors are tuned to SDSS and other observed galaxy samples; ADDGALS trains empirically at low redshifts, then extrapolates to higher redshifts using a spectral energy distribution template approach [for more details, see Wechsler et al., 2021]. While the method reproduces well the magnitude counts and two-point clustering statistics of individual galaxies [DeRose et al., 2021], massive clusters are somewhat underpopulated as compared to observations, and behaviors of the Buzzard universe at higher redshifts are less rooted in observation than those at $z < 1$.

The Buzzard Flock has been central to validating DES galaxy clustering analysis [DeRose et al., 2022], providing important quality assurance for precise cosmological constraints from 3x2 clustering signals [Abbott et al., 2020, Amon et al., 2022, Secco et al., 2022, Abbott et al., 2022].

[2] SDSS/mid-z sample extracted from SDSS SkyServer with this SQL script.

The synthetic catalog is not an exact replica of the DES sky, of course, but it provides a useful testbed for our method. We intend to report on application of Red Dragon to DES-Y3 photometry in a companion paper (Black *et al.*, in prep.).

3.3 Methods: Algorithm Construction

Red Dragon is a novel method for calculating red sequence membership probabilities. In its most general construction, a Red Dragon RS selector uses a Gaussian mixture in multi-color space to select populations of galaxies (RS, BC, and optionally additional components). In this section, we outline the algorithm (§3.3.1), introduce the core likelihood function for Red Dragon (§3.3.2), detail interpolation of GMM parameters across redshift (§3.3.3), and define membership probability assignment (§3.3.4).

3.3.1 Algorithm Overview

There are two broad stages to the Red Dragon algorithm, illustrated in Figure 3.2. In the first stage, galaxy colors are modeled within discrete redshift shells, yielding mixture component information at these redshifts[3] and finds GMM parameterizations for each slice (steps 1–4). In the second stage, the algorithm matches components across redshift bins and interpolates mixture model parameters (steps 4–7). The end result is a continuous and consistent definition of galaxy populations across redshift.

3.3.1.1 First stage: redshift-discrete fits.

Red Dragon reads in photometry and redshift information supplied by the user. Using input redshift estimates $z \pm \delta_z$ (Gaussian uncertainties), Red Dragon randomly samples (with replacement) galaxies within each user-defined redshift bin. From input magnitudes $\vec{m}$ and magnitude errors $\vec{\delta}_m$, Red Dragon calculates primary color vector $\vec{c}$ and its corresponding noise covariance matrix Δ (detailed below). This set of input variables is then sent to a GMM to find fit parameters $\vec{\theta}_\alpha$ for each component α in each redshift bin. Section 3.3.2 offer details of the likelihood model parameters and fitting process.

At this point, the components across redshift are unordered, meaning the $\alpha = 0$ component in redshift bin 1 may not correspond to the $\alpha = 0$ component in redshift bin 2. This set of Gaussian parameterizations must be linked across redshift to form continuous representations of each mixture component.

[3] Though we describe fits as functions of redshift here, any secondary variable may be used (given a thin enough redshift extent for the data), such as stellar mass or a single photometric band.

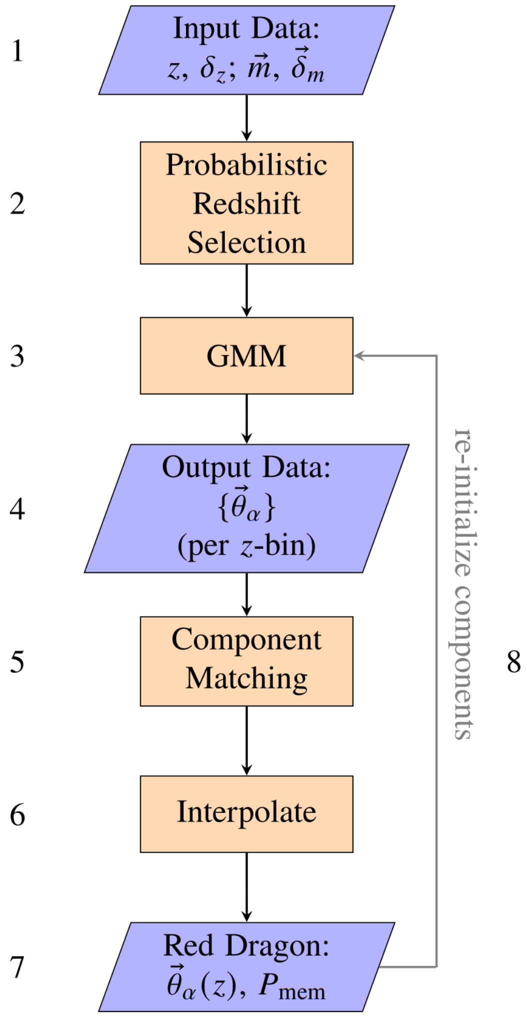

1
Input Data:
$z,\ \delta_z;\ \vec{m},\ \vec{\delta}_m$
2
Probabilistic Redshift Selection
3
GMM
4
Output Data:
$\{\vec{\theta}_\alpha\}$
(per z-bin)
5
Component Matching
6
Interpolate
7
Red Dragon:
$\vec{\theta}_\alpha(z),\ P_{\rm mem}$
re-initialize components
8

Figure 3.2: Work flow for Red Dragon algorithm. In steps 1–4, GMMs fit input data, yielding parameterizations $\{\vec{\theta}_\alpha\}_i$ at each redshift slice i for each component α. Steps 4–7 match components across redshift from the disarranged collection $\{\vec{\theta}_\alpha\}$ and then interpolate, yielding continuous parameterization of GMM components $\vec{\theta}_\alpha(z)$ and thereby membership probabilities $P_{\text{mem},\alpha}$. The grey arrow (8) indicates that a trained dragon may be used to initialize components for a new one. By default, the algorithm fits sparsely with an `sklearn` GMM (1–7), then uses that rough fit to inform initial conditions for a `pyGMMis` GMM (8), which gives a final, error-cognizant fitting.

Table 3.3: Variables of the likelihood model, equations (3.6) and (3.7). Left column shows GMM model parameters θ_α characterizing component α (one of K); right column shows input photometry data for galaxy j (one of N_{gal}). Galaxy color and noise covariance are calculated from input magnitudes and magnitude errors (see text for details).

θ_α	model parameters	x_j	galaxy data
w_α	weight (where $\sum_\alpha w_\alpha = 1$)	$\vec{c}_j$	color
$\vec{\mu}_\alpha$	mean color	$\vec{\delta}_j$	errors on galaxy colors
Σ_α	intrinsic covariance	Δ_j	noise covariance

3.3.1.2 Second stage: redshift-continuous fits.

Using the calculated set of Gaussian mixtures, the algorithm now matches similar Gaussian components across redshift bins. While distinguishing continuous components across redshift is relatively straightforward for two-component models ($K = 2$), matching for even three components ($K = 3$) can be challenging. (Does the wide, redder portion of the BC connect to the narrower but similar in color component in the adjacent redshift bin, or does it connect to the component with similar scatter despite its bluer mean color?) Despite these challenges, the matching process can be largely automated by using relative location and extent information in color space, with quality assurance checks done by the user. As detailed in Section 3.3.3, the interpolated components present the user with a trained dragon, smoothly characterizing each of the populations across redshift.

3.3.2 Likelihood model

Red Dragon employs a multi-dimensional Gaussian mixture model with parameters listed in the left column of Table 3.3. The right column lists the input galaxy photometric information.

A Gaussian mixture of K components in a single color c has a set of parameters θ that constitute the model. For each component α, this parameter set includes the component weight w_α, mean color μ_α, and the intrinsic population scatter σ_α. The weights are normalized such that $\sum_\alpha w_\alpha = 1$. For a set of N_{gal} galaxies with input colors $\{c_j\}$ and color errors $\{\delta_j\}$ for each galaxy j, the model parameters maximize the likelihood

$$\mathcal{L}(\theta|x) = \prod_{j=1}^{N_{\text{gal}}} \left\{ \sum_{\alpha=1}^{K} \frac{w_\alpha}{\sqrt{2\pi({\sigma_\alpha}^2 + {\delta_j}^2)}} \exp\left[-\frac{1}{2} \frac{(c_j - \mu_\alpha)^2}{{\sigma_\alpha}^2 + {\delta_j}^2} \right] \right\}. \tag{3.1}$$

This type of error-corrected Gaussian mixture model (ECGMM) was introduced by Hao et al. [2009] with SDSS $g - r$ as the color classifier.

Expanding this model into an N-dimensional color space requires that we employ for each component α an intrinsic color covariance matrix Σ_α. The errors then must be handled as a noise covariance matrix Δ_j for each galaxy.

Consider the DES four-band optical *griz* photometry used in Buzzard, with input magnitudes $\vec{m} = [m_g,\, m_r,\, m_i,\, m_z]$. We define a vector of **primary colors** based on neighboring photometric bands:

$$\vec{c} = [g - r,\, r - i,\, i - z]. \tag{3.2}$$

Colors are derived from magnitudes by the matrix operation $\vec{c} = A\,\vec{m}$, where the transform matrix is

$$A = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}. \tag{3.3}$$

We assume that the photometric errors of each galaxy are determined independently in each band, and so take them to be uncorrelated.[4] The magnitude error covariance matrix M_j for each galaxy is then diagonal. Transformed to the space of primary colors, the noise covariance of galaxy j is then

$$\Delta_j = A\,M_j\,A^\top = \begin{bmatrix} {\delta_g}^2 + {\delta_r}^2 & -{\delta_r}^2 & 0 \\ -{\delta_r}^2 & {\delta_r}^2 + {\delta_i}^2 & -{\delta_i}^2 \\ 0 & -{\delta_i}^2 & {\delta_i}^2 + {\delta_z}^2 \end{bmatrix}_j \tag{3.4}$$

where each δ_x above refers to the photometric error of band x. Note that for this matrix to be non-singular, the selection of colors must be linearly independent (e.g. one cannot use each of $g-r$, $r-i$, and $g-i$ in an error-inclusive model). For symmetry, simplicity, and to avoid singularity, we employ the set of primary colors.

The derivation of Δ_j is similar for SDSS photometry (which includes u band). The primary color vector for *ugriz* is then

$$\vec{c} = [u - g,\, g - r,\, r - i,\, i - z] \tag{3.5}$$

and the corresponding Δ_j matrices come from a straightforward extension of the above matrices.

Let x_j include the primary colors $\vec{c}_j$ as well as the noise covariance matrix Δ_j for each galaxy (Table 3.3). The likelihood of the error-cognizant N-dimensional Gaussian mixture is then the sum over galaxies and components

$$\mathcal{L}(\theta|x) = \prod_{j=1}^{N_\text{gal}} \sum_{\alpha=1}^{K} \mathcal{L}_\alpha(\theta_\alpha|x_j) \tag{3.6}$$

[4] In hindsight, this turns out to be a bad assumption for drift-scan surveys such as SDSS, where atmospheric noise correlates highly between neighboring bands, potentially causing strong off-diagonal terms in M_j.

where the specific component likelihood for galaxy j is

$$\begin{aligned}\mathcal{L}_\alpha(\theta_\alpha \big| x_j) = &\frac{w_\alpha}{\sqrt{(2\pi)^N} \left|\Sigma_\alpha + \Delta_j\right|} \\ &\times \exp\left[-\frac{1}{2}(\vec{c}_j - \vec{\mu}_\alpha)^{\mathrm{T}} (\Sigma_\alpha + \Delta_j)^{-1} (\vec{c}_j - \vec{\mu}_\alpha)\right] .\end{aligned} \tag{3.7}$$

At individual redshift slices, we use the error-inclusive Gaussian Mixture package pyGMMis [Melchior and Goulding, 2018] to find best-fit parameters θ_α for each component α. Without a reasonable input for a first guess at parameters, pyGMMis sometimes struggles to properly characterize populations. To provide a rough first guess, we first sparsely fit the data using `sklearn`'s error ignorant GaussianMixture package [Pedregosa et al., 2011]. This extremely quick fit gives a rough initial guess to the fit parameters, yielding better results than running `pyGMMis` blind.

3.3.3 Fit interpolation

Red Dragon interpolates best-fit parameters across redshift bins, continuously defining populations. Fitting is linear by default (with flat endpoint extrapolation), but other methods such as smoothed spline interpolation [SciPy: Virtanen et al., 2020] or kernel-localized linear regression [KLLR: Farahi et al., 2022] are available to give smoother fits. After fitting weights, the normalization $\sum_\alpha w_\alpha(z) = 1$ is re-enforced. To interpolate the covariance matrix, log variances are interpolated first, followed by interpolating the correlations (enforcing correlation magnitude $|\rho| \leq 1$), which together then provide a better fit than fitting the covariance matrix values directly (which could result in unphysical negative variances or correlations larger than unit magnitude).

3.3.4 Galaxy Classification

These redshift-continuous fits can then be used to classify galaxies based on its membership likelihood for each component. The probability that galaxy j is a member of GMM component α is

$$P_\alpha(x_j) = \frac{\mathcal{L}_\alpha(\theta_\alpha \big| x_j)}{\sum_\beta \mathcal{L}_\beta(\theta_\beta \big| x_j)}. \tag{3.8}$$

For example, a two-component model would then have red sequence membership probability $P_{\mathrm{RS}} = \mathcal{L}_{\mathrm{RS}} / (\mathcal{L}_{\mathrm{RS}} + \mathcal{L}_{\mathrm{BC}})$. Classification can be based on the maximum value among components or some chosen threshold. We show in Section 3.5.2 that maximum RS membership probabilities are strongly peaked near 1 when $K = 2$ or 3, declining slightly with $K = 4$.

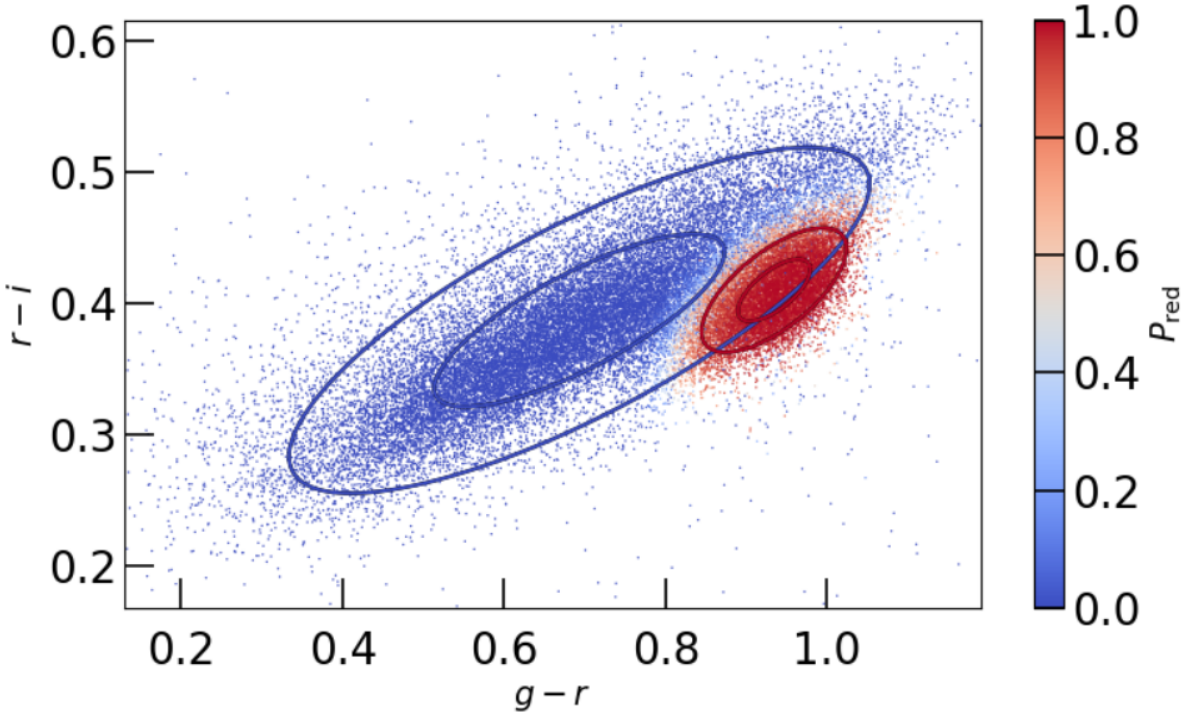


Figure 3.3: Red sequence membership probabilities P_{red}, calculated by Red Dragon for SDSS/low-z galaxies shown in Figure 3.1, here using $g-r$ as the x-axis color. Ellipses show in this projected primary color plane the two Red Dragon components, RS (red) and BC (blue), with a relative offset of mean colors and differing covariance structures.

3.4 Results

Here we show results of running Red Dragon on SDSS and Buzzard datasets. Our SDSS analysis highlights the accuracy of selecting the quenched population and RS continuity across the first D_{4000} filter transition. The Buzzard analysis focuses on fit parameter evolution across a wider range of redshift, derived from a much larger galaxy sample.

3.4.1 Sloan analysis

Fits to SDSS datasets used two-component Red Dragon models, analyzing the four primary colors derived from SDSS *ugriz* photometry in equation (3.5).

Figure 3.3 shows galaxy classification for the low-z data. Ellipses indicate the mean colors and covariance of the RS and BC components (red and blue, respectively) in the projected space of $g-r \times r-i$. The location and extent of the RS overlaps well with the low-sSFR region displayed in Figure 3.1.

The following sections quantify this level of agreement for both the low-z (§3.4.1.1) and mid-z

(§3.4.1.3) SDSS samples.

3.4.1.1 Balanced Accuracy

To quantify Red Dragon's success in identifying the quenched population, we use the binary classification measure of *balanced accuracy*. Balanced accuracy (bACC) averages sensitivity and specificity, weighing equally the true positive rate TPR≡TP/(TP+FN) and the true negative rate TNR≡TN/(TN+FP). As the prediction, we use RS membership probability $P_{\mathrm{RS}} > 0.5$. As the truth label, we define the quenched population as galaxies with sufficiently low sSFR values, setting a threshold as a function of redshift

$$\log_{10}\left(\mathrm{sSFR}\cdot\mathrm{yr}\right) < -11 + z \tag{3.9}$$

[adapted from Moustakas et al., 2013]. The bACC gives us a straightforward metric for scoring various methods of selecting the RS.

Because the distribution of sSFR values is skew-lognormal rather than truly bimodal (as photometric colors, D_{4000}, the Sérsic index [Krywult and Pollo, 2018], Hα equivalent width, and the log radial slope of circular velocity [Kalinova et al., 2022] all are between RS and BC), a hard cut in sSFR is not a robust method for splitting between RS and BC. That is, sliding the threshold separating quenched and star-forming galaxies can yield significant differences in measured accuracies. Appendix D.1 shows that the magnitude of such shifts is generally small, typically a few percent at most, but more importantly, it shows that the values are somewhat consistent relative to each other. Therefore, all bACC values should be taken as relative to one another, rather than as absolute.

3.4.1.2 Selection quality of the quenched population

Here we use balanced accuracy to compare quenched population selection of Red Dragon to previous approaches that employ hard cuts. Our CM and CC selections use the following methods. 'Typical' CM selection follows Hao et al. [2009]. After fitting the red sequence population with a Gaussian mixture in color space, we fit a line to the red sequence population (in the CM space of $g - r$ vs m_i), find its scatter, then select all galaxies within 2σ of the mean relation. 'Typical' CC selection follows Adhikari et al. [2020]. After finding population means via Gaussian mixtures, we draw a line between maxima (in the CC space of $g - r$ vs $r - i$), then plot a perpendicular line at the minimum likelihood point between the two components (i.e. where a galaxy is equally likely to belong to either component). These two methods give benchmark comparisons for standard efficiency of selection in CM and CC spaces for comparison to Red Dragon selection.

Figure 3.4 compares bACC values of these hard cut methods to Red Dragon using either two or three components. Even optimized CM selection typically incurs $\gtrsim 10\%$ error (i.e. $\sim 10\%$ of the

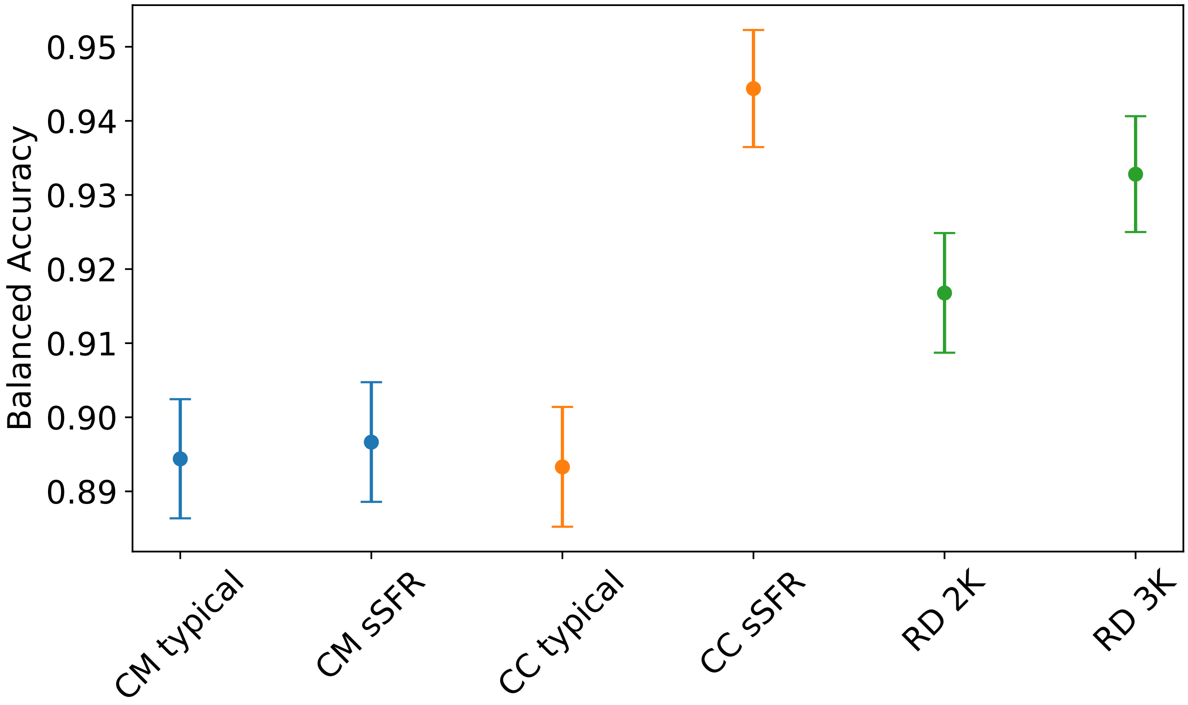


Figure 3.4: Balanced accuracy in selecting the quenched population of SDSS/low-z galaxies. "CM" and "CC" methods draw hard cuts through color-magnitude and color-color spaces respectively, whereas the "RD 2K" and "RD 3K" methods use Gaussian mixtures of two and three components, respectively. Error bars are generated from Poisson error estimates on each of the classification components.

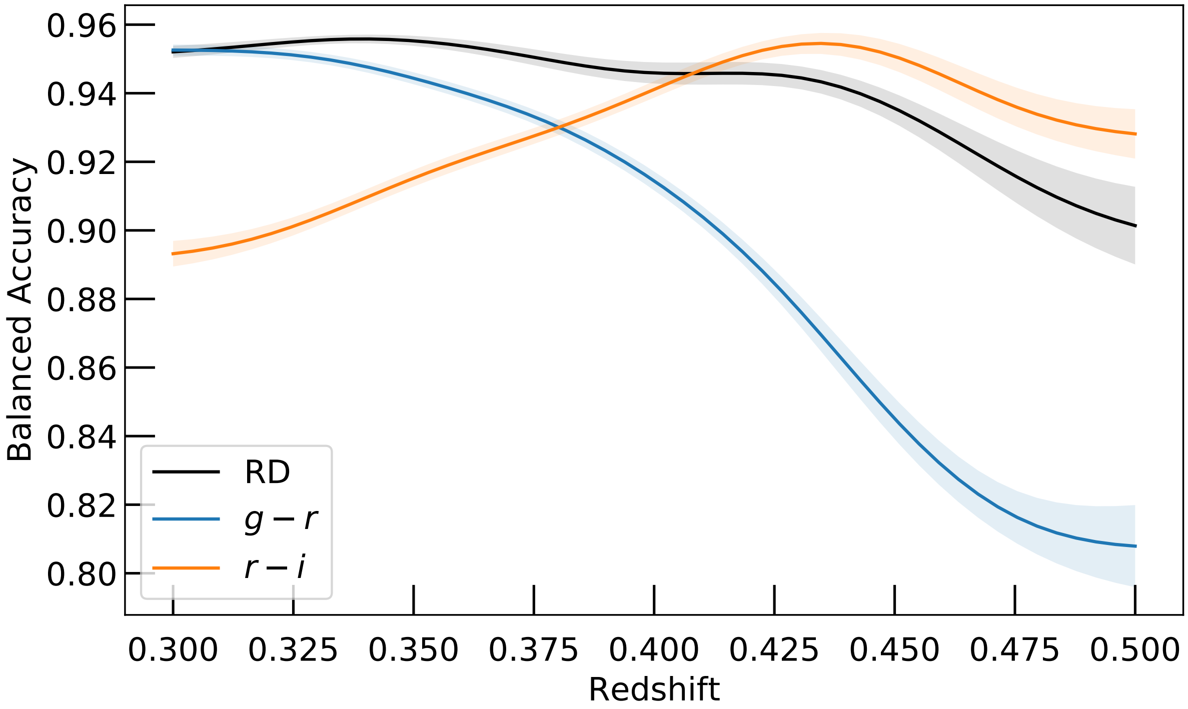


Figure 3.5: Balanced accuracy in selecting the quenched population of bright SDSS galaxies: Red Dragon (RD; black) performs similarly or superior to hard cuts in single colors. Bootstrap $\pm 1\sigma$ error shown with transparencies; these increase with redshift chiefly due to decreasing number counts (rather than from increased intrinsic scatter). Values localized by Gaussian kernel (width $\sigma_z = .02$).

RS and BC are contaminated by star-forming or quenched galaxies respectively) while optimized CC selection typically incurs $\gtrsim 6\%$ error, showing that two colors (CC space) work significantly better than one (CM space).

Gaussian mixtures (without any optimization from sSFR truth) generally perform on par with optimized CM and CC fits. While Figure 3.4 shows that if sSFR values are known, one could define hard cut CC selections of the RS which would have accuracies similar or superior to a GM selection of the RS. However, when sSFR values are unknown, GMs select the quenched population significantly better.

3.4.1.3 Redshift continuity of quenched galaxy selection

Figure 3.5 compares balanced accuracy in selecting the quenched population between Red Dragon and two (redshift-evolving) choices of single-color cuts for defining the RS. As the 4000 Å break passes from g band to r band near $z = 0.36$, the ability of $g - r$ to select the quenched population wanes while that of $r - i$ waxes, as expected. If using single-color selection, $z = 0.38$ would be the best redshift to transition from selecting the RS with $g - r$ to selecting with $r - i$, if your goal is to select the quenched population with greatest fidelity.

In comparison to these single-color selection methods, Red Dragon performs similarly to best-case single-band selection (within 3σ), and is vastly superior ($> 6\sigma$) in accuracy across $z = 0.38$, the optimized transition redshift. We note here that the high-redshift side of the plot has significantly lower number counts, and lacks statistical power compared to the low-redshift side.

Red Dragon preserves accuracy in selecting the RS across redshift transitions while maintaining a continuous red fraction (by algorithm construction). This then evades the discontinuities inherent in swapping bands, continuously selecting RS galaxies with high fidelity.

3.4.2 Buzzard Flock analysis

Extending our analysis to a wider redshift range, we turn to the synthetic galaxy catalogs of the Buzzard Flock, using the three primary colors of equation (3.2). In this section, we detail fit $\theta_\alpha(z)$ interpolation across redshift for a minimal two-component model. This includes for each component α: weight w_α (§3.4.2.1), means $\vec{\mu}_\alpha$ (§3.4.2.2), scatters $\vec{\sigma}_\alpha$ (§3.4.2.3), and correlation coefficients $\overset{\leftrightarrow}{\rho}_\alpha$ (§3.4.2.4).

The galaxy sample is magnitude-limited using the redshift-evolving cut of $0.2\,L_{*,i}(z)$ from Rykoff et al. [2014] within the redshift range $0.05 < z < 0.84$. The galaxies are divided into narrow cosmological redshift bins of width 0.025, resulting in counts per redshift bin of 60k to 7.6M galaxies. Red Dragon is run on 100 bootstrapped samples of 10^6 galaxies (undersampling at high redshifts for the sake of speed, efficiency, and easing computational burden); the resulting median parameters $\theta(z_i)$ and $\pm 2\sigma$ quantile range for each bin are shown in the figures below. The discrete redshift parameters are then interpolated using KLLR with a Guassian kernel of width $\sigma_z = 0.02$, shown as lines in the following parameter evolution plots.

With a sample size of 94M galaxies (see Table 3.1) we are able to extract precise estimates of all model parameters. The small statistical errors must be considered in context of Buzzard's algorithm to produce a synthetic DES catalog. Different galaxy catalog construction methods, such as MICE [Carretero et al., 2015] or cosmoDC2, populated by the GalSampler algorithm of Hearin et al. [2020], may have systematically different behaviors, especially at high redshifts.

For the mean and scatter of the RS component, we compare at low redshifts to the SDSS results of Hao et al. [2009] and across a wider redshift range to analysis of the DES-Y3 catalogue made available by E. Rykoff (2022, private comm.). These methods both give error-corrected Gaussian fitting to the RS, similar to Red Dragon's methods. However, as both fit only the RS, no information on red fraction nor any fits of the BC are available for comparison.

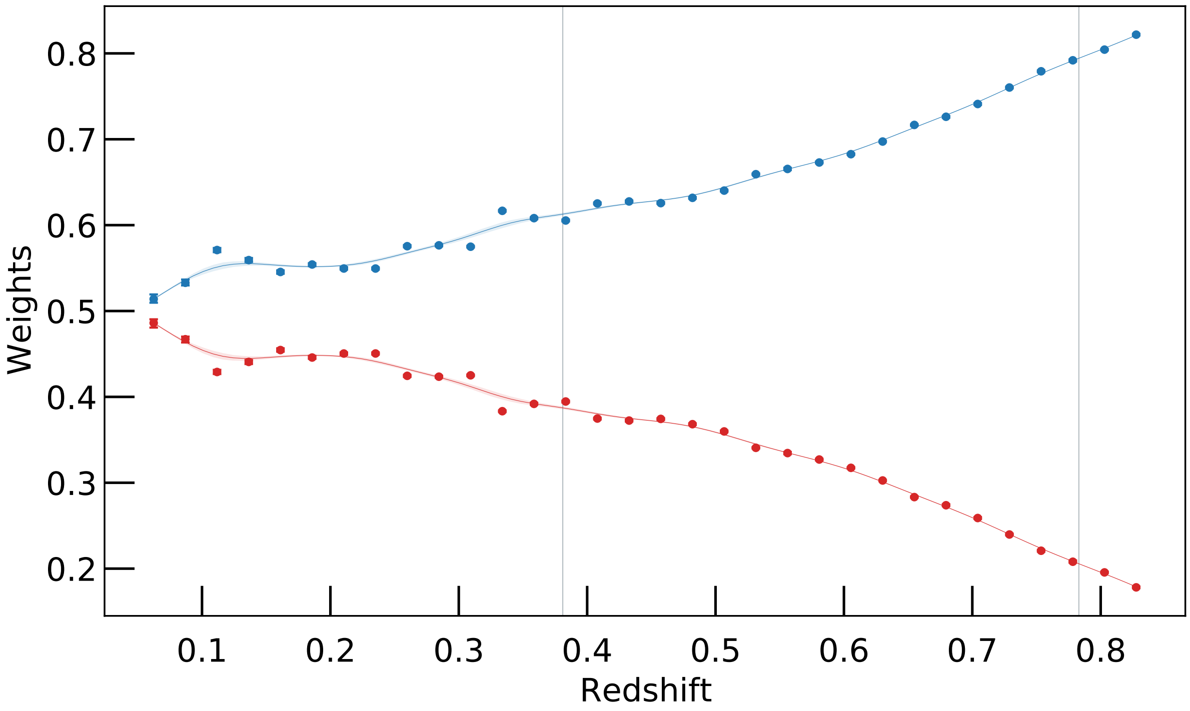


Figure 3.6: Component weight for RS (red) and BC (blue) as a function of redshift for the Buzzard flock. Points show parameter fits from individual redshift bins with ±2σ bootstrap quantiles shown as error bars. Fit line interpolation smoothed with KLLR (Gaussian kernel width $\sigma_z = 0.02$) with uncertainty in fit shown as transparencies about the line (generally smaller than the line width). Vertical grey lines indicate transition redshifts of the 4kÅ break from Table 3.2.

3.4.2.1 Component Weights

Figure 3.6 shows that the RS weight consistently decreases with redshift, ranging from roughly 50% at redshift $z = 0$ down to 20% at the highest redshift of 0.84. Since the red fraction is luminosity dependent (as discussed in Appendix E), one should remember that the weight reported here represents a weighted average of all galaxies above $0.2\,L_*(z)$, which will be dominated by magnitudes near the cutoff. Choosing a brighter magnitude cutoff would uniformly raise the RS weights whereas a dimmer cutoff would lower them.

A variety of deep observations of the real universe indicate that star formation rates per unit baryon mass were much higher in the past [Madau et al., 1996, Connolly et al., 1997, Madau and Dickinson, 2014]. Astrophysically speaking, while the Butcher-Oemler effect of reddening over time [Butcher and Oemler, 1978] applies primarily to galaxy clusters, the entire population of galaxies ages and tends to redden as a whole. Red Dragon applied to Buzzard extracts behavior consistent with this general trend.

3.4.2.2 Mean Colors

Figure 3.7 shows the redshift evolution of the mean colors of the two components. The three panels each show measured BC and RS means in different colors. Comparisons to observations of mean RS color are shown in grey. Hao et al. [2009] fit only to $g - r$ at low redshift, and the agreement with Buzzard is good, but both Buzzard and DES-Y3 mean $g - r$ colors are slightly bluer than the low-redshift SDSS values. Comparing to the Rykoff et al. [2014] method, Buzzard only deviates by .01 mag on average.

Note that each color has different vertical scaling: $\langle g - r \rangle$ spans the largest range while $\langle i - z \rangle$ spans the smallest range, exaggerating its features. Since the DES-Y3 analysis includes magnitude dependent-colors, shading illustrates the magnitude gradient by showing mean colors at L_* (upper bound) and at $0.2\,L_*$ (lower bound). The latter is the limiting luminosity of our Buzzard galaxy analysis. The color gradient is largest in $g - r$ at redshifts beyond the first D_{4000} transition, and Buzzard galaxies tend to be slightly redder in this regime.

At redshifts $z < 0.7$ the mean $r - i$ color of the RS in Buzzard tracks well the mean value found for DES-Y3 RS galaxies. At the highest redshifts, where D_{4000} transitions into the i-band, the mean $i - z$ color of RS Buzzard galaxies tracks that of DES.

At transition redshifts (see Table 3.2), the slope of mean colors (with respect to redshift) changes rapidly, necessitating narrow redshift analysis bins and careful fitting. Mean colors for both RS & BC follow a general shape of rising as the 4000 Å break enters the color's minuend, then plateauing or falling as the break enters the color's subtrahend, as expected.

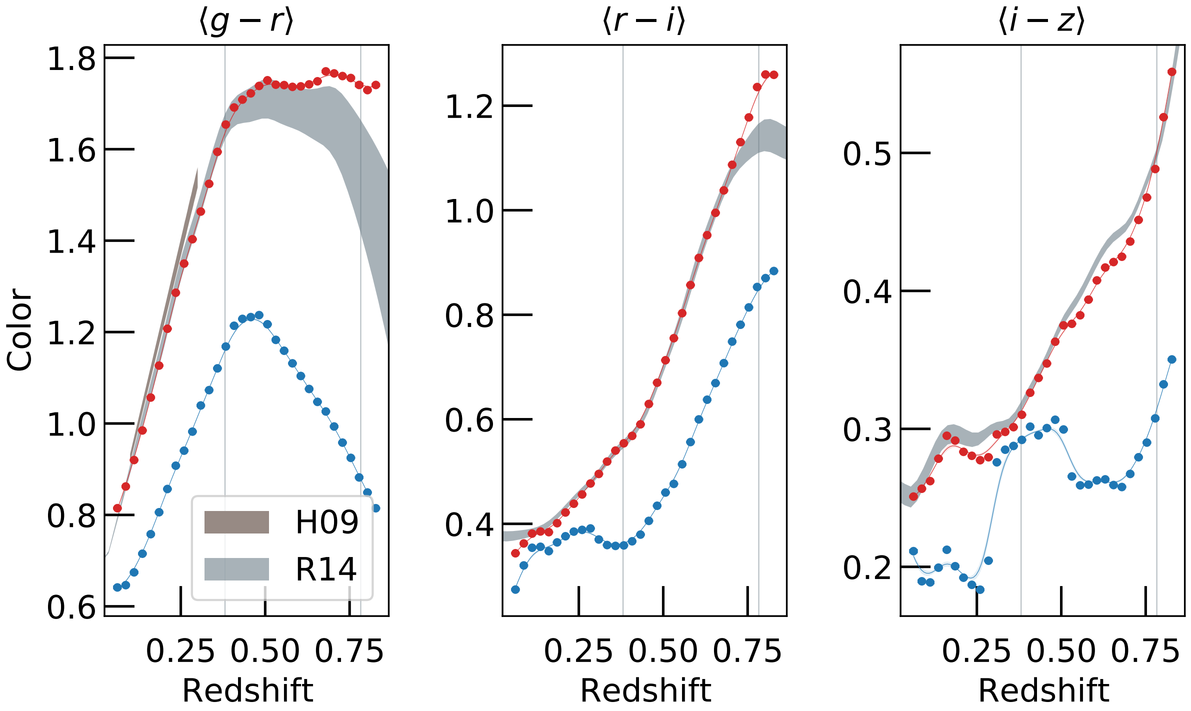


Figure 3.7: Buzzard mean colors. Points show mean colors for each GMM component (RS & BC) for Buzzard galaxies. Coloring and interpolation as in Figure 3.6. RS mean measurements from Hao et al. [2009] (SDSS, $g-r$ only) and E. Rykoff (2022, private comm.) (DES Y3, all colors) are shown for comparison. To illustrate dependence on magnitude, we present mean color ranges from $0.2\,L_*$ (the magnitude limit of Buzzard; lower edge of transparencies) up to L_* (upper edge of transparencies) for DES-Y3. Besides $g-r$ at $z \gtrsim 0.4$, the RS slope strength is less than 0.04 (shift in color per magnitude), so the spread is generally minimal (see also Appendix E.1 for a comparison of slope strength to scatter).

3.4.2.3 Color Scatter

The top row of Figure 3.8 shows the redshift evolution of the intrinsic scatter in each color for both GMM components. Observational estimates are shown in grey (absent magnitude dependence). Green shaded regions give $\pm 1\sigma$ and $\pm 2\sigma$ ranges of the statistical error on individual galaxy colors for the $0.2 L_*$ selected sample. At high redshifts, errors are somewhat larger in $g - r$ than in the other colors.

In accordance with expectations, at redshifts where the color's minuend contains the 4000 Å break, the RS has significantly lower scatter than the BC, by a factor of ~ 3, indicating a more uniform stellar population. Compared to observed RS scatter, Buzzard colors are generally slightly broader. The running of the RS mean color with magnitude implies that a magnitude-ignorant fitting of the population scatter should find larger values than a magnitude-cognizant model, so Red Dragon's measurement of generally wider scatters is expected.

Since Hao et al. [2009] and Rykoff et al. [2014] only fit the RS, a Red Dragon fitting of the RS, which also accounts for and fits the BC, is not guaranteed to identify an identical parameterization of the RS. Encouragingly, there is good agreement in the RS $r - i$ scatter over most of the redshift range shown. The RS intrinsic scatter in $i - z$ becomes small relative to statistical color uncertainties at $z > 0.4$, clouding comparison of Buzzard and DES-Y3 behavior in that regime. The BC intrinsic scatter is larger than the median statistical color error at nearly all redshifts shown.

3.4.2.4 Color Correlations

The bottom row of Figure 3.8 shows intrinsic correlations between colors inferred from the covariance matrix, giving $\rho(g - r,\, r - i)$, $\rho(g - r,\, i - z)$, and $\rho(r - i,\, i - z)$ from left to right.

At higher redshifts, typical photometric uncertainties on colors begin to dominate population scatters for the RS (see the green transparencies in the upper panel of Figure 3.8). Above $z \sim 0.5$ instead of measuring intrinsic correlation of each population, Red Dragon is primarily measuring the correlations between photometric errors. For this reason there is limited astrophysical information available to be gleaned from the correlations at these error-dominated redshifts.

Though SPS modelling suggests that intrinsic color correlations of the RS are expected to be $\gtrsim 90\%$ (E. Rykoff 2021, private comm.), the GMM fitting of Buzzard's RS consistently has lower correlations between colors—even at low redshifts, where photometric uncertainties are diminutive compared to the intrinsic scatter. In contrast to the RS, the BC has an intrinsic scatter which exceeds typical photometric uncertainties at nearly all redshifts, so these correlations correspond to intrinsic population correlations rather than correlations between photometric uncertainties. We find relatively high correlations between BC colors across all redshifts, typically $\gtrsim 75\%$.

The fits of Figures 3.6–3.8 show Red Dragon mapping out the RS and BC to high redshift

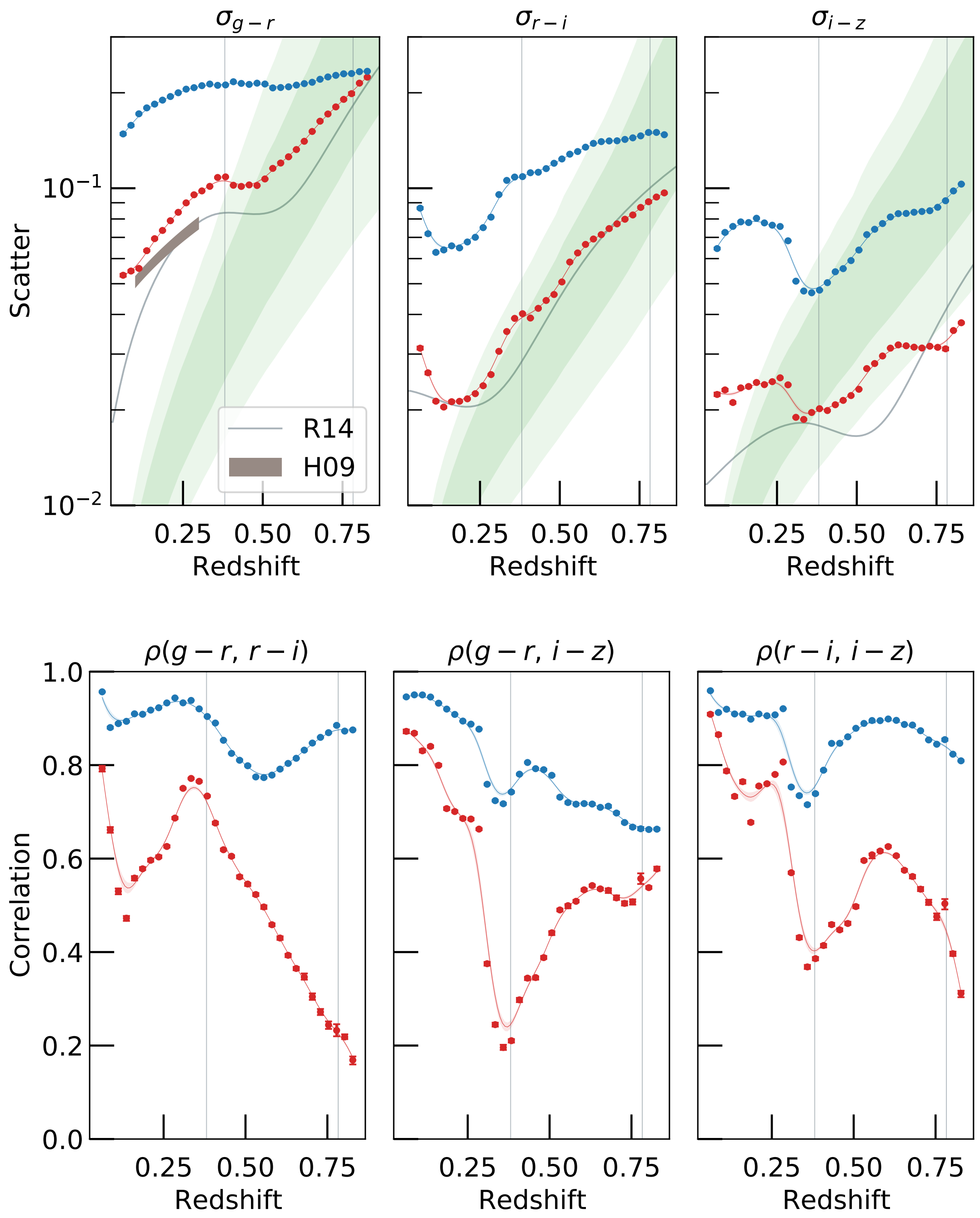


Figure 3.8: Decomposition of the covariance matrix for Red Dragon run on Buzzard, showing intrinsic color scatter (upper panel) and correlations between colors (lower panel). Green transparencies give one and two-sigma quantile distributions of photometric color errors in Buzzard. Coloring of components and interpolation as detailed in Figure 3.6; comparison to observations as in Figure 3.7, though width now indicates uncertainty rather than magnitude spread (R14 has no uncertainty estimate).

(spanning across multiple transition redshifts), continuously parameterizing important aspects of each population. As future studies create more complete samples of galaxies to higher redshifts, Red Dragon will be able to detail population characteristics smoothly across even wider redshift ranges.

3.5 Discussion: Algorithm Considerations

In this section, we consider various options for the Red Dragon algorithm. We first discuss choice of color vector, such as the optimal number of colors to include, measured by the accuracy gains from added bands (§3.5.1). We then consider changing the number of Gaussian components (§3.5.2) and quantify its effect on RS selection (§3.5.3). Finally, we discuss whether Gaussian features must be allowed to run with magnitude to accurately select the quiescent population (§3.5.4).

3.5.1 Accuracy gains from added colors

Here we detail gains in accuracy from adding more colors to the mixture. In short, we find that using all colors at once selects the quenched population with similar accuracy as an optimized two-color selection. That is, using the best combination of any two possible photometric colors didn't perform significantly better than a simple selection using all primary colors at once.

Using SDSS/low-z as an example, we measure gains in quenched selection accuracy with increased dimensionality of color space. We find the highest accuracy of two-component Gaussian mixtures using all possible single-color, double-color, and triple-color combinations (including the band-jumping secondary colors, like $u - i$, in addition to the primary colors). Those optimized maxima are then compared to default Red Dragon selection (using the primary color vector of equation (3.5) alone, with no optimization).

We find that the gain from optimal single color selection to optimal dual color selection is substantial (+5.8% in bACC, $> 12\sigma$ significant), but that gains are minimal beyond two colors ($\lesssim \pm 1\%$, $< 5\sigma$), with statistically similar selection accuracy as the primary color vector. While these particular findings apply only at a fixed redshift, the primary color vector serves well as a baseline for selecting the quenched population in a continuous manner across a wide range of redshifts.

3.5.2 Three or More Gaussian Components

Though historically galaxy classification has been binary, galaxies transitioning from RS to BC are sometimes classified as members of the green valley (GV), adding a third category. From an

Table 3.4: Distribution of maximum membership probabilities, i.e. $\{\max(P_{\text{mem}})\}$, in the SDSS/low-z sample for various component counts K. The first two rows show what fraction of galaxies have high and low maximum membership probabilities, corresponding to galaxies that are poorly-classified and well-classified respectively. Mean and median values for each distribution are also given.

K	2	3	4
$> .977$	70.7%	67.6%	18.8%
$< .841$	9.85%	11.3%	34.8%
mean	0.954	0.947	0.855
median	0.994	0.991	0.922

agnostic view of the color space data, components beyond two can simply be seen as an attempt to better model inherent non-Gaussianities in the populations [see e.g. Carretero et al., 2015]. From an astrophysics view, galaxies quenched by different mechanisms have distinct population characteristics [Peng et al., 2010, Davies et al., 2021, Dacunha et al., 2022]. High-mass galaxies have different trends for mean and scatter of colors than those of low-mass galaxies (which are primarily merger- or environment-quenched, rather than mass-quenched) [Baldry et al., 2004]. Modeling these populations with distinct Gaussians may better represent the underlying populations. For any of the above reasons, one may desire to model components beyond two.

In a BIC analysis of the SDSS/low-z sample, we found that though Figure 3.1 shows clear bimodality visually, and indeed, using two components gives a fair fit to the photometric color data, using three components fits the distribution of galaxies in photometric color space significantly better, with the third component modeling the outlier scatter around the core RS+BC population. We leave detailed analysis of sub-populations for future papers, but in short, using more than three components gave no significant improvement in fit.

As component count increases beyond $K = 3$, galaxy membership classification becomes less distinct. To illustrate this, we find for each galaxy the maximum membership probability $\max(P_{\text{mem}})$ across all GMM components. Table 3.4 shows summary statistics of the $\{\max(P_{\text{mem}})\}$ distributions for the cases of $K = 2$ to $K = 4$ applied to the low redshift SDSS data. For $K = 2$, the classification ability of Red Dragon is strong; the large majority of galaxies ($\gtrsim$ 70%) have extremely high maximum membership probabilities, $\{\max(P_{\text{mem}})\} > .977$. For the $K = 3$ case, the classification is only slightly degraded, with the fraction of uncertain classification ($\{\max(P_{\text{mem}})\} < .841$) rising to 11.3%, from 9.9% for $K = 2$. Adding a fourth component substantially increases this cohort, to 34.8% of all galaxies.

Viewed another way, the mean value of membership probability approximately dropped from 95% with two components down to 85% with four components. This suggests, as did our BIC analysis, that while two and three components characterize the underlying distributions well,

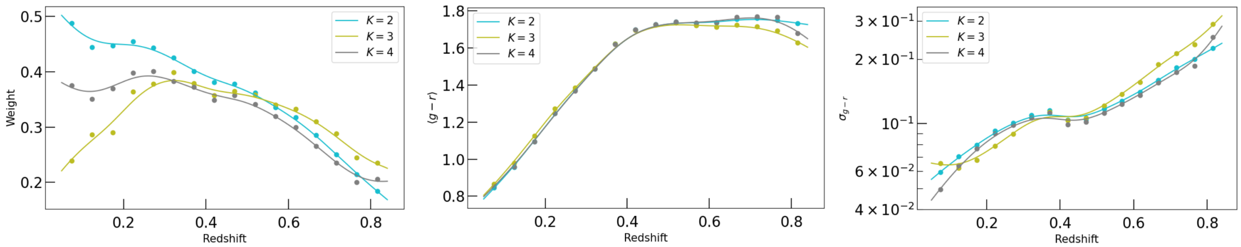


Figure 3.9: RS characteristics for different component counts, K. **Left**: Red fraction changes significantly with the addition of more components below $z = 0.3$, whereas weights are more consistent at higher redshifts. **Middle**: Mean RS $g - r$ color is extremely consistent across K, especially for $z \lesssim 0.6$, varying by $\lesssim 0.05$ mag on adding a component. **Right**: Scatter in the $g - r$ color of the RS.

adding a fourth component may be doing more harm than good in this sample. Adding the fourth component created more overlap, increasing ambiguity of selection. For simplicity of discussion and comparison, we chiefly employed the minimal two-component model in our analyses.

3.5.3 RS robustness to component count

Here we investigate consistency in Red Dragon's characterization of the RS in models with more than two components. On adding additional components to a Gaussian mixture model, one generally expects (1) a reduction in weight of each component, (2) a reduction in scatter, and (3) a shift of the means as the new component displaces the old. Figure 3.9 compares red fraction, RS mean $g - r$ color, and RS $g-r$ scatter for varying component counts $K = \{2, 3, 4\}$ as a function of redshift in the Buzzard catalog. We find that while the RS has a highly consistent mean and scatter, independent of component count, the RS weight is diminished at low redshifts as additional components subdivide part of the two-component RS. At $z \gtrsim 0.3$ weights are consistent to $\lesssim 5\%$. Mean color changed on average by $\lesssim 0.01$ mag on adding a new component. Scatter reduced on average by $< 5\%$ with each added component, and correlations (not shown) varied by less than 12% on average.

At a single redshift slice centered on $z = 0.2$, Figure 3.10 shows galaxy classification by Red Dragon into components in the $g - r$ and $r - i$ plane for cases of two, three and four-component mixtures. Classification colors are assigned by maximum likelihood rather than absolute probability (as was done in Figure 3.3), resulting in sharp boundaries between groups. Adding a third component to Buzzard at this redshift captures GV galaxies in the region between the RS and BC. Adding a fourth component captures the reddest (upper right region) portion of BC galaxies (12%).

Figure 3.11 illustrates how the mean colors of the $K = 3$ model behave across redshift (similar to Figure 3.7). The middle component (green line) transitions from being more similar to the RS at low redshift to being more similar to the BC at high redshifts. Given the Buzzard galaxies colors

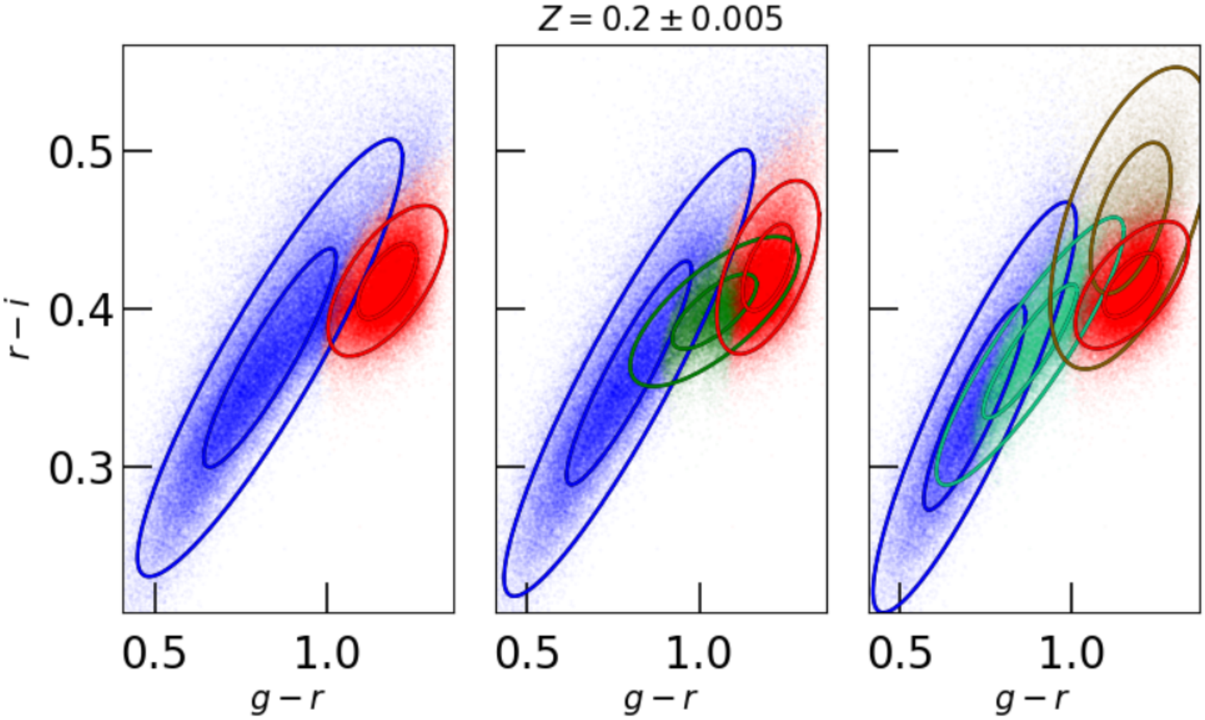


Figure 3.10: Effects of increasing component count on a Buzzard slice. Red Dragon population components for Buzzard galaxies in the thin redshift slice $z = 0.200 \pm 0.005$ for different component counts. **Left:** $K = 2$ ($f_R \sim 45\%$), **Middle:** $K = 3$ ($f_R \sim 33\%$, green component $\sim 23\%$), **Right:** $K = 4$ ($f_R \sim 38\%$, brown component $\sim 12\%$, green and blue component $\sim 25\%$ and $\sim 24\%$ respectively).

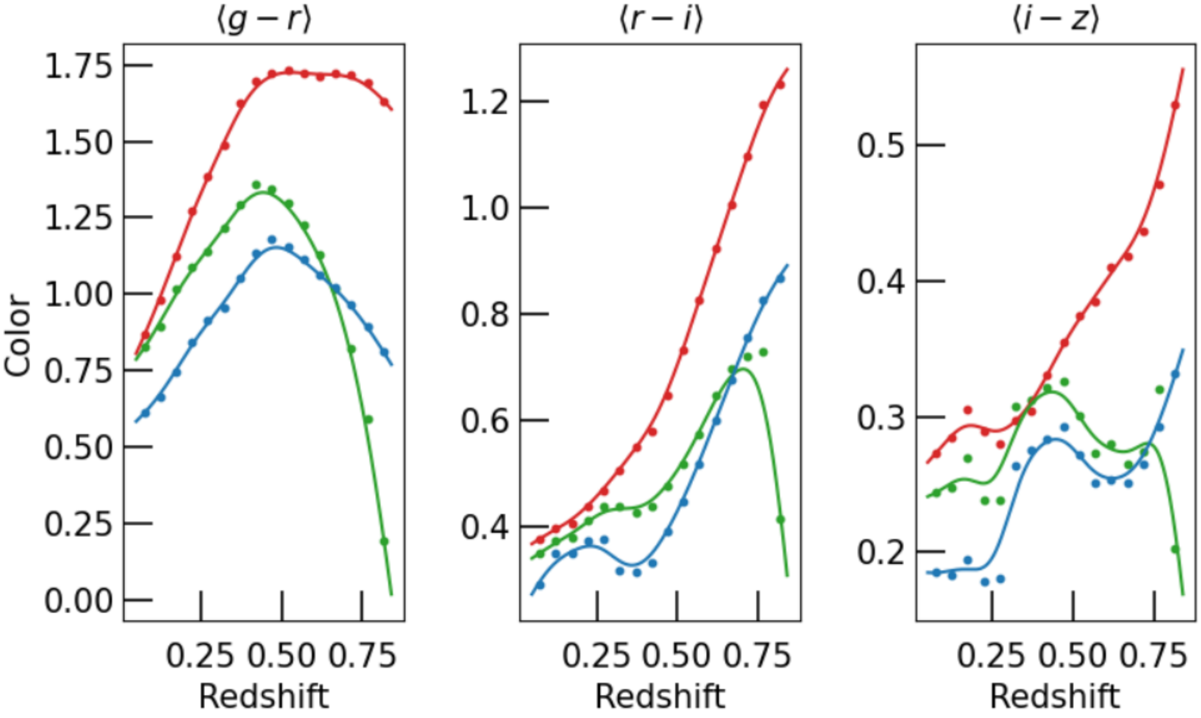


Figure 3.11: Mean color values for all three components of a $K = 3$ fitting of Buzzard photometry. The red and blue lines consistently track the redder and bluer parts of the galaxy population, whereas the green component moves from matching the bluer population at high redshifts but matching the redder population at lower redshifts.

are assigned purely statistically, it is doubtful that this transition reflects astrophysical evolution of galaxies from BC to RS over time. This change in characterization reflects the evolving structure in color space, with the RS being more non-Gaussian than the BC at low redshifts, and vice-versa at high redshifts. Therefore, the third component then does not consistently trace a transitioning population, and should not be identified as a GV across all redshifts. This result calls for care in labeling the RS if using three or more components, since the quiescent population may be shared among components.

As discussed earlier (see §3.5.2), the optimal count of Gaussian components to use in modeling depends on the dataset; increased signal-to-noise may reveal non-Gaussianities in color space that invite more complex models. For SDSS and DES Buzzard catalogs, the basic two-component model characterizes the galaxy population fairly well. Adding extra components needs to be done with care if the sample spans a wide range of redshifts.

3.5.4 Running with magnitude

Gaussian mixture parameters (population weight, mean color, and scatter for RS and BC) are known to depend on magnitude at a fixed redshift. Nearly 100% of the brightest galaxies are red, while very few of the faintest galaxies are red [Baldry et al., 2004], so component weight will vary with magnitude. The mean color of the RS in clusters is known to run significantly, albeit weakly, with magnitude [Kodama and Arimoto, 1996, Gladders et al., 1998, Repp and Ebeling, 2018], with the slope modeled explicitly in the RS fitting of e.g. Hao et al. [2009] and Rykoff et al. [2014]. The scatter of the RS and BC also runs non-linearly with magnitude [Baldry et al., 2004, Balogh et al., 2004], though this is less often modeled. Therefore, a magnitude-ignorant fitting of the populations will have all parameters somewhat dependent on the limiting magnitude of the sample. The brighter the sample, the higher the f_R, the redder the mean RS color, and the smaller the RS scatter. Though parameters do significantly evolve with magnitude, how significantly does magnitude ignorance affect selection of the RS?

Using thin redshift slices of Buzzard, we quantify differences in RS selection between magnitude-cognizant and magnitude-ignorant models. Appendix E details results of this analysis. In short, while magnitude running of GMM parameters is indeed statistically significant, their running had a relatively minimal impact on selection for this $0.2\,L_*$ sample in Buzzard. For thin redshift slices, selection of red sequence galaxies (where $P_{\rm RS} > .5$) was 95% identical between the standard redshift-running and the niche magnitude-running versions of Red Dragon. Since this difference in selection is relatively small, we leave magnitude running out of the current version of Red Dragon in favor for prioritizing smooth redshift evolution. For those who wish to explicitly account for magnitude running or other secondary parameters, several workarounds exist, as detailed in

Appendix E.2.

3.5.5 Possible future extensions

Red Dragon is introduced in the context of binary classification of galaxies based on their multi-band optical colors. As shown above, this simple classification has value in separating those galaxies that are actively forming stars from those that are quenched. A much wider range of phenomenology, particularly signatures of active galactic nuclei and local intergalactic gas environment, can be addressed by expanding the wavelength coverage into far-infrared and radio, or to higher energy emission in UV / X-ray / γ-ray emission.

Gaussian mixtures that couple optical colors to wider wavelength bands could, for example, separate galaxies hosting X-ray or radio-loud quasars, or identify galaxy members of protocluster regions with hot intergalactic plasma. Ratios of X-ray to optical emission have long been known to separate spiral and elliptical galaxies [Fabbiano, 1989], and spectral energy density models of AGN samples [e.g., Symeonidis et al., 2022] support an expectation of multiple components in a widened optical-IR space.

Source confusion will become important when attempting to combine observations with differing angular resolution and signal-to-noise, but methods such as `TRACTOR` [Lang et al., 2016] have been used to associate sources across a wide range of photon energy in the Deep Drilling Fields to be explored by the Rubin Observatory's Legacy Survey of Space and Time [LSST, Lacy et al., 2021].

A continuous RS definition across redshift will be critical as LSST scans deeper in the optical and space-based galaxy surveys, particularly *Euclid*[5] [Laureijs et al., 2011] and the *Nancy Grace Roman Space Telescope*[6] [formerly WFIRST, Spergel et al., 2015] provide infrared observations of galaxies at redshifts above one.

Quiescent galaxies have already been detected beyond a redshift of two [Kriek et al., 2008, Williams et al., 2009, Gobat et al., 2011, Golden-Marx et al., 2021], with a quenched protocluster population established at $z \sim 3.5$ [McConachie et al., 2022]. Forecasts for 5σ cluster detections with Euclid anticipate a catalog of 200,000 systems, roughly one-quarter of which will lie at $z > 1$ [Sartoris et al., 2016]. Redshift-continuous definitions of the red sequence will support improved definitions of galaxy richness, an important mass proxy for optical-IR cluster cosmology.

[5] Planned launch date: Q1 2023; its Y, J, H filters span 900 to 2000 nm.

[6] Planned launch date: by May 2027; its six filters span 480 to 2300 nm.

3.6 Chapter Conclusions

We present Red Dragon, a novel method for galaxy population characterization, which evolves Gaussian mixture models in the space of broad-band optical colors across redshift. With the red sequence of quiescent galaxies as a target population in both observed and simulated galaxy samples, we demonstrate the method's ability to identify the quenched population with similar accuracy to previous approaches, though now with smooth continuity across redshift and characterization of the BC (and optionally additional components).

Jumping from using one color alone to another as a RS selector (e.g the transition from $g - r$ to $r - i$ near $z = 0.4$) gives an inherent discontinuity in red fraction $f_R(z)$ or in the accuracy of selecting the quenched population. Since Red Dragon interpolates Gaussian mixture parameters across redshifts in multi-color space, the resulting characterization of galaxy photometry yields a continuous red sequence definition, thereby providing a more continuous measure of galaxy richness in massive halos.

Red Dragon also offers a new way to explore RS, GV, and BC photometric behavior through its explicit fits to population weight, mean color, scatter, and correlation between colors. This fitting allows investigations into the photometric sub-populations of galaxies, such as fraction of transitioning galaxies as a function of redshift, or the evolution of mean color and scatter of blue cloud galaxies. By enabling estimates of the intrinsic covariance of colors in each mixture component, the Red Dragon algorithm also offers potential insights into the evolution of galaxy properties such as dust, stellar age, and metallicity. Potential extensions to broader wavelength coverage could support classification of galaxies containing active galactic nuclei in a variety of astrophysical states.

Acknowledgements

The authors acknowledge support from Chandra Award TM9-20006X, the Leinweber Center for Theoretical Physics (LCTP) at University of Michigan, and Department of Energy NERSC computing resources. We thank the referee for their extremely helpful review, which not only improved the quality and presentation of this paper, but also helped us discover a bug in the handling of the noise covariance matrix.

WKB thanks Johnny Esteves, Eric Bell, Eli Rykoff, Oleg Gnedin, and Peter Melchior for their insights and assistance in this project. WKB also thanks his wife Eden and his son Fletcher, by whom his days are vastly brightened and to whom he owes his life.

This work was made possible by the generous open-source software of NumPy [van der Walt et al., 2011], Matplotlib [Hunter, 2007], h5py [Collette et al., 2017], Sci-kit Learn [Pedregosa et al., 2011], and KLLR [Farahi et al., 2018, Anbajagane et al., 2020].

Data Availability

The current iteration of Red Dragon is available on Bitbucket at wkblack/red-dragon-gamma. Other methods used in this analysis are available at wkblack/red-dragon on Bitbucket. Data from SDSS are publicly available from their online SkyServer. Data from Buzzard is available upon request.

CHAPTER 4

Cosmic Dragons: Color Evolution of Galaxies in the COSMOS Field

Chapter Summary

Using the photometric population characterization method **Red Dragon**, we characterize the Red Sequence (RS) and Blue Cloud (BC) of DES galaxies in the COSMOS field. Red Dragon measures the distribution of photometric colors using a redshift-evolving Gaussian Mixture Model (GMM), smoothly parameterizing the two populations with mean colors, intrinsic scatters, and inter-color correlations. This resulting fit of RS and BC yields RS membership probabilities P_{RS} for each galaxy. Even using only DES main bands $griz$, this selects the quiescent population (defined here as galaxies with $\lg \mathrm{sSFR} \cdot \mathrm{yr} < -11$) with $\gtrsim 90\%$ accuracy out to $z = 2$; adding extended photometry from VIRCAM improves this accuracy to $\sim 95\%$ out to $z = 3$. We measure redshift evolution of both sSFR and galactic age, finding that the BC is consistently more star-forming (by ~ 1 dex) and younger (by ~ 1 Gyr) than the RS. This characterization of both RS and BC as functions of redshift and stellar mass improve our understanding of both populations and open the door to more precise galaxy characterization in the future.

4.1 Introduction

Galaxies come in two main flavors: a “blue cloud” (BC) of actively star-forming galaxies and a “red sequence” (RS) of passively evolving (“red and dead”) galaxies which have largely ceased star formation [Strateva et al., 2001, Bell et al., 2004]. Between these two populations lies the so-called “green valley” (GV). The BC typically is made of young, spiral galaxies which are often found in low-density environments. In contrast, RS galaxies are typically massive, old, bright ellipticals, often found in high-density environments—especially the interiors of galaxy clusters.

If cold gas is unable to condense into new stars, the galaxy ceases star formation, and is considered *quenched*. Various quenching mechanisms dominate at different galactic mass ranges and at different redshifts [see e.g. Figure 15 of Peng et al., 2010], either heating up gas so it is unable to condense, or blowing out gas, so there is none to collapse. The review by Somerville and Davé [2015] states that though the exact physics behind the quenching of satellite galaxies is unknown, several processes play crucial roles. For galaxies of decimal log stellar mass $\mu_\star \equiv \log_{10} M/M_\odot \gtrsim 10.5$, feedback from central supermassive black holes blows gas out of the galaxy, preventing formation of new stars. Especially at earlier times ($z > .5$), merging events play crucial roles in galaxy quenching—disrupting, heating, and ejecting gas from galaxies. In addition to initiating merging, gravity continually pulls galaxies towards denser regions. Denser environments quench galaxies more. As a galaxy barrels through the hot gas of a galaxy cluster, the resulting ram pressure strips away gas. Furthermore, close interactions between galaxies (or between a galaxy and a massive cluster) cause tidal stripping, pulling off outer layers of cold gas from the galaxy. Both processes prevent further star formation. Each of these quenching mechanisms align to place galaxy clusters as hotbeds for the creation of quiescent galaxies (galaxies which have largely ceased star formation). As clusters are the endpoint of large-scale gravitational collapse, they also serve as the terminal location of such galaxies. Therefore, as compared to other locations of the cosmic web, galaxy clusters tend to have quite high red fraction ($f_{\rm RS}$; fraction of galaxies belonging to RS).

Though no absolute boundary exists between RS and BC, various means exist to classify galaxies in a strongly bimodal fashion. The most striking difference in the spectra of RS and BC galaxies in the optical regime (roughly 3100–11,000 Å) is "the 4000 Å break." In particular, D_{4000} measures the relative spectral intensities on either side of the 4kÅ break [Balogh et al., 1999]: this estimates well a galaxy's current star formation rate. The break is caused by both a coincidence of several metallic ionized absorption lines and the Balmer break, near $\lambda_{\rm B} = 3645$ Å [Hathi et al., 2008, Kim et al., 2018]. These all work to suppress stellar spectra at shorter wavelengths—especially so for stars in RS galaxies—leading to a clear dichotomy in D_{4000} between RS and BC.

Far from the 4kÅ break, other features further distinguish RS from BC. At very short wavelengths, RS and BC show significant differences in X-ray luminosity L_X [Fabbiano, 1989]. In intermediate wavelengths, differences such as dust content and metallicity further distinguish the spectra of RS from BC [Blanton and Moustakas, 2009]. At longer wavelengths, RS and BC are distinguished in the far infrared [Symeonidis et al., 2022]. Differences such as these invite models to use more than D_{4000} alone to distinguish between RS and BC.

Because measuring spectroscopy becomes prohibitively expensive and time-consuming as galaxy count increases, surveys turn to **photometry**—measuring (negative log) net flux across a wide band of wavelengths. In this paper, we use photometry from both DECam (*ugriz*, spanning $\lambda \sim$ 3100 to 10,000 Å) and VIRCAM ($JHK_{\rm s}$, spanning $\lambda \sim$ 11,700 to 23,000 Å). See Table F.1 for more information on these bandpasses.

In addition to RS appearing generally brighter than the BC, the two populations distinguish themselves on investigating photometric "colors" (subtracted magnitudes). Because photometric bands measure log flux, subtracting two photometric bands yields the log of the relative flux in either portion of the spectrum. For example, with the color $r - i$, we have

$$r - i = -2.5 \log_{10} \frac{F_r}{F_{r,0}} + 2.5 \log_{10} \frac{F_i}{F_{i,0}} = 2.5 \log_{10} \frac{F_i}{F_r} \tag{4.1}$$

(since zero points in the AB magnitude system used by DES are equal, $F_{r,0} = F_{i,0}$ here). Using such photometric colors about the 4kÅ break (and in many colors beyond it), RS galaxies tend appear redder and tighter-clustered than BC galaxies (hence their names of "red sequence" and "blue cloud" respectively). At low redshifts, colors such as $u - r$ or $g - r$ span the 4kÅ break (essentially measuring log D_{4000}), showing significant differences in distribution. However, the photometric color which best measures D_{4000} changes over redshift, so no single color suffices for all galaxies.

The observed bimodality of RS and BC in photometric colors (at fixed galactic stellar mass and redshift) arises for two main reasons. First, specific star formation rates (sSFR; star formation rate per galactic stellar mass) follow a skew-lognormal distribution, with its peak at high star formation ($\lg \text{sSFR} \cdot \text{yr} \sim -10$ near $z = 0$) for the galactic main sequence (BC galaxies) and a long tail towards lower star formation for the RS [Wetzel et al., 2012, Eales et al., 2018, Leja et al., 2022]. Second, at low sSFR values ($\lg \text{sSFR} \cdot \text{yr} \lesssim -11.3$), the scatter in color at fixed sSFR decreases significantly, such that low-sSFR galaxies all tend towards the same location in color space [Eales et al., 2017]. These two effects then conspire to result in a dual Gaussian distribution of galaxies in photometric color space [Strateva et al., 2001, Bell et al., 2004, Baldry et al., 2004, Williams et al., 2009, Hao et al., 2009, Krywult and Pollo, 2018]. We therefore can use Gaussian mixtures to characterize these populations in multi-color space.

Gaussian mixture models (GMMs) characterize each component α with a component weight w_α, mean colors $\vec{\mu}_\alpha$, and color covariance Σ_α (including scatters σ and correlations ρ). For a two-component model of the RS and BC, this then entails the total parameter set $\{f_{\text{RS}}, \mu_{\text{RS}}, \mu_{\text{BC}}, \Sigma_{\text{RS}}, \Sigma_{\text{BC}}\}$, where red fraction $f_{\text{RS}} \equiv w_{\text{RS}}/(w_{\text{RS}} + w_{\text{BC}})$. Each of these parameters have been shown to evolve over redshift and stellar mass for both RS and BC populations [Baldry et al., 2004, Balogh et al., 2004]. As mentioned previously, red fraction additionally depends on local overdensity (with e.g. clusters containing a higher fraction of red galaxies than filaments and the voids between clusters), but dependence of the other parameters on local density have yet to be significantly measured separately from stellar mass (high-mass galaxies are born and bred in denser environments, so the two factors correlate).

In this paper, we characterize GMM fit parameters for galaxies in the COSMOS field using the Red Dragon algorithm. In §4.2, we summarize the algorithm of Red Dragon and discuss expected dependence of variables with stellar mass and redshift. In §4.3, we detail the galactic data used in this analysis, including its mass completion, redshift limits, and color selection. In §4.4, we display in brief the stellar mass and redshift dependence of fit parameters (§4.4.1), then use that parameterization to select RS & BC galaxies and characterize physical properties of both populations (§4.4.2). In §4.5, we discuss quiescent population selection accuracy as well as detail the resulting distributions of sSFR and galactic age. We conclude with a summary of our core findings in §4.6.

4.2 Red Dragon Algorithm

To parameterize the RS and BC, we use the galaxy population modeling tool **Red Dragon** [RD, Black and Evrard, 2022]. Red Dragon uses redshift-evolving Gaussian mixtures in photometric multi-color space to characterize both RS and BC (as well as optionally additional components). In this section, we give a brief overview of the algorithm and results of previous characterization on galaxies from the Sloan Digital Sky Survey (SDSS) and galaxies from Buzzard, a synthetic catalogue [see DeRose et al., 2019, 2021, Wechsler et al., 2021].

As previously mentioned, in sufficiently small mass and redshift bins, the distribution of optically-selected galaxies have bimodal behavior in photometric color space, well-characterized by two Gaussian components; the exact fit parameterization (e.g. red fraction, value of mean BC color, scatter of RS) may depend on redshift, stellar mass, and local density. Red Dragon characterizes galaxies under variable conditions, such as at different redshifts, then smoothly interpolates fit parameters across that variable, resulting in a continuous parameterization of galaxies.

4.2.1 Algorithm overview

Red Dragon uses an evolving error-corrected Gaussian Mixture Model (GMM) in multi-color space to characterize the RS and BC galaxy populations. In particular, for a given component (e.g. RS or BC) α, the specific component likelihood of the parameterization θ_α for a galaxy x_j is

$$\begin{aligned}\mathcal{L}_\alpha(\theta_\alpha \big| x_j) = & \frac{w_\alpha}{\sqrt{(2\pi)^N}\left|\Sigma_\alpha + \Delta_j\right|} \\ & \times \exp\left[-\frac{1}{2}(\vec{c}_j - \vec{\mu}_\alpha)^{\mathrm{T}}(\Sigma_\alpha + \Delta_j)^{-1}(\vec{c}_j - \vec{\mu}_\alpha)\right],\end{aligned} \tag{4.2}$$

where the parameter set θ_α is composed of component

- weight w_α (with constraint $\sum w_\alpha = 1$),

- mean color vector $\vec{\mu}_\alpha$, and
- color covariance matrix Σ_α;

each galaxy x_j has inputs of its

- measured photometric colors $\vec{c}_j$ and
- noise covariance matrix Δ_j (which encodes color uncertainties).

The parameterizations $\{\theta_\alpha\}$ for each of the K components can then be optimized by finding the maximum likelihood parameterization for all $N_{\rm gal}$ galaxies with data $\{x_j\}$:

$$\mathcal{L}(\{\theta_\alpha\}\big|\{x_j\}) = \prod_j^{N_{\rm gal}} \sum_{\alpha=1}^{K} \mathcal{L}_\alpha(\theta_\alpha|x_j). \tag{4.3}$$

By doing bootstrap resampling of the input galaxy population $\{x_j\}$, one can then obtain uncertainties on all fit parameters.

The Red Dragon algorithm measures these parameterizations across redshift (or, given a sufficiently thin redshift bin, across some other variable, like stellar mass) and smoothly interpolates, yielding a continuous characterization of RS, BC, and optionally additional components. With fit characterization in hand, one can then predict for a galaxy x_j at redshift z_j to which component it more likely belongs. The probability of it belonging to component α is

$$P_\alpha(x_j) = \frac{\mathcal{L}_\alpha(\theta_\alpha|x_j)}{\sum_\beta \mathcal{L}_\beta(\theta_\beta|x_j)}. \tag{4.4}$$

For the standard two-component model of galaxy populations, RS and BC, this then yields $P_{\rm RS} = \mathcal{L}_{\rm RS}/(\mathcal{L}_{\rm RS} + \mathcal{L}_{\rm BC})$ as the probability of the galaxy belongs to the RS.

Code for Red Dragon is freely available on BitBucket.[1]

4.2.2 Expected parameter dependence

Previous GMM fitting of RS and BC have shown dependence of red fraction, mean color, and color scatter on redshift, galactic stellar mass (or magnitude, largely a proxy of stellar mass), and local density [in the case of red fraction; see Balogh et al., 2004].

Of particular note for this paper is the significant reddening at higher masses. Baldry et al. [2004] quantified for a rest-frame sample of low-z galaxies the evolution of red fraction, mean $u-r$ color, and color scatter as a function of M_r magnitude (a proxy of galactic stellar mass). They

[1] bitbucket.org/wkblack/red-dragon-gamma

fit color and scatter as linear functions with a hyperbolic tangent transitioning the fit from one intercept to another (resulting in a slanted sigmoid function).

Colors redden with increased log stellar mass $\mu_\star \equiv \lg M_\star/M_\odot$. In particular, they found that the RS mean color undergoes a shift around $\mu_\star = 10.3$ (BC mean colors undergo a similar shift at $\mu_\star = 10.4$), from following a bluer trend to a redder with increased stellar mass. This implies that linear extrapolation from high-masses ($\mu_\star > 11$) down towards lower masses ($\mu_\star < 10$) will predict significantly redder galaxies than observed. Color scatters tend to increase with stellar mass, but only monotonically so for the RS. Near $\mu_\star = 10.0$, BC scatter increases momentarily, then continues its decreasing trend with increasing stellar mass. However, these measurements only included a single color, though more are available from core DES $griz$ photometry.

Using data from DES Y3, redMaPPer [RM; Rykoff et al., 2014] characterized bright members ($L > 0.2\,L_{*,z}(z)$, corresponding to $m_z - m_{*,z}(z) < 1.75$, where $L_{*,z}(z)$ and $m_{*,z}(z)$ are the characteristic z-band luminosity and magnitude for a galaxy at a given redshift z) of the RS in a multi-color × magnitude space, evolving a Gaussian parameterization of the RS across redshift. RM fit mean RS colors to drift linearly with magnitude (such that higher stellar masses were redder), allowing the slope to evolve with redshift. As these galaxies were generally far heavier than the sigmoid transitions mentioned above, such linear modeling fits the high-mass characterizations of Baldry et al. [2004]. As the fit only considers the RS (the BC is dimmer and has less accurate redshift estimates than the RS), it lacks red fraction and BC outputs. In regions where color uncertainties exceeded intrinsic scatter (which was nearly always, since it trained on a photometric dataset), RM-measured inter-color correlations reflected correlations between *uncertainties* rather than reflecting *intrinsic* correlation between colors (within only the RS population). For these reasons, in future comparisons to this fit, we will only show the RS mean color and scatter measured by this model.

While RM lacks BC characterization, a sufficiently clean sample, with accurate redshift estimates for the BC, could potentially use a GMM to model both populations across redshift. Such a model could potentially give more informed cluster memberships, photo-z values, and reveal astrophysics of the BC.

4.2.3 Fit Results from B22

Previous RD fitting of RS and BC done by Black and Evrard [2022, hereafter B22] used bright ($L > 0.2\,L_{*,i}(z)$, where $L_{*,i}(z)$ is the i-band characteristic magnitude at a given redshift z) galaxies from SDSS and a DES-like synthetic galaxy catalog [Buzzard; DeRose et al., 2019, 2021, Wechsler et al., 2021] with primary colors $\vec{c}_{\text{DES}} = [g-r,\, r-i,\, i-z]$. The SDSS sample focused on low redshifts $z = .1 \pm .005$ and intermediate redshifts $z|(.3, .5)$; in contrast, the Buzzard sample extended from redshift $z = .05$ out to $z = .84$, focusing on the redshift evolution of parameters.

Several key results follow.

With the low-z SDSS sample, Red Dragon selected the quiescent population, defined as $\lg \mathrm{sSFR} \cdot \mathrm{yr} < -11 + z$, with $\sim 92\%$ accuracy. Increasing the threshold to $\lg \mathrm{sSFR} \cdot \mathrm{yr} < -10.7$ decreased selection accuracy to $\sim 89\%$ while decreasing the threshold to < -11.3 increased selection accuracy to $\sim 94\%$. A three-component fit was found to best explain the data; the third component, beyond RS and BC, had lower weight ($w < 10\%$), higher scatter (roughly twice that of the BC), and had low (small positive or consistent with null) inter-color correlations (often the lowest). This could indicate that the third component merely captured 'noise', i.e. galaxies that didn't fit well in either component. In the mid-z SDSS sample, we found that RD selects the quiescent population similar to or superior to single-color selection and at transition redshifts (redshifts where the 4kÅ break redshifts from one photometric band into its neighbor), RD performed $> 6\sigma$ superior to either color alone.

In Buzzard, we found that f_{RS} decreased near linearly with redshift. Mean colors generally agreed with redMaPPer fits (RM; E. Rykoff 2022, private communication), though at $z \gtrsim .7$ colors were somewhat redder. Scatters agreed roughly, though RD tended to have larger scatters than RM, especially at $z \lesssim .25$.

RD made the first measurement of inter-color correlations: correlations between photometric colors within the RS population and also in the BC population. In both SDSS as well as Buzzard, Red Dragon measured inter-color correlations of $\rho_{\mathrm{RS}} < \rho_{\mathrm{BC}} < 95\%$ typically. In Buzzard, these correlations generally decreased with redshift. It was expected from SPS modeling that color correlations for both RS and BC should be $\rho \gtrsim 90\%$ [Rykoff et al., 2014], but results showed a different story, with correlations even dipping negative at times. Very roughly, for their bright sample of synthetic galaxies out to $z = .84$, B22 found $\rho_{\mathrm{BC}} \sim .95 - .2z$ and $\rho_{\mathrm{RS}} \sim .8 - .5z$. Deviating from a smooth trend, B22 found a significant dip in correlation of $\rho(g - r, i - z)$ and $\rho(r - i, i - z)$ near $z = .38$, where the 4kÅ break is leaving g band and entering r band. These results show considerably lower RS inter-color correlations than expected, along with interesting features near 4kÅ redshift transitions (vertical grey lines of Figure 4.1).

4.3 Data

In this paper, we use galaxies from one of the DES deep fields known as the COSMOS patch. In addition to photometry, this sample includes estimated stellar masses, star formation rates, and ages. In this section, we detail the COSMOS dataset and the particular selection of galaxies we utilize in this paper.

4.3.1 The COSMOS2015 Catalog

Our data come from a region of the sky known as the COSMOS (Cosmic Evolution Survey) field [Yoshiaki et al., 2005, Scoville et al., 2007], covering a solid angle of the sky about 16 times larger than the moon, found in the constellation Sextants. This region has few local stars nor gas clouds from the Milky Way intervening to obstruct imaging of other galaxies. It is centered at approximately J2000 (RA, Dec.) = $(150.1166, 2.2058)$.

The COSMOS field has received much attention across the spectrum, with observations ranging from short-wavelength X-rays all the way to long-wavelength radio waves [Schinnerer et al., 2004, Hasinger et al., 2007]. We use the COSMOS2015 catalog: observations from the COSMOS DES (Dark Energy Survey) Deep Field [Laigle et al., 2016, 2018, Hartley et al., 2022]. This makes use of data from both the Dark Energy Camera [DECam; Flaugher et al., 2015], which supplies bands $ugriz$, as well as data from the VISTA InfraRed CAMera [VIRCAM; Dalton et al., 2006, Emerson et al., 2006], which supplies bands JHK_{s} [as part of the UltraVISTA survey; see McCracken et al., 2012, Caputi et al., 2015].

Figure 4.1 displays these bands, showing both their observing wavelengths at $z = 0$ as well what galactic spectral features would be picked up by each band at higher redshifts. Central wavelengths and widths for each band are given in Table F.1, along with the redshifts at which the 4kÅ break enters and exits each band. Because it is the most prominent optical feature of galaxies, the points at which the 4kÅ break exits each band (vertical lines in the figure) give approximate redshifts at and just after which to expect significant changes in band flux.

Redshifts in this dataset are photometric, estimated by the code LePhare [PHotometric Analysis for Redshift Estimations; Arnouts et al., 2002, Ilbert et al., 2006], using a suite of Bruzual and Charlot [2003] spectrum templates as in Ilbert et al. [2013] to model expected redshifts with high precision (median redshift uncertainty $\sigma_z/(1+z) = .036$ in our mass range). This SED fitting of photometry as a byproduct includes estimations of stellar mass, specific star formation rates, and galactic ages.

4.3.2 Stellar mass completion

Ilbert et al. [2013, Table 2], gives the minimum masses at which the RS & BC are complete for several redshift bins. As the BC minimum masses lie below those of the RS, we use the former. The minimum mass at a given redshift is well fit by

$$\frac{1}{1+z_{\mathrm{max}}} = .39 - .19\,(\mu_{\star,\mathrm{complete}} - 9.4). \tag{4.5}$$

This gives us the maximum redshift to which we can extend the sample while remaining mass-

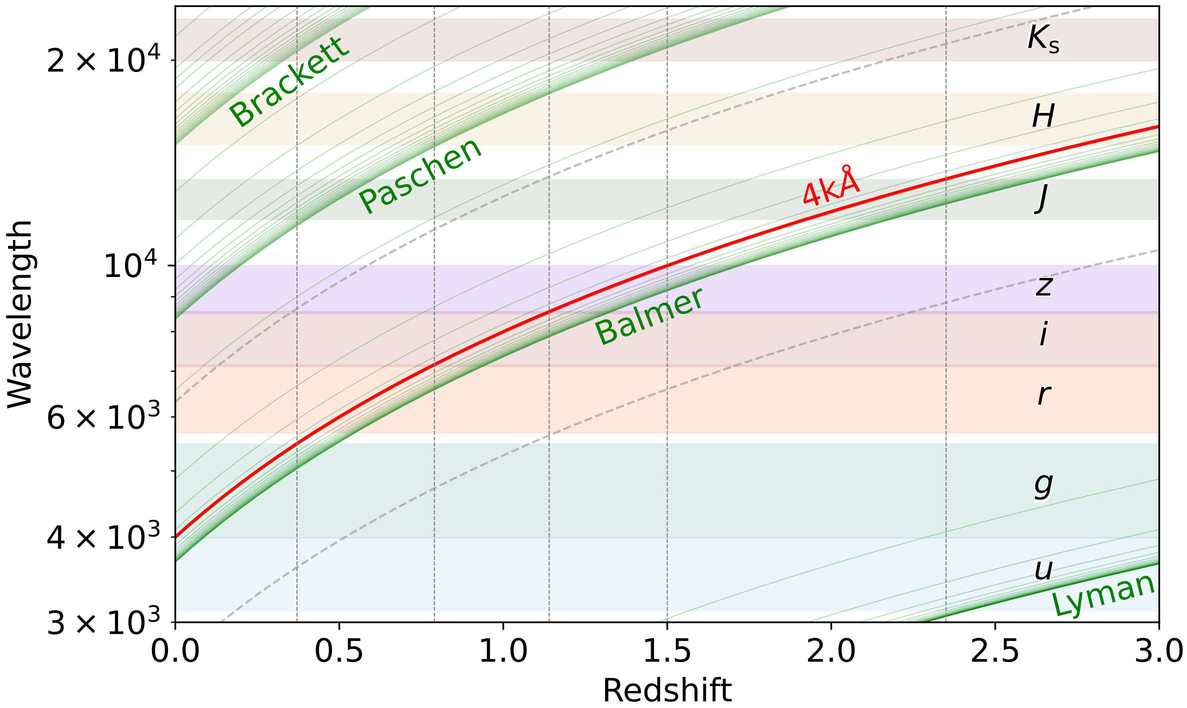


Figure 4.1: Redshift drift of several galactic spectral features, compared to photometric bands from DECam and VIRCAM. In red is the 4kÅ break, the strongest feature in the RS and BC. Vertical grey lines indicate redshifts at which the break exits each band. Green curves display Hydrogen spectral lines, while grey dashed curves show Fe ii (2635 Å) and [O i] (6310 Å) emission lines (discussed in §4.5.2).

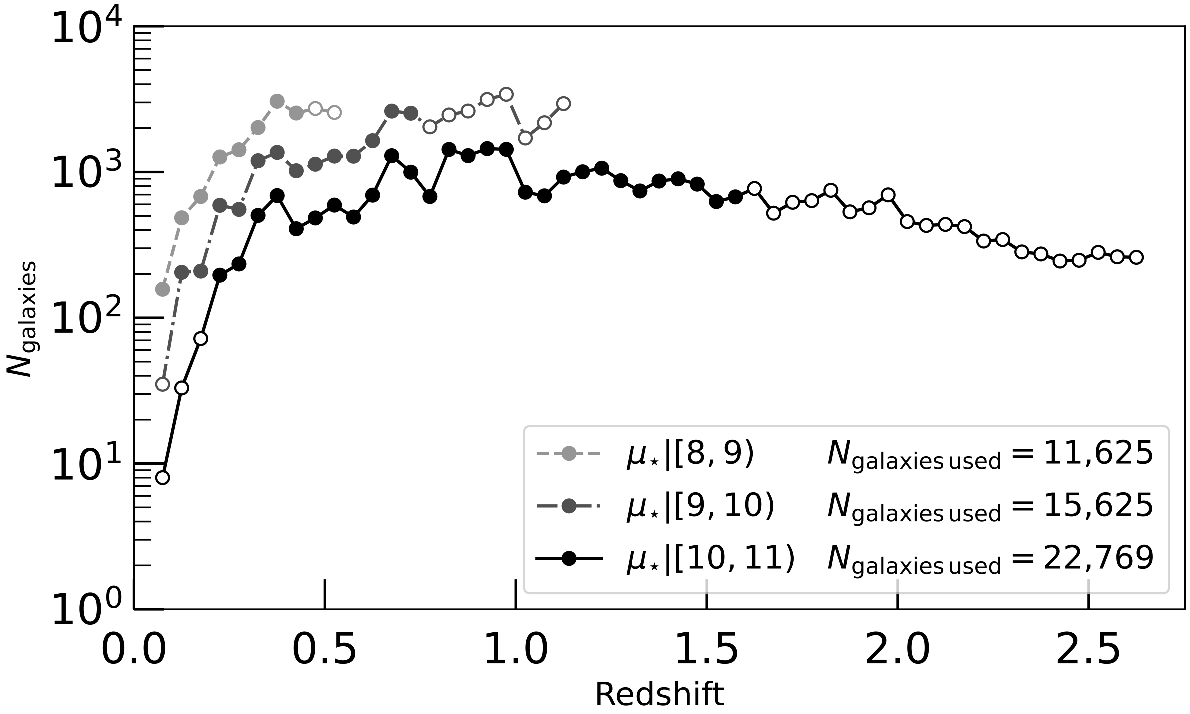


Figure 4.2: Galaxy count per redshift and mass bin. For each decadal stellar mass bin used in this analysis, count of galaxies presented per each $\Delta z = 0.05$ bin (width used here for RD fitting). The legend includes galaxy count of each sub-sample used for core results. Open circles indicate bins with insufficient count of quiescent galaxies, which were not displayed in the main results that follow.

complete at a given threshold. A lower stellar mass completion requirement therefore limits us to lower redshifts whereas a high stellar mass completion limit extends the sample to higher redshifts.

Low-mass galaxies are more abundant than high-mass galaxies, yet a lower mass threshold decreases the redshift range; these two factors compete to yield a peak number of galaxies when using a mass completion limit of $\mu_{\star,\text{complete}} \sim 9.14$, extending out to $z = 0.56$, with a total galaxy count of $N \sim 54,000$. For our analysis, we use three stellar mass decades in that neighborhood, spanning $\mu_\star|[8, 11)$. The minimum masses of each bin $\mu_\star = \{8,\ 9, 10\}$ require maximum redshifts of $z_{\max} = \{.55,\ 1.15,\ 2.65\}$ respectively to ensure mass completion.

Figure 4.2 shows the resulting number counts as a function of redshift for each of these three mass decades. Counts rise as more volume is enclosed, then gradually falls with redshift. As expected for mass-complete samples, we see significantly more low-mass galaxies than high-mass galaxies at a fixed redshift.

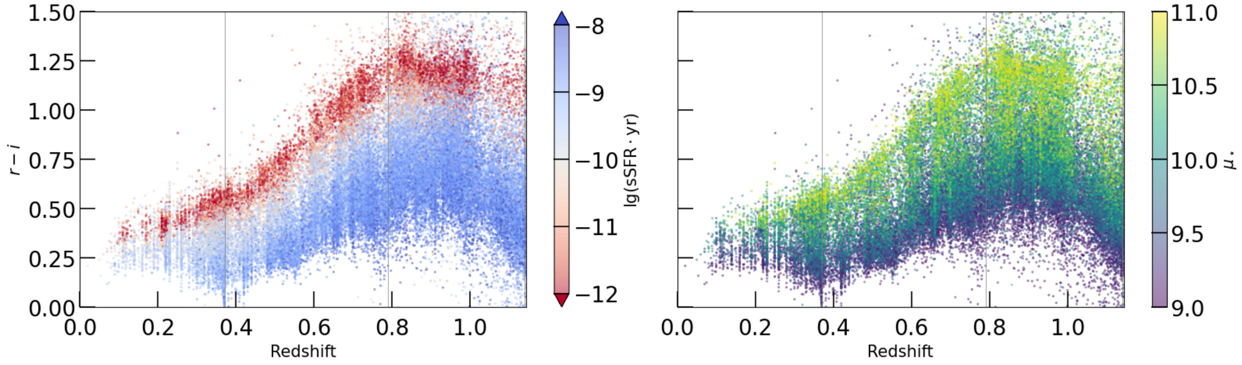


Figure 4.3: Distribution of galaxy photometric color $r-i$ over redshift from the COSMOS2015 dataset. Colored by the physical properties of specific star formation rate (left) and stellar mass (right). As in Figure 4.1, vertical grey lines indicate redshifts at which the 4kÅ break leaves each photometric band. Mass-complete down to $\mu_\star = 9$ and therefore redshift-limited to $z < 1.14$. Compare to Red Dragon selection of the Red Sequence in Figure 4.9.

4.3.3 Redshift limits

In order to ensure trustworthiness of fitted data, we further limit our focus to only those redshift bins in which reside a sufficient population of galaxies, excluding any bin with < 100 galaxies.

In order to ensure we constrain the quiescent population of galaxies, we further demand that there be a sufficient quantity of low sSFR galaxies in each redshift bin. We use the simple definition from Ilbert et al. [2013] of

$$\lg \mathrm{sSFR} \cdot \mathrm{yr} < -11 \tag{4.6}$$

as a definition of quiescent galaxies. We exclude redshift bins with significantly (allowing for $\pm 3\sigma$ Poisson uncertainty) less than 100 quiescent galaxies (as defined by equation (4.6)), ensuring a sufficient population exists within each bin.

These requirements limit the redshift extent of each mass bin from that shown in Figure 4.2 down to $z|[.05, .45)$, $z|[.1, .75)$, and $z|[.2, 1.6)$ for the mass bins in increasing order (solid points on the figure).

4.3.4 Color selection

Myles et al. [2021] analyzed DES Y3 data for weak lensing analyses and enumerated several important limits to enforce in order to exclude unphysical results. Relevant to our analysis, they restrict their color ranges to exclude unphysical colors ("assumed to be caused by catastrophic flux measurement failures"). Mirroring their selection, we restrict each color of our primary color vector

$$\vec{c} = [g-r,\, r-i,\, i-z] \tag{4.7}$$

and of our extended primary color vector

$$\vec{c} = [u - g,\ g - r,\ r - i,\ i - z,\ z - J,\ J - H,\ H - K_S] \tag{4.8}$$

to be within the range $[-1.5, 4.0]$. These bounds turn out to be quite generous, as we find scatter off of mean colors ($\vec{c} \pm 5\vec{\sigma}$) lies entirely in the range $(-.75, 3.0)$ across all redshifts considered. Roughly 1% of galaxies in our main sample are removed by this cut.

Figure 4.3 displays $r - i$ color across redshift for a mass-complete sample of galaxies, as colored by the physical properties of specific star formation rate (sSFR) and log stellar mass ($\mu_\star$). Note that, especially at high redshift, some low-mass and star-forming galaxies are redder in $r - i$ than the core RS, indicating significant scatter in observed color at fixed physical properties.

4.4 Results

Here we display and discuss several results of running Red Dragon (RD) on DES deep-field data from the COSMOS patch. In §4.4.1 we give RD fit parameterization for the main DES bands *griz*, including component weights w, mean colors μ, scatters σ, and correlations ρ between photometric colors for both RS and BC galaxy populations. In §4.4.2 we use the RD fit to characterize RS membership probabilities P_{RS}, then use those probabilities to measure median sSFR and galactic age for each population.

4.4.1 Red Dragon fit parameterization

In this section, we describe the results of fitting RS & BC galaxy populations from the COSMOS field with Red Dragon. Only fits using the DES main bands *griz* are shown here for simplicity of visualization and discussion, with a focus on $r - i$; see fits to other mean colors, scatters, and correlations in Appendix G. In order to show dependence of fit parameterization with stellar mass, we simultaneously visualize these fits for each of the three aforementioned decadal stellar mass bins (see Figure 4.2).

4.4.1.1 Component weights

Figure 4.4 shows the fraction of red galaxies f_{RS} as a function of redshift for each of the three decadal mass bins. For the lower two mass bins, we see clear decreasing trends, as expected due to the continual evolution of BC galaxies towards RS galaxies. Though in the highest-mass bin we find a net negative slope ($df_{\mathrm{RS}}/dz = -.09 \pm .03$), the population exhibits less consistent redshift evolution compared to the lower-mass bins.

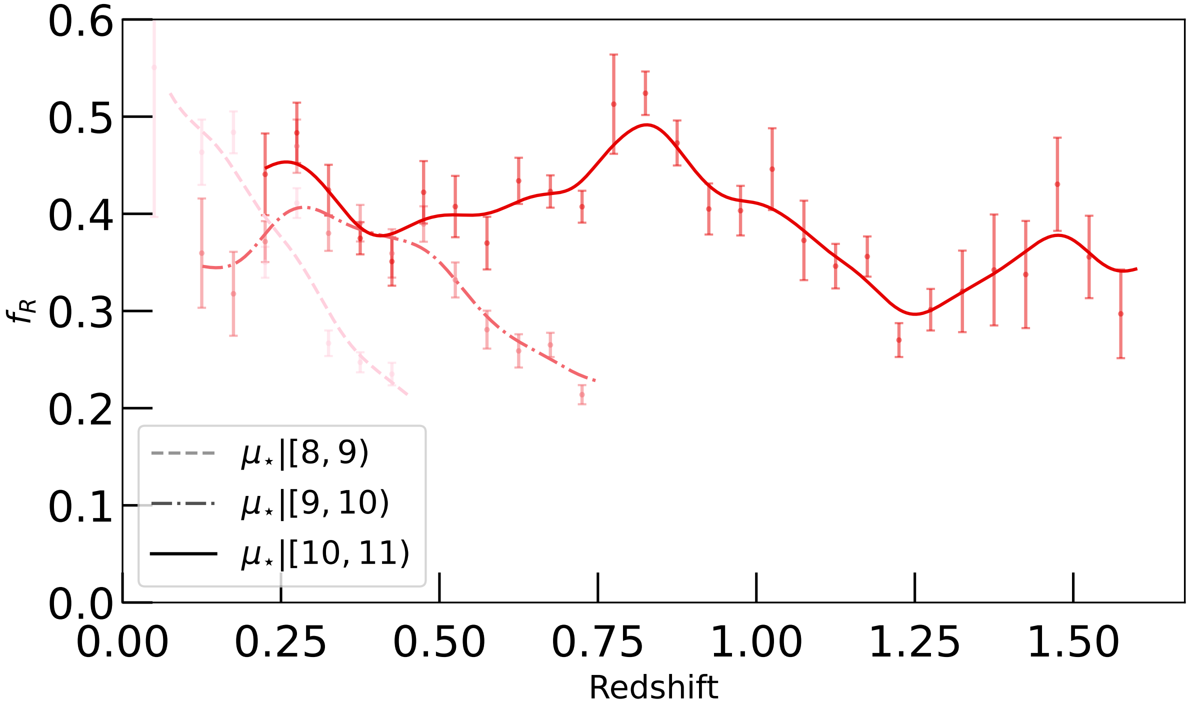


Figure 4.4: RD-measured redshift evolution of component weights for RS, i.e. red fraction f_{RS}. Shown for each of the three decadal stellar mass bins used in this analysis; coloring gets lighter and lines get sparser with decreasing mass. KLLR kernel widths are $\sigma_z = .05$ in redshift.

The fraction of red galaxies in a given sample depends heavily on how you select your galaxies. The probability of a galaxy being red increases as you move nearer to cluster centers (denser environments), at higher stellar masses, and at lower redshifts (i.e. later times, as BC galaxies continuously evolve towards RS galaxies) [Butcher and Oemler, 1978, Madau et al., 1996, Connolly et al., 1997, Baldry et al., 2004, Balogh et al., 2004, Peng et al., 2010, Madau and Dickinson, 2014]. While we focus this analysis on the latter two effects for our patch of the sky considered, proximity to clusters indirectly effects $f_{\rm RS}$, due to cosmic variance.

Because the COSMOS patch is a relatively small patch on the sky, it is highly subject to matter density fluctuations on the length scale subtended. Figure 4.3 visually depicts this, with individual clusters identifiable at low redshifts ($z \lesssim .45$) as vertical lines in the color–redshift plots (indicating galaxies within the same galaxy cluster, lying at near-identical redshifts). At $z < .2$, the square root of the COSMOS field's small area on the sky is over 100 times smaller than the scale of homogeneity [$\sim$ 100 Mpc$/h/(1+z)$; see Avila et al., 2022]. Due to the increased variability in large-scale structure in smaller volumes as well as increased uncertainty due to low-count Poisson uncertainty, we see increases in red fraction variability at the lowest redshifts ($z \lesssim .25$). Because voids take up more of the universe's volume than walls, filaments, and nodes, we're more likely to find field galaxies than cluster galaxies in small volumes. This effect is especially pronounced for luminous red galaxies, as they are far more likely found at large nodes than dim or blue galaxies. As field galaxies are more likely blue than red, this is likely to depress the red fraction at low redshifts. (Contrariwise, if a cluster *does* happen to be found in the small volume, especially at low redshifts, red fraction would be boosted tremendously, especially for high-mass galaxies.)

The slight peak of $f_{\rm RS}$ in the high-mass sample near $z = .8$ is driven by a sudden dearth of BC galaxies at $z = .775$ and a sudden surplus of RS galaxies at $z = .825$; the significance of the peak is low compared to its neighbors, but $> 5\sigma$ significantly larger than $z \sim .3$ values of red fraction. This divergence from Butcher-Oemler expectations is likely driven by cosmic variance, which is particularly strong at high masses. At $z \sim .8$, the COSMOS patch has only reached a diameter of $\sim$ 30 Mpc, less than half the scale of homogeneity ($\sim$ 79 Mpc at that redshift), implying that the survey is still highly susceptible to fluctuations due to cosmic variance (especially so at high masses). Therefore, the peak in $f_{\rm RS}$ for the high mass sample around $z = .8$ should not be taken as an indication of a universal increase at that time.

4.4.1.2 Mean colors

If we assume perfectly neighboring and square photometric bands, and assume a galaxy's spectrum is a step function at the 4000 Å break, then the resulting photometric color (as a function of z) is a single triangle wave: rising from zero as the break moves into the shorter-wavelength band then falling back towards zero as the break moves through the longer-wavelength band. Despite the

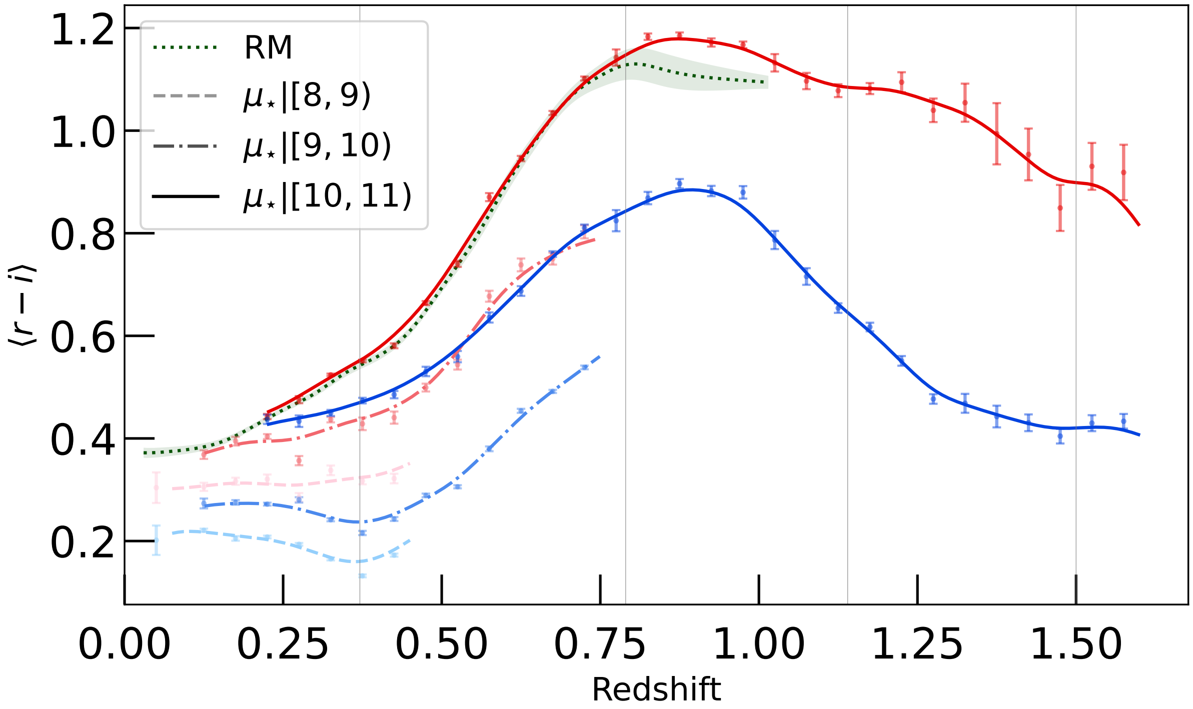


Figure 4.5: RD-measured mean component $r-i$ color for each decadal mass bin (mass-complete out to the redshifts shown). RS shown in tints of red; BC shown in tints of blue; coloring grows lighter with decreasing mass. The dotted green line is a fit from the redMaPPer algorithm to the RS; as colors are magnitude-dependent, we show for the mean and scatter of magnitudes in our highest-mass sample the estimated fit. As in Figure 4.1, vertical lines indicate 4kÅ exit redshifts for *griz* bands. See Figures G.1 & G.2 for $\langle g-r\rangle$ and $\langle i-z\rangle$ characterizations.

extreme simplicity of this model, it roughly explains the location of the observed major peak for each color.

Figure 4.5 shows measured mean $r-i$ color evolution for each of the three mass bins for both RS and BC. We see the expected rising behavior starting at $z \sim .4$ (near where the 4kÅ break enters r band), the peak at $z \sim .8$ (near where the break exits r band and enters i band), and falling behavior afterward. Imperfections and asymmetries in filters and other features in galactic spectra make mean measured color diverge from the simple triangle wave model; see §4.5.2 for a more detailed analysis of features observed in measured mean colors. Both RS and BC are known to redden nonlinearly with increasing galactic stellar mass [though for $\mu_\star > 10.3$ a linear model may suffice; see Baldry et al., 2004, Balogh et al., 2004]; our results agree with this finding, as measured mean colors at a given redshift consistently redden with mass.

As mentioned previously, we compare RS mean color and scatter to measurements made by redMaPPer (RM; fits made available by E. Rykoff 2022, private comm.). Mean colors drift

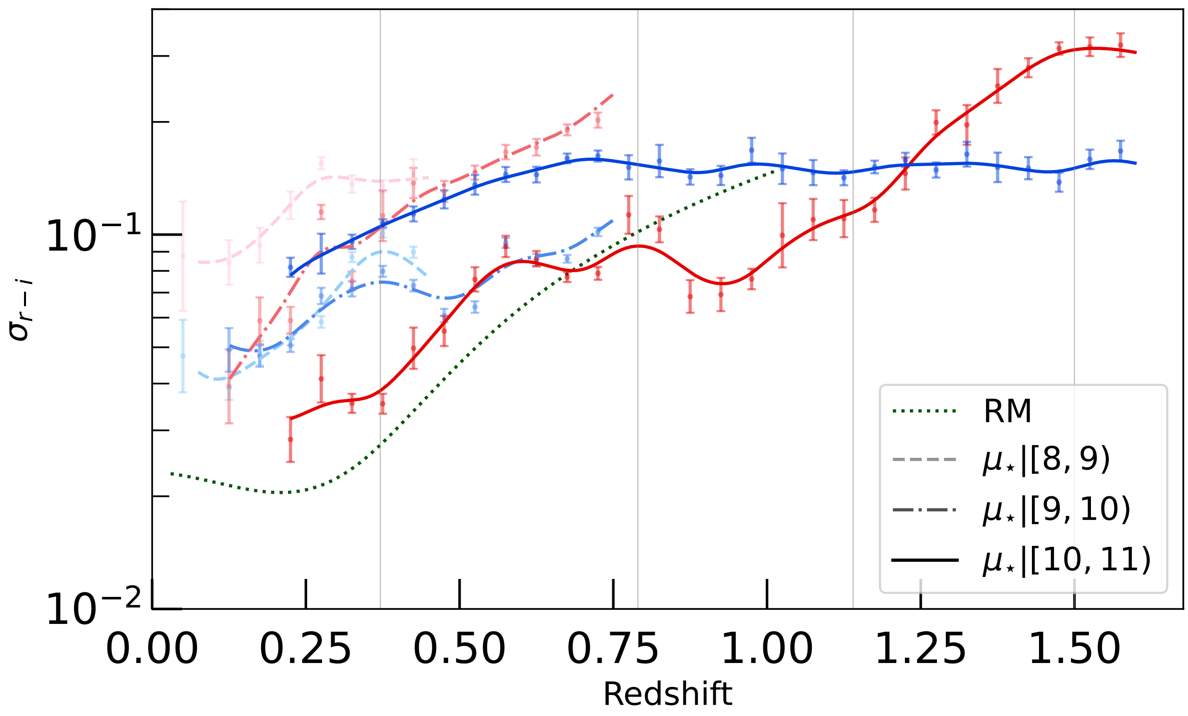


Figure 4.6: RD-measured intrinsic color scatter in $r - i$; lines as in Figure 4.5. (RM scatters do not run with stellar mass, hence only the single curve.) See Figures G.3 & G.4 for σ_{g-r} and σ_{i-z} characterizations.

linearly with z-band magnitude m_z, so to compare with our mass-binned sub-samples, we use the characteristic magnitude $m_{*,z}(z)$ of Rykoff et al. [2014] to measure mean and scatter of $m_z - m_{*,z}(z) = 1.26 \pm .88$ in the highest mass bin, which is used to plot expectations. We note that even in this highest mass bin, 27% of galaxies are fainter than the RM minimum luminosity threshold (of $m_z - m_{*,z}(z) = 1.75$), so we focus our RM comparison exclusively on the highest-mass bin. We find reasonably good agreement, though with some divergence at higher redshifts.

Fits for $\langle g - r\rangle$ and $\langle i - z\rangle$ are included in Appendix G. We see several relatively consistent features across each color, albeit offset in redshift (due to the differing rest frame band wavelength ranges), discussed in §4.5.2. Each of the color plots roughly agree with fits from B22 (up to its terminal redshift of $z = .84$), each diverging by $\lesssim .1$ mag. While $\langle i - z\rangle_{\rm RS}$ matches the RM fit well, $\langle g - r\rangle_{\rm RS}$ differs significantly beyond $z \sim .75$, as the RM fit diverges towards bluer colors.

4.4.1.3 Color scatters

Figure 4.6 details the RD fitting of $r - i$ scatter as it evolves with redshift for each stellar mass bin, for both RS and BC components. The most massive stellar mass bin performs similar to the RM fit, consistent within a factor of $\lesssim 2$ for $\mu_\star|[10, 11)$ galaxies. In contrast, the lower-mass bins showed considerably larger scatter than high-mass galaxies, by a factor of $\gtrsim 2$, as expected from Baldry et al. [2004].

Scatters for $g - r$ and $i - z$ are shown in Appendix G. Across the board, RS color scatters tend to increase with redshift. BC scatters tend to increase until the color's transition redshift, after which it is roughly constant. This implies that at wavelengths shorter than the 4kÅ break, BC spectra have constant scatter, indicating uniformity of the BC across these regions. For the RS, scatters almost always reduced with increasing $\mu_\star$, but BC scatters exhibited non-monotonic behavior with $\mu_\star$. Baldry et al. [2004] measured non-monotonic behavior within $\sim .23$ dex of $\mu_\star \sim 9.95$; our results broadly agree.

For each color of the high-mass galaxy sample, the RS scatter begins to exceed BC scatter near the redshift where the 4kÅ break exits a color's subtrahend. Here, for $r - i$, this is at $z \sim 1.2$, where the 4kÅ break exits i-band. In contrast, for the lightest galaxies, we found consistently that $\sigma_{\text{RS}} \gtrsim \sigma_{\text{BC}}$ for their full redshift extent. This could be caused by the plurality of quenching mechanisms present at low galactic mass; while high-mass galaxies ($\mu_\star \gtrsim 10.5$) are essentially all quenched by AGN, lower-mass galaxies are quenched by a variety of mechanisms, such as supernovae, stellar winds, or reionization [Wechsler and Tinker, 2018]. The increased variety in quenching mechanisms could be the driving reason behind increased RS scatter towards lower masses. The narrowness of the RS was first observed for bright galaxies; as most all galaxies even in the highest mass bin were below $.4\,L_*$, it should be relatively unsurprising that our findings in lower mass bins run contrary to the high-mass expectations [especially considering the expectations of increased RS scatter at lower stellar masses; see Baldry et al., 2004, Fig. 5].

4.4.1.4 Color correlations

Positive intrinsic inter-color correlations here imply that if one color is redder than average (i.e. the longer-wavelength band captures more net flux), then the other color considered will also be redder than average. Perfect correlation would come from a pure power-law spectrum, implying no significant divergences between color behaviors. Negative correlations imply that if one color is redder, the other color will be bluer. If the correlation is between overlapping colors, such as $\rho(g - r,\ r - i)$, where the two colors share the middle band r, then negative correlations could be caused by variable peaks in the spectrum within the shared band. If a narrow variable feature (e.g. oxygen emission line) existed in r for an otherwise power-law spectrum, then for $\rho(g - r,\ r - i)$ as the feature is redshifted through g-band the correlation would drop to null; once the feature enters r-band, the correlation would drop to -1; as it then enters and passes through i-band, it would symmetrically return towards high correlations.

Figure 4.7 shows redshift evolution of intrinsic inter-color correlations between the overlapping

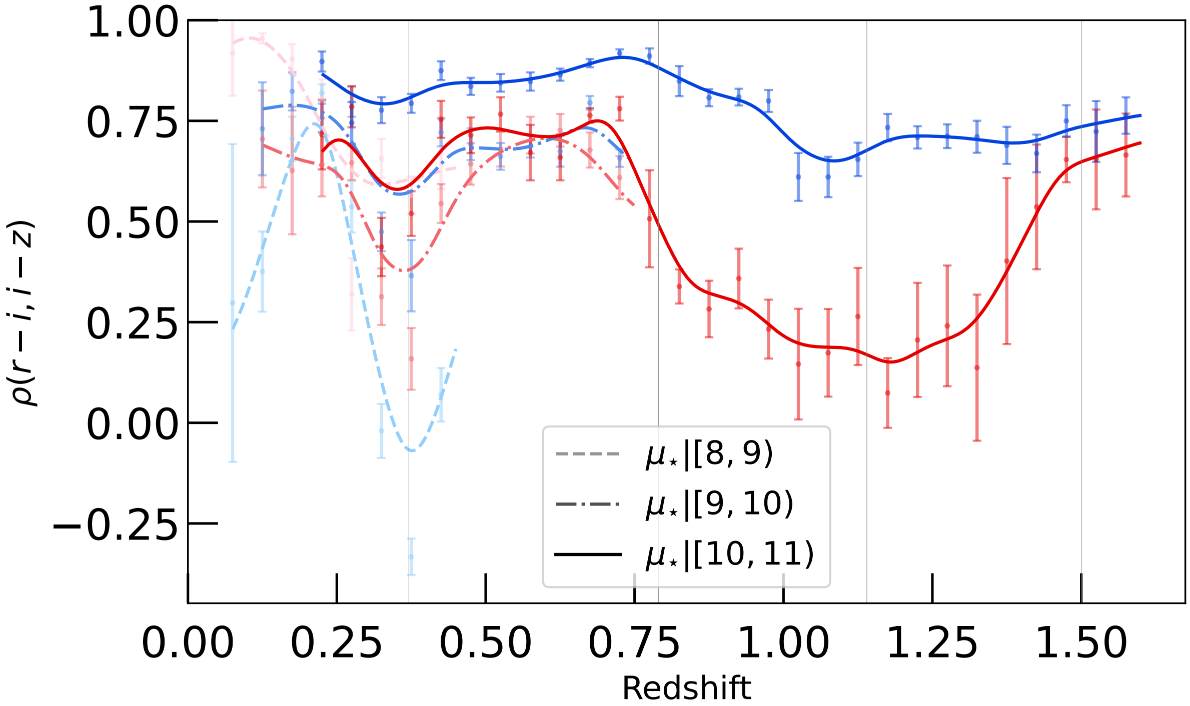


Figure 4.7: RD-measured inter-color correlation between $r-i$ and $i-z$ photometric colors; RS and BC lines for each mass bin as in Figure 4.5. Across all color combinations, BC color correlations ρ_{BC} on average are roughly 10% larger than RS color correlations ρ_{RS} ($\sim$ 25% larger for high-mass galaxies). See Figures G.5 & G.6 for $\rho(g-r,\ r-i)$ & $\rho(g-r,\ i-z)$ characterizations.

colors $r - i$ and $i - z$ for each mass bin and for each component as they evolve across redshift. Figures G.5 & G.6 show correlations between the other two color combinations.

Broadly speaking, correlations tend to decrease over redshift ($d\rho/dz \sim -.16$) and increase with stellar mass ($d\rho/d\mu_\star \sim +.12$). We find that the RS inter-color correlations ρ_{RS} tend to be less than or similar to those of the BC ρ_{BC}. This is most apparent for the high-mass sample, with BC correlations significantly larger than the RS; in contrast, the RS correlations of the low-mass sample are occasionally larger than that of the high-mass sample. Very roughly, we find inter-color correlation for the RS $\rho_{\mathrm{RS}} \sim 60\%$; in contrast, we find $\rho_{\mathrm{BC}} \sim 75\%$, still lower than expectations from SPS modeling.

Comparing to Figure 4.1, we see that the wavelength region of r to z in the redshift range of $z|[.776, 1.35]$ is quite tumultuous, with the 4kÅ break and other narrow features passing through i-band, the shared band of this color correlation. Variation in the strengths of the Balmer series and 4kÅ break (e.g. caused by differences in metallicities or temperature) may explain the lower correlation in the RS seen in this redshift region.

As mentioned earlier, a dip in correlation of overlapping colors could be caused by a narrow feature with variable intensity. The dip in correlation at $z = .38$ could thus be caused by a strongly varying emission line near the central wavelength of i-band at that redshift, namely $\lambda_c/(1+z) = 7835$ Å$/(1+.38) \sim 4100$ Å, which incidentally matches the H-δ line. Strong variation in H-δ strength could thus cause such a feature.

Due to the skew-normal distribution of $\lg \mathrm{sSFR} \cdot \mathrm{yr}$, the star-forming population has relatively tightly-knit spectral features compared to the quiescent population (which approach null star formation). This could in part drive the stronger correlations between BC colors as compared to RS colors. Furthermore, there are various quenching pathways from the BC to the RS, be it by aging or feedback or merging. These various paths to the RS from the BC could imprint themselves differently on various RS spectra, leading to increased variability (and therefore decreased correlation) in RS colors as compared to BC colors.

4.4.2 Red Dragon fit results

With the photometric fits to the galaxy populations in hand, we can calculate component membership likelihoods for each galaxy. In particular, we focus on the RS membership probability P_{RS} (see eqn. (4.4)), using $P_{\mathrm{RS}} \geq .5$ to distinguish RS from BC galaxies. Such binary classification allows us to measure redshift evolution of sSFR values and galactic age for each population, characterizing typical quiescent galaxies as well as tracing the star-forming main sequence.

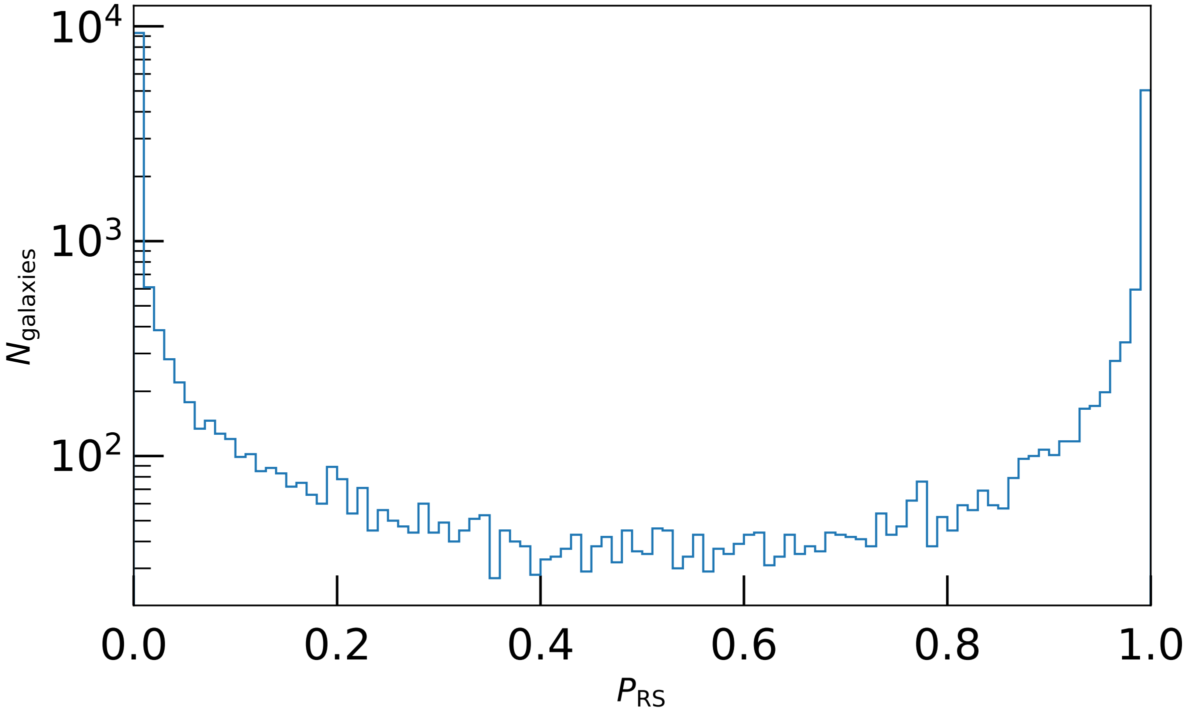


Figure 4.8: Histogram of P_{RS} values for the highest-mass dragon. Bin width of .01 highlights at $P_{\mathrm{RS}} < .01$ and $P_{\mathrm{RS}} > .99$ the order of magnitude increase, indicating high certainty of characterization for the lion's share of galaxies.

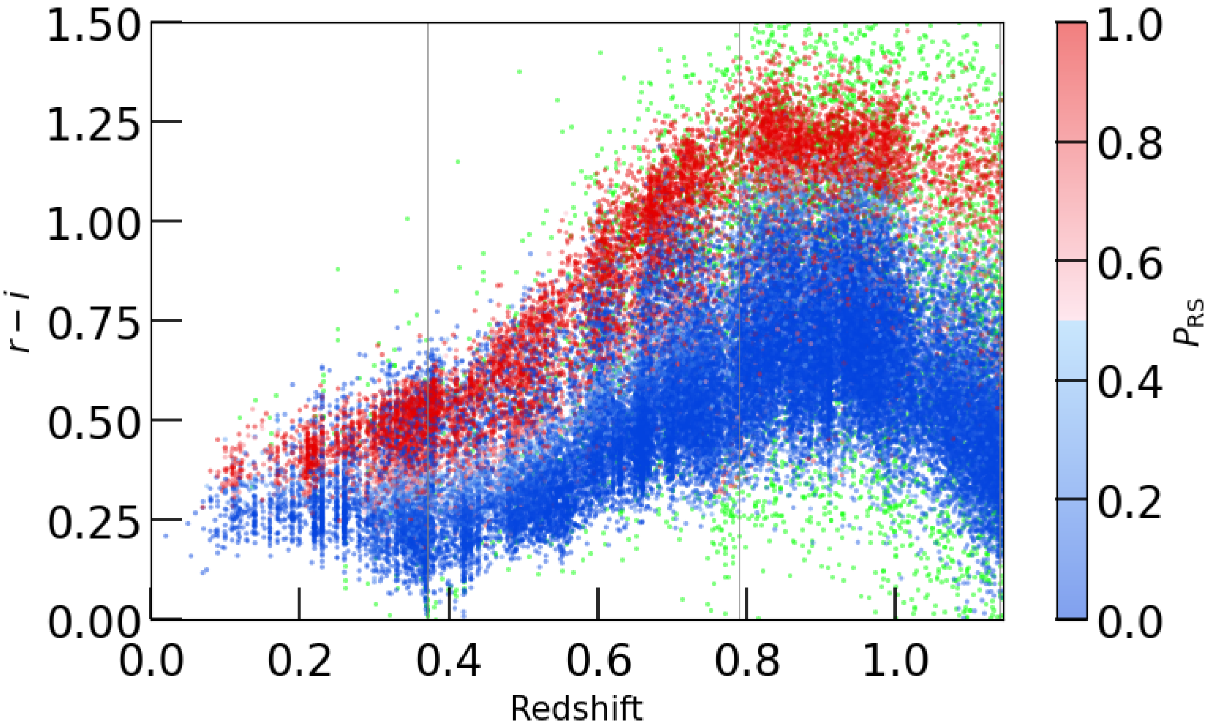


Figure 4.9: Galaxy $(r-i)(z)$ colored by P_{RS}; as panels of Figure 4.3 but with points colored by RS membership probability P_{RS}. Galaxies with $\max(\mathcal{L})$ less than the 10th percentile are marked in green, indicating galaxies which were favored by neither RS nor BC as strong members.

4.4.2.1 Galaxy selection and characterization

Using the RD modeling of COSMOS patch galaxies of §4.4.1, we now ascribe to each galaxy a red sequence membership probability P_{RS} (see equation (4.4)). Figure 4.8 shows for the highest mass bin that these values are strongly bimodal: roughly 90% of galaxies across our three mass bins have probabilities $P \equiv \max(P_{\text{RS}},\ P_{\text{BC}})$ greater than 0.75 and roughly 75% of galaxies have probabilities $P > 0.90$, so this distribution lends itself naturally towards binary selection of the RS and BC. As maximum likelihood values $\mathcal{L}_{\text{max}} \equiv \max(\mathcal{L}_{\text{RS}}, \mathcal{L}_{\text{BC}})$ decrease, galaxies are more likely ascribed middling values of P_{RS}.

Figure 4.9 shows the characterization of P_{RS} for galaxies in the space of $r-i$ across redshift, visually showing the bimodality of probabilities. This selection, visually compared to Figure 4.3, shows that the RD-defined RS here largely isolates the quiescent population.

Shown as green points on this figure are galaxies which didn't fit well to either population. In particular, galaxies with the lowest 10% of $\mathcal{L}$ values were neither likely candidates for RS nor BC. Such galaxies tend towards higher redshifts and had photometric color uncertainties typically $\gtrsim 2.7$ times larger than higher-likelihood galaxies. Intrinsic scatter of this low-$\mathcal{L}$ sample far exceeds that

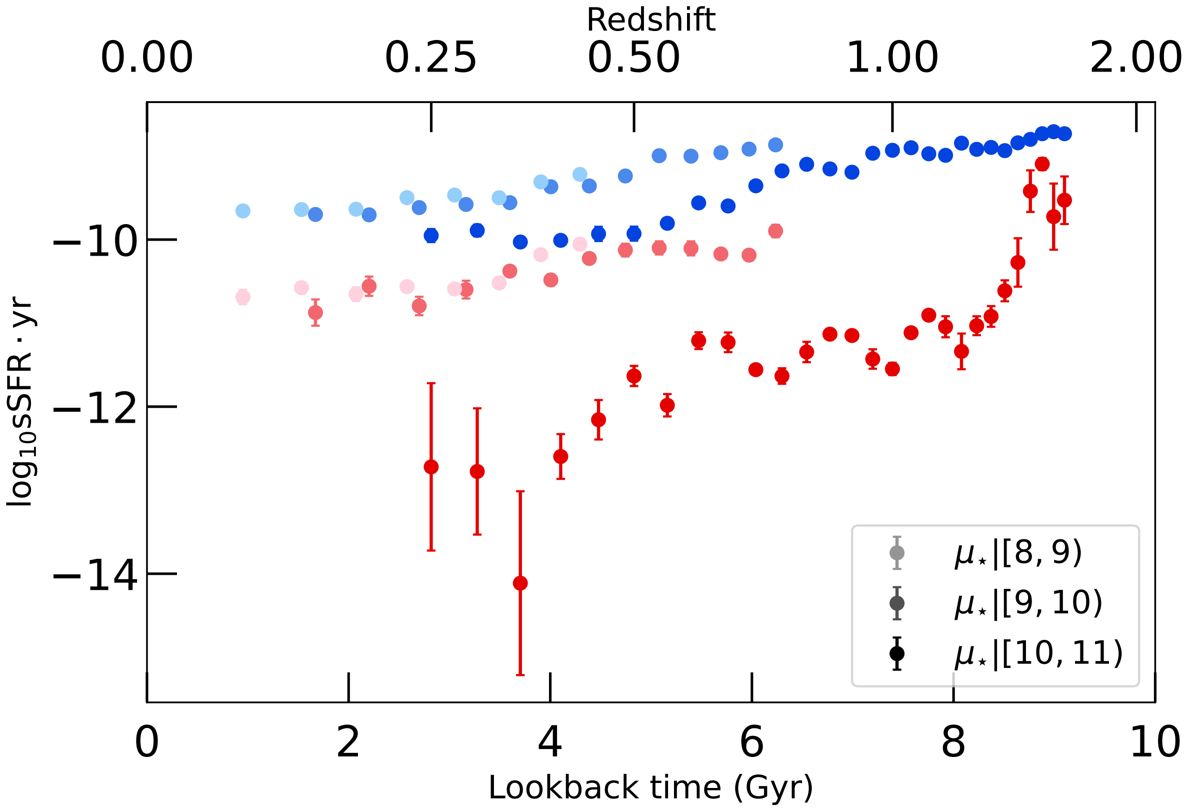


Figure 4.10: Median sSFR values (with bootstrap uncertainty) for RS and BC for each of the three mass-binned samples as they evolve over time. Only redshift bins with $N_{\text{galaxies}} \geq 100$ are shown.

of the RS or BC, indicating that these galaxies tend to be outliers in color space.

4.4.2.2 Evolution of star formation rates

Using binary selection of the RS and BC from P_{RS}, we can now find sSFR values for RS and BC as functions of redshift.

Figure 4.10 shows median sSFR values for each population with bootstrap uncertainties, showing a clear trend of increasing with redshift and decreasing with stellar mass. Each mass bin shows a clear distinction between quiescent and star-forming galaxies, with roughly a factor of 10 to 100 between the two populations (where higher masses tend towards a larger separation in sSFR values). Star formation rates increase with redshift, indicating more active populations in the past, towards cosmic noon (at $z \sim 2$).

The uptick in median sSFR in the high-mass population at high redshifts / early times could relate to the lack of quiescent galaxies at high redshifts. While this could be an actual feature (perhaps related to cosmic noon), this is similar to what would be expected due to a lack of quiescent galaxies, that the RS approaches a noise term which looks more similar to the BC. In the last five redshift bins displayed in Figure 4.10, quiescent galaxies (using the definition of eqn. (4.6), thus differing from RD characterization) make up only $\sim 10\%$ of the population (decaying sharply with redshift beyond that point). At such a low red fraction, it becomes increasingly difficult for GMMs to detect the RS, favoring fitting noise terms or sub-dividing the BC instead. Until we have additional high-quality data to draw from at high redshifts, forming a sizeable RS population, it will be difficult to discern.

The star-forming main sequence (SFMS) is roughly fit by

$$\lg(\text{sSFR} \cdot \text{yr})_{\text{BC}} = -10 + (.2\ \text{Gyr}^{-1})\, t \tag{4.9}$$

(where t is lookback time). The time slope of this relation is in rough agreement with findings from Speagle et al. [2014], who found the time slope ranged in $[.098, .176]$ Gyr^{-1} over our mass range. A slightly more precise quiescent definition than equation (4.6) could use 1 dex below this fit as a dividing line between RS and BC, allowing for time evolution of the quiescent population. However, as discussed in Leja et al. [2022, see §7.2 therein], enforcing a hard cut selection of the RS and BC using a threshold in sSFR has considerable downsides: as sSFR is skew-lognormal (rather than bimodal, as are galactic colors), a slight shift in threshold can lead to a significant shift in population characterization. This favors using photometry to characterize population sSFR values, rather than the converse, of using sSFR values to characterize RS & BC populations.

4.4.2.3 Evolution of galactic age

Similar to the previous section, we can map out median estimated galactic age for each population (typical length of time since the galaxy's formation). This is shown in Figure 4.11, mirroring the style of Figure 4.10. Individual galaxy ages lack uncertainty estimates here, so these findings should be interpreted with discretion. We find that RD-selected RS galaxies in each mass bin are consistently older than BC galaxies, by $O(1\ \text{Gyr})$; this is the expected quenching timescale [Bell et al., 2004, Blanton, 2006].

If all galaxies in a certain stellar mass bin were born at the same time and didn't grow or merge enough to leave the mass bin, nor move from BC to RS, then we would expect a slope of negative one on the plot for each population, indicated by the diagonal grey lines. Though not universal, slopes tend to be slightly shallower than the assumption of a single creation epoch and unchanging population membership. This indicates that the populations are being joined by younger galaxies (more recently formed) or that older galaxies are leaving the mass bin. (In contrast, steeper slopes indicate the reverse, that the mass bin is either joined by older galaxies or that younger galaxies are leaving the mass bin.)

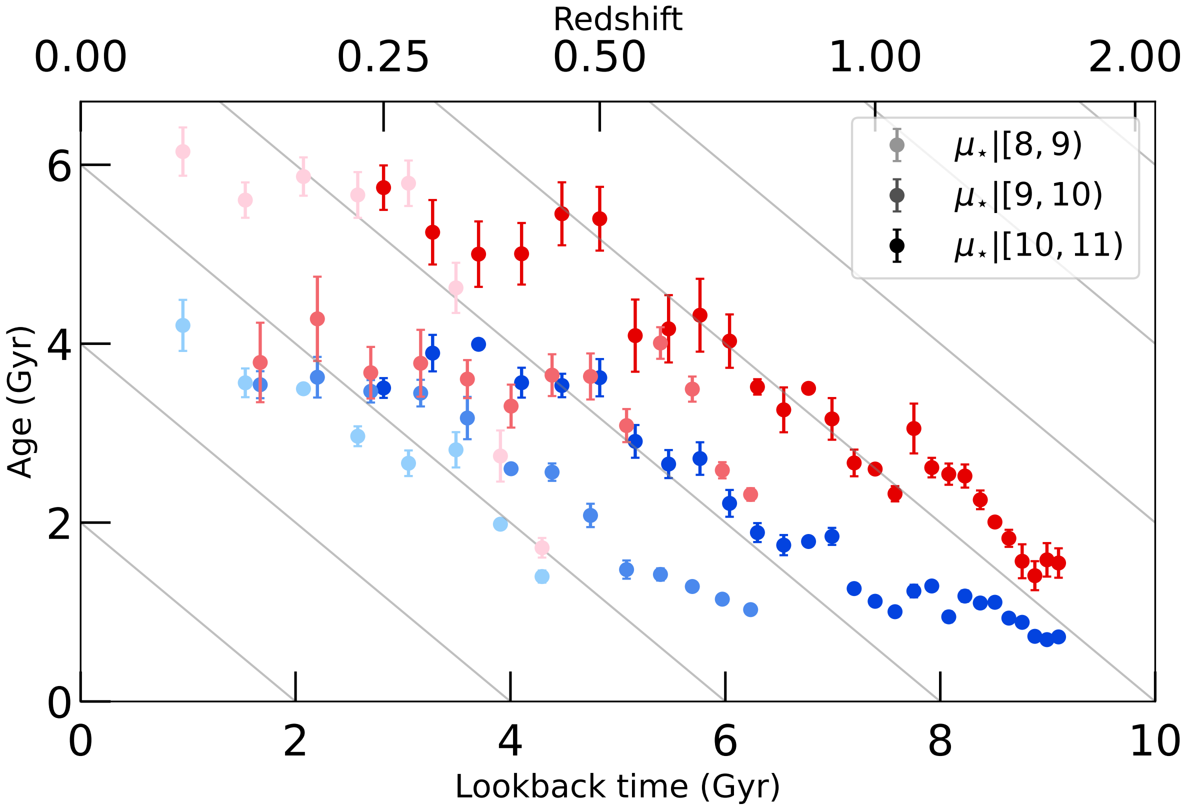


Figure 4.11: Median galactic age, formatted as Figure 4.10. Diagonal grey lines indicate equal-epoch growth.

Focusing on the high-mass sample (dark points), RS and BC show distinct trends, even out to high redshift. This is somewhat in contrast to the high-mass high-redshift regime of Figure 4.10, where the RS seems to move towards the BC in what could be a degradation towards noise or a sub-division of the BC. The significant distinction of ages in Figure 4.11 indicates that Red Dragon is still selecting significantly different populations in such circumstances.

4.5 Discussion

In this section, we discuss the accuracy wherewith RD selects the quiescent population (§4.5.1), the interpretation of mean color characterizations (§4.5.2), and the choice of using two components vs more to characterize galaxy populations (§4.5.3).

4.5.1 Accuracy in selecting the quiescent population

Rather than use RD selection of the RS as a 'truth' to characterize median sSFR values, we can use sSFR values as a 'truth' wherewith to select the RS, and measure RD selection accuracy of this quiescent population.

Ilbert et al. [2013] used $\lg \mathrm{sSFR} \cdot \mathrm{yr} < -11$ as a truth label for the RS and measured selection accuracy as a function of redshift. Using *balanced accuracy* (bACC; the average of specificity and sensitivity) to measure this, their two-color hard-cut selection of the RS results in a bACC which is consistent with linearly decaying over redshift, moving from $\sim 95\%$ accuracy at redshift zero down such that at $z = 2.5$ their accuracy was $\sim 65\%$ (where random selection results in a bACC of 50%).

For comparison, we allow dragons to fit populations out to high redshift, beyond the point of losing a large population of $\lg(\mathrm{sSFR} \cdot \mathrm{yr}) < -11$ galaxies. To increase RS size, we fit all $\mu_\star > 9$ galaxies here, rather than limit ourselves to a single mass decade. To account for the significant evolution of GMM parameters with stellar mass for the BC in this sample, we perform a three-component fit. Though these fits may degrade somewhat towards higher redshifts (as discussed in §4.4.2.2), we will show that the fits retain utility nonetheless.

Figure 4.12 compares balanced accuracies of selecting quiescent (equation (4.6)) galaxies as measured by the two-color selection of Ilbert et al. [2013] (black errorbar points; no uncertainties in bACC available) as compared to the Red Dragon algorithm, for both DES main photometry $griz$ (blue points) and extended photometry $ugriz + JHK_\mathrm{s}$ (orange points) models of the RS. In contrast to the linearly decaying RS selection accuracy of Ilbert et al. [2013], we find consistently high accuracies of RS selection using Red Dragon. Across the board, we find bACC $\sim 95\%$ for the two dragons, with the extended-photometry dragon outperforming the $griz$ dragon significantly at the

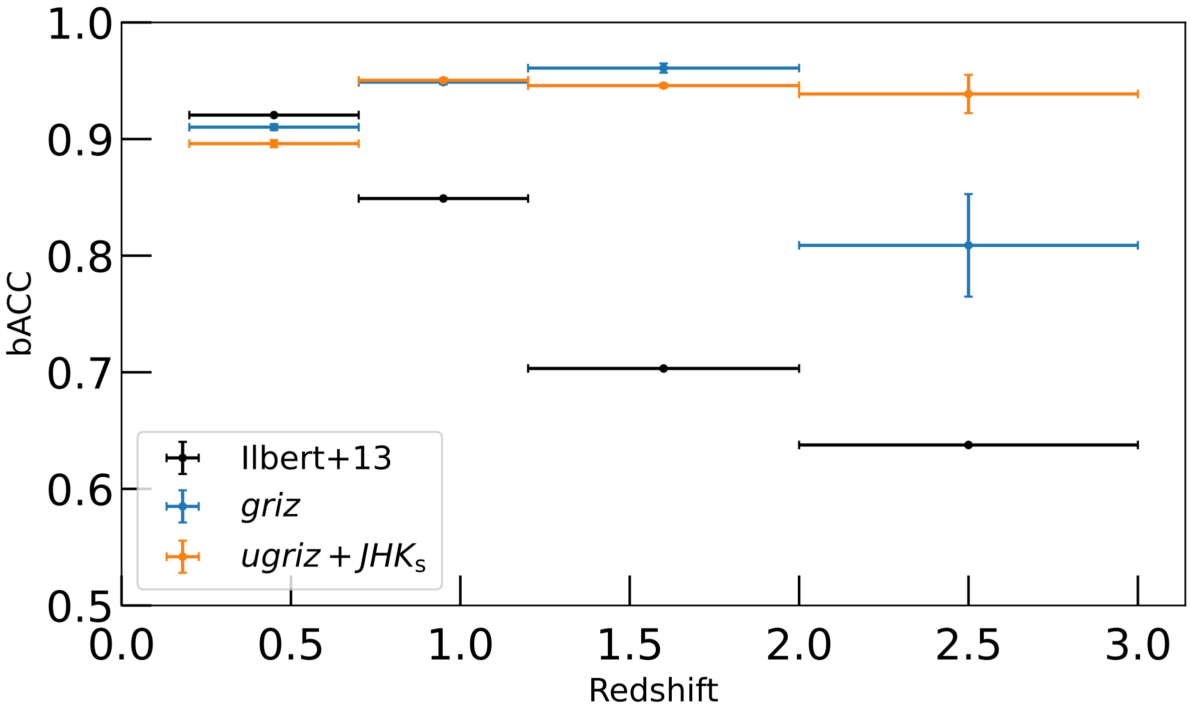


Figure 4.12: Balanced accuracy in selecting quiescent galaxies (see equation (4.6)). Shown for $K = 3$ dragons with standard DES bands ($griz$; blue) and extended photometry ($ugriz + JHK_s$; orange), trained on and evaluating $\mu_\star > 9$ galaxies. Ilbert et al. [2013] selection show in black (no uncertainties available). Note that uncertainties in sSFR are not well measured and are not taken into account in this analysis.

highest redshifts. Though the highest redshift bin has very few[2] quiescent galaxies, and therefore has wider uncertainty in bACC, using RD in the intermediate redshift bins spanning $z|(0.7, 2.0)$ yields significant gains in accuracy. Particularly at $z \sim 1.6$, we see a gain in accuracy of $\sim 25\%$, indicating far superior selection of the quiescent population. We therefore see clear superiority in GMM selection of the RS as compared to using hard cuts in color–color space, especially at higher redshifts.

As cautioned in Leja et al. [2022, see §7.2 therein], due to the skew-lognormal distribution of sSFR, hard cut selection of the RS vs BC in sSFR space has considerable dependence on the particular threshold used. For example, wiggling the truth threshold by .3 dex (a factor of two) results in a $\sim$ 5% change in bACC values. Rather than focus on a "quiescent accuracy" of RS selection, we find more utility in using the RS to measure quiescence of its constituent members, as was performed in §4.4.2.2.

4.5.2 Interpretation of mean color parameterization

Photometric colors are proportional to the logarithm of the ratio of integrated fluxes in each band of the color (see equation (4.1)). If photometric filters were infinitesimally narrow box functions, then a photometric color would measure a log spectrum slope at the wavelength between the bands. A measurement of a given color at redshift z would then correspond to a rest wavelength feature at $\lambda_{\rm rest} = \lambda_{\rm tr}/(1 + z)$, where $\lambda_{\rm tr}$ is the transition wavelength between bands—the point where a signal begins being picked up more by the next-longest wavelength neighboring filter (see transition wavelengths in Appendix F). For our photometry, each filter has substantial width, asymmetries (both within and between filters), and irregular edges. Photometric colors therefore measure a smoothed version of the spectrum slope, somewhat deformed due to various filter divergences from ideal, identical box functions. This means that (approximate) spectral slopes measured by different photometric colors may be somewhat offset from each other, despite measuring the same rest wavelength range (even if the spectrum is unchanging over time). This makes comparison of vertical offsets between different colors generally less profitable than comparing relative shapes between colors or same-color differences between mass bins.

Figure 4.13 shows measured mean colors transformed from their measured domain of redshift (as shown in Figures 4.5, G.1, & G.2) to an imputed rest-frame wavelength $\lambda_{\rm rest}$. The most notable feature for both RS & BC is the major peak, aligning near wavelength $\lambda \sim 3900$ Å (corresponding to the major RS and BC peaks seen in Figure 4.5 above, at $z \sim .9$ for $r - i$). In addition, a significant second peak emerges for BC galaxies near $\lambda = 6300$ Å (the small BC bump seen at $z \sim .13$ *ibid.*).

[2] Using equation (4.6), only 58 of 35,216 galaxies are quiescent (roughly 1 in 600 galaxies). However, using 1 dex below equation (4.9) as a quiescent definition, the count drastically increases, to 12,422 (roughly 35% of galaxies).

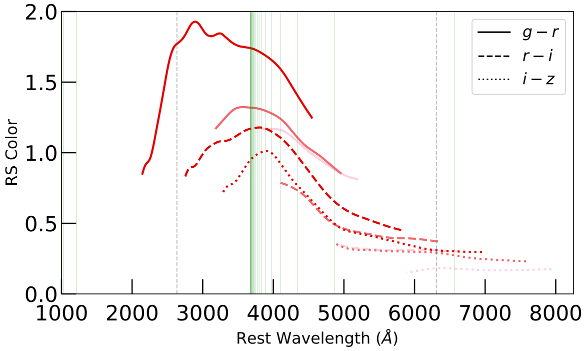


Figure 4.13: Measured mean colors as a function imputed rest-frame wavelength. Lighter colors indicate lower mass bins, as in Figure 4.5, while solid / dashed / dotted lines indicate Green solid lines show hydrogen spectral series lines, with most belonging to the Balmer series (beginning with Hα at 6550 Å). Grey dashed lines show Fe ii (2635 Å) and [O i] (6310 Å) emission lines. Note that at a fixed imputed rest wavelength, each color is measured at a different redshift, with $g-r$ at the lowest redshift and $i-z$ at the highest. Redshift at which each color measurement was made increases to the left (towards shorter rest wavelengths).

Several tertiary features are also present, with unresolved peaks at wavelengths such as 2500 Å and others.

The primary peak, near 3900 Å for both RS and BC, corresponds to the 4kÅ break. This relates directly to the Balmer series termination wavelength (at $\lambda \geq \lambda_B \doteq 3647$ Å) as well as several coincidences of metallic lines. The secondary peak could correspond to the H-α line (at $\lambda \doteq 6550$ Å) of the hydrogen spectral series, but [O i] at 6310 Å also aligns. Tertiary peaks hint at other features such as iron (e.g. Fe ii at 2635 Å), oxygen, and helium emission lines, but it is beyond the scope of this current paper to detail these peaks.

While some differences between colors may be due to filter inhomogeneities, it may be that the differences are driven by redshift evolution. For example, around 3250 Å in the RS plot, not only are different bands and masses offset vertically (which could easily be caused by asymmetries in filter widths), but the curves show distinct shapes, with the $\langle g-r \rangle$ curves redder than other colors (which, at the same imputed rest wavelength, are at higher redshift). This could imply a significant reddening of spectra (steepening of spectral slope) about 3250 Å over cosmic time.

A more careful analysis, correcting for filter irregularities, could better impute the rest-frame spectrum. This could decipher which features are truly detected as well as which differences between colors are redshift-dependent or merely filter-dependent. We leave such detailed analysis for future papers.

4.5.3 Optimal component count

Red Dragon allows for fitting of not only RS and BC, but additionally of any number of Gaussian mixture model components. Hypothetically, one could model either "green valley" galaxies or a noisy background (e.g. from galaxies with bad photo-z estimates), or even further components, dividing RS or BC into sub-populations in order to catch non-Gaussianities.

B22 quantified optimal component count using the Bayesian Information Criterion (BIC). BIC measures relative information loss of different models (run on the same data); it increases with the number of model parameters and decreases with increased maximum model likelihood. BIC thus increases with model complexity and decreases with improved fit, so models with lower BIC better minimize information loss. Log relative likelihoods $\ln \mathcal{L} = (\mathrm{BIC}_a - \mathrm{BIC}_b)/2$ of model b minimizing information loss as compared to model a tend to have values in the hundreds, leading to probabilities of model superiority tending strongly towards zero and one (i.e. ϵ and $1 - \epsilon$, with $\epsilon \lesssim 10^{-20}$). However, bootstrap uncertainties tend to overshadow ΔBIC such that comparison of *significance* of a non-zero difference in BIC tends to be more useful than the measure of $\mathcal{L}$ itself. In the spirit of BIC—minimizing both model complexity as well as information loss—we only prefer a more complicated model if it *significantly* diminishes information loss.

For both $griz$ and $ugriz + JHK_\mathrm{s}$ photometries, we find no significant differences in BIC for our three mass-binned dragons on moving from a two component model ($K = 2$) to more components ($K \geq 3$).[3] The lack of significant preference for a higher-component model implies that the simpler, two-component model should be preferred. While these results hold for our mass resolution of $\Delta\mu_\star = 1$ and our redshift resolution of $\Delta z = .05$, a GMM run on too wide a $\mu_\star$ or z range will fail to model populations well (as is discussed in Appendix H). In such circumstances, a third component may be significantly favored for inclusion.

4.6 Conclusions

We present Red Dragon fitting of the RS and BC populations in the COSMOS2015 dataset. We consider three decadal mass bins which span stellar masses $\mu|[8, 11)$. The mass-complete sample focuses on redshift bins with a significant population of quiescent galaxies. Our model measures mean colors, color scatters, and the covariance between bands for each galactic population.

With these populations characterized, we measure median specific star formation rates and galactic ages (Figures 4.10 & 4.11), finding that the RD-selected RS is consistently more quiescent (by $\gtrsim$ 1 dex) and older (by $\gtrsim$ 1 Gyr) than the BC across all redshift within each mass bin. The

[3] The strongest preference for a $K = 3$ model came from the highest mass bin, using the extended photometry, around $z = .75$, but this preference was only $\sim 2\sigma$ significant, consistently $< 3\sigma$. At high and low redshifts, the simpler model was preferred.

tendency towards a shallow slope of galactic age over time indicates the populations considered here in each mass bin are continually joined by younger galaxies over time (and/or left by higher mass galaxies). Though using a hard cut in sSFR as a truth label for the RS has complications (as discussed in §4.5.1), we find high selection accuracy, especially at higher redshifts, as compared to typical two-color selection of the quiescent population.

Despite being only a rough first pass, our imputation of rest-frame spectral slopes (see §4.5.2) revealed significant spectral differences between RS and BC on average; in addition to the 4kÅ break, we find a notable feature near 6300 Å in the BC. Future analyses could use more detailed methods to align slopes between bands, accounting for differences between bandpass transmissions. Accounting for these asymmetries and deformities in bandpass shape (as compared to flawless box functions) would allow for more accurate spectrum reconstruction, leading to an improved characterization of populations.

In our models we find no significant evidence for using a ≥ three-component model over a two-component model of galaxy populations (see §4.5.3). However, in the low-z SDSS dataset used in B22, we found significant evidence towards using three components instead of only two. The third component consistently had lower weight ($w < 10\%$), higher scatter (roughly twice that of the BC), and usually had the lowest (small positive or consistent with null) inter-color correlations. This indicates that the third component merely captures 'noise', i.e. galaxies that didn't fit well in either component (like the green points of Figure 4.9). Whether such galaxies are excluded due to low likelihoods or chosen as a third component of the mixture model, the two core populations of RS and BC remain dominant, with insufficient non-Gaussianities to warrant their sub-division in our samples.

Finally, we turn to the future of the Red Dragon algorithm. In its current state, it is designed to run GMM parameters $\boldsymbol{\theta}$ with a single variable smoothly—in this paper, we run with redshift. Any other fields must be binned—in this paper, the three mass decades. An improved version of the algorithm could allow $\boldsymbol{\theta}$ to evolve with N fields (e.g. redshift, stellar mass, and local density). Such a multidimensional parameterization of the RS and BC would yield insights e.g. into whether evolution of RS & BC mean colors with local density observed by Balogh et al. [2004] are purely due to the correlation of stellar mass with local density (i.e. mean colors are invariant to local density) or whether RS mean colors do indeed depend on local density. This could also give a global red fraction function $f_{\rm RS}(z, \mu_\star, \delta)$, measuring mean red fraction as a function of redshift, stellar mass, and local density.

As JWST and other telescopes increasingly yield quality high-redshift data of the quiescent population, Red Dragon will yield increasingly precise characterization of galaxy populations, leading to improved understanding of our universe.

Acknowledgements

WKB thanks Peter Melchior for providing pyGMMis, the backbone of Red Dragon, along with crucial insights in parameter fitting and interpretation. He also thanks Johnny Esteves for editing assistance and the stellar name of “Red Dragon”.

Funding from NASA Grant 80NSSC22K0476 provided crucial support for this research. The authors additionally thank the National Energy Research Scientific Computing Center (NERSC) for access to computing resources (including COSMOS data) used to carry out the analyses of this paper.

This work was made possible by the generous open-source software of NumPy [van der Walt et al., 2011], Matplotlib [Hunter, 2007], h5py [Collette et al., 2017], KLLR [Farahi et al., 2018, Anbajagane et al., 2020], and pyGMMis [Melchior and Goulding, 2018].

Data Availability

The Red Dragon algorithm is available at `bitbucket.org/wkblack/red-dragon-gamma`. Routines used for the analyses this paper are found in `bitbucket.org/wkblack/rd_des`.

The COSMOS dataset is available publicly, on the DES Data Management page.

CHAPTER 5

Practice Makes Better: Quantifying Grade Gains of Practice Study

Chapter Summary

Problem Roulette (PR), an online study service at the University of Michigan, offers points-free formative practice to students preparing for examinations in introductory STEM courses. Using four years of PR data involving millions of problem attempts by thousands of students, we quantify benefits of increased practice study volume in introductory physics. After conditioning mean final grade on standardized (ACT/SAT) math test score, we analyze deviations based on student study volume. We find a strong effect; mean course grade rises quadratically with the logarithm of the total number of PR questions encountered over the term ($N_{\mathrm{Q,tot}}$), with an overall gain of 0.77 ± 0.12 grade points between $1 < N_{\mathrm{Q,tot}} < 1000$. The gains are persistent across the range of math test score represented in our sample. While $N_{\mathrm{Q,tot}}$ surely correlates with other study habits, the benefits of increased study in general still hold. A model for final grade using test score and study volume largely accounts for demographic stratification, including by sex, parental education level, number of parents at home, nationality / underrepresented minority status, and regional income level, with two significant exceptions: students whose parents did not earn a college degree, who earn -0.27 ± 0.04 grade points (6.1σ) below expectations and underrepresented minority students at -0.14 ± 0.04 points (3.6σ). Residual scatter in final grade remains comparable to the maximal study gains, implying that the model is far from deterministic: individual variation trumps mean trends. Our findings can help motivate students to study more and help teachers to identify which types of students may especially need such encouragement.

5.1 Introduction

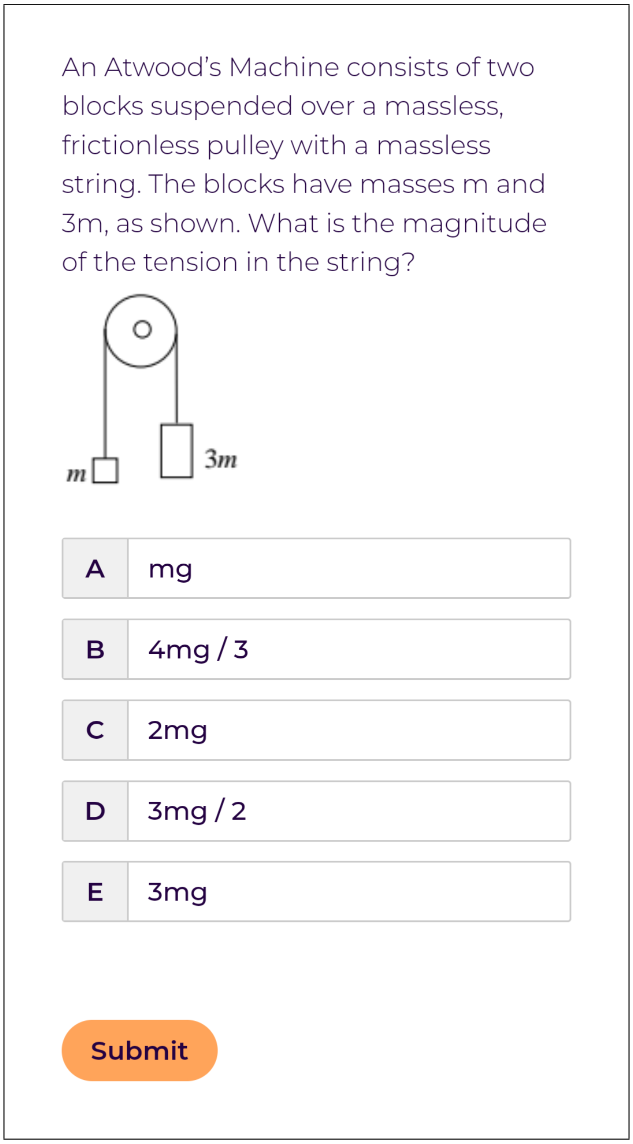
Physics is a notoriously difficult subject in the eyes of many undergraduates [Wong et al., 2022]. Introductory physics courses are typically among the most-failed courses on college campuses. According to student evaluations of teaching at our university[1] the workload of the first-semester physics course is perceived as considerably heavier than the workload of introductory courses in general chemistry or statistics. In a study of past examination problems for these three subjects [Weaverdyck et al., 2020], we found that introductory physics questions are both *more complex* (take a longer time to solve on average) and *harder* (have a lower average correct response rate) than questions in chemistry and statistics. Asked by students how to succeed in physics, instructors often recommend practicing more problems, among other strategies. We seek here to measure how such practice study affects final grade.

Benefits of practice study are well-documented, though some forms of study benefit students more than others. In 2013, Dunlosky et al. [2013] performed a meta-analysis on hundreds of studies on learning, measuring the utility of various study methods. Of the ten main study methods considered, they found that **practice testing** ("self-testing or taking practice tests over to-be-learned material") and **distributed practice** ("implementing a schedule of practice that spreads out study activities over time") were of the highest utility. Both methods benefited learners of different ages and abilities and were shown to boost students' performance across many different kinds of tasks and contexts. (In contrast, study methods such as summarizing, highlighting, and rereading were less effective.) A more recent meta-analysis [Yang et al., 2021] of several hundred studies finds significant gains from testing (a term they use to include quizzing), measuring boosts in student attainment that benefit all kinds of students in similar manner. Such findings help motivate use of a study tool at the University of Michigan known as Problem Roulette [Evrard et al., 2015].

Problem Roulette (PR) is an optional, points-free study service at the University of Michigan (UM) that provides students open access to a large library of locally-authored, topically-organized problems in multiple subject areas. Most of the content consists of multiple-choice questions used in past examinations in introductory science, technology, engineering, and mathematics (STEM) courses. "Roulette" refers to both its random selection of questions (an example is shown in Figure 5.1) as well as its reflection of high-risk assessments. Students can optionally activate settings which simulate timed tests, mirroring the difficulty and stress of actual exams; see Appendix I for more details on study modes and instructor options. PR's equality of accessibility to UM students sidesteps paywalls of online repositories and bypasses access limits to large banks of past exams held by exclusive student groups such as fraternities, sororities, and honor societies.

Our group previously examined study behaviors and grade benefits of practice based on PR

[1] Made available to campus members by the Atlas service, http:atlas.ai.umich.edu.



An Atwood's Machine consists of two blocks suspended over a massless, frictionless pulley with a massless string. The blocks have masses m and 3m, as shown. What is the magnitude of the tension in the string?

A	mg
B	4mg / 3
C	2mg
D	3mg / 2
E	3mg

Figure 5.1: Example Problem Roulette question, as viewed on a mobile device for PHYSICS 140.

usage data from 2013 to 2017 in three subject areas. Weaverdyck et al. [2020] measured study gains using a median split of high vs low study, finding in the high study group a moderate benefit of ~ 0.15 grade points in chemistry and statistics and a weaker benefit in physics. They also found that female students worked $\sim 25\%$ more problems on average than male students in all three subjects; they additionally received a slightly higher grade benefit than male students from higher volumes of PR practice.

To quantify these study benefits they used a measure called "grade anomaly", comparing each student's grade points earned (GPE) in a particular course relative to their end-of-term cumulative GPA computed from all *other* classes (termed GPAO) [Huberth et al., 2015, Koester et al., 2016]. "Grade anomaly" is then simply GPE − GPAO. While measures such as GPAO are commonly used to control for student performance, "grade anomaly" masks a large portion of grade benefit from practice study due to its correlation with study habits (see §5.3.3), so subtracting off GPAO from GPE subsumes a significant portion of otherwise measureable study gains. It is therefore not optimal for our current study, which seeks to quantify gains of study.

In this paper, after modeling GPE as a function of standardized math exam score (concordant ACT and SAT math subscores), we quantify the extent to which deviations from the mean trend correspond to increased PR study volume. We then consider five student demographic characteristics: sex, parental education level, single parent status, nationality and underrepresented minority (URM) status, and high school zip code median income (a proxy for estimating family income). Grade differences within each of these subgroups have been noted; linear regression and studies of linear correlations have shown that these differences in grades are ameliorated to varying degrees on accounting for test scores, personality, study habits, and other factors [Walpole, 2003, Aluja and Blanch, 2004, Pascarella et al., 2004, Delaney et al., 2011, Richardson et al., 2012, McLanahan et al., 2013, Matz et al., 2017, Simmons and Heckler, 2020] In contrast, in this paper we use the non-linear tool KLLR to investigate to what extent differences between these demographic sub-groups are explained or exacerbated by accounting for test scores and study volume.

We describe our methods and data in §5.2 before presenting our key findings in §5.3. That section begins with grade gains from practice conditioned on ACT/SAT math score (§5.3.1) followed by an exploration of demographic differences (§5.3.2) and a comparison to grade gains conditioned on GPAO (§5.3.3). We offer reflection in §5.4 and conclude with succinct recommendations for teachers and students in §5.5.

5.2 Methods

In this section, we describe the flexible population modeling method known as kernel-localized linear regression (KLLR), introduce the data and scope of our study, demographic groupings, and introduce the math test score measure (T) used to model mean GPE as a base condition.

5.2.1 Measuring trends with KLLR

While any differentiable function is linear at small enough scales, life tends to be non-linear. As one considers increasingly wider domains, behavior of any variable moves from a simple mean to a linear trend to a quadratic one and so forth (as is the nature of Taylor series). Our data in student math scores and study habits cover wide enough ranges that simple averages and even linear fits become statistically disfavored.

Furthermore, there are reasons to expect significant non-linearity in the trends we consider in this paper. One might expect grade points earned (GPE) as a function of math test score to have a sigmoid relation (if test scores ranged high enough), as there's a definite floor and ceiling to GPE (and a practical floor of grades given or accepted, as only some grades are passing). One might expect complex behavior in the space of grade benefit versus study volume in a given term, showing steady increases until a student begins to overstudy (eventually even foregoing sleeping and eating to increase study volume), followed by a decrease in performance. These expected non-linearities motivated us to use a statistical method beyond simple averages, linear trends, or even polynomial fitting—a method capable of measuring local, non-linear trends in a relatively agnostic fashion.

To examine mean relationships of grades and grade benefits as smooth functions of secondary variables, we employ Kernel-Localized Linear Regression (KLLR) [Farahi et al., 2018, 2022], a method that determines parameters of a locally linear fit (mean, slope, and variance) within a sliding Gaussian window. This approach to population modeling, originally developed in the 1970s [Cleveland, 1979, Cleveland and Loader, 1996, Takezawa, 2005], allows for more detailed analysis than polynomial fitting as it does not enforce a particular global behavior.

A single parameter is required for the method, the width of the Gaussian filter. As the Gaussian window slides across the domain, KLLR produces continuous and smooth fits to data. The kernel widths are chosen to be roughly a fifth of their range,

$$\sigma_{\mathrm{KLLR}} = \frac{1}{5}(q_{99\%} - q_{1\%}) \tag{5.1}$$

where $q_{n\%}$ is the nth quantile of the logged data. For our data, this kernel size minimizes noise in the fit while maximizing allowed internal variation, lying on the threshold of non-monotonic behavior.

5.2.2 Courses investigated

Problem Roulette currently supports sixteen courses with over ten thousand unique questions available in aggregate. Our primary focus here is on the introductory, calculus-based physics sequence for scientists and engineers: PHYSICS 140 (General Physics I: Mechanics) and its continuation, PHYSICS 240 (General Physics II: Electricity and Magnetism). These courses are primarily (~ 80%) taken by engineering students. [2]

While our focus is on study in physics, in this section only we offer some context by including basic PR usage and mean grade behavior in three other PR-supported STEM courses: General Chemistry: Macroscopic Investigations and Reaction Principles (CHEM 130), Elementary Programming Concepts (EECS 183), and Introduction to Statistics and Data Analysis (STATS 250).

Table 5.1: PR usage statistics, shown for selected PR-supported courses over a seven-semester period (Winter 2018 to Winter 2021). We focus on PHYSICS 140 and 240 here but include three other courses for context. Statistics reflect the ~ 93% of students with ACT or SAT scores available.

Course	$N_{\mathrm{Q,course}}$ [3]	N_{students} [4]	f_{use} [5]	$\sum N_{\mathrm{Q,tot}}$ [6]	$\sum N_{\mathrm{sess}}$ [7]
PHYSICS 140	854	3856	82.5%	348208	27471
PHYSICS 240	931	2772	60.8%	131868	10312
CHEM 130	685	5786	90.0%	1036986	59358
EECS 183	953	3803	71.3%	246239	13371
STATS 250	526	9435	74.9%	614700	38242

3. Number of unique questions available.
4. Number of students enrolled during the study period.
5. Fraction of students encountering at least one question on PR.
6. Total number of questions completed per term by all students.
7. Total number of sessions completed per term by all students.

Table 5.1 shows PR usage statistics for the seven-semester study period beginning Winter 2018 and ending Winter 2021 (inclusive). Spring and Summer terms during this period are also included, but these have much smaller enrollments than the traditional Fall/Winter terms. The student count includes only those with either an ACT or SAT scores recorded, representing ~ 93% of the full student enrollment.

Because it is an optional service that is not employed for summative course assessments, not all students use PR. The usage fractions are generally high, with values ranging from 60% in PHYSICS 240 to 90% in CHEM 130. Note that the student numbers are unique only within each class; many students enroll in several of these courses over the course of their careers. The intensity of study is highest in CHEM 130, with an average of 200 questions attempted per term by PR-using students.

[2] UM also offers a similar sequence designed for life sciences students as well as an honors sequence for physics majors. The former sequence was rebuilt during the study period and so is not yet supported by PR. The latter sequence has a much smaller enrollment which limits its statistical power.

By this metric, study volume in PHYSICS courses is considerably lower, with a mean value of 110 and 78 questions per term in 140 and 240, respectively. Median study volumes for physics courses were also smaller by more than a factor of two compared to CHEM 130.

The study was determined to be exempt from ongoing review by our Institutional Review Board (HUM00158291).

5.2.3 Study volume indicators

We use several indicators of study volume in this paper, each calculated on a per term basis and summarized in Table 5.2. The primary measure we use is the total number of unique questions encountered per session, summed over all sessions on Problem Roulette. This measure, $N_{\mathrm{Q,tot}}$, is given for each student individually over the full academic term. As shown in Figure 5.2, the distribution of $N_{\mathrm{Q,tot}}$ is close to log-normal (as are the distributions of N_{sess} and $N_{\mathrm{Q,mean}}$). For the combined sample of PR users in both physics courses, the log-mean value of $N_{\mathrm{Q,tot}}$ is $\exp\langle\ln N_{\mathrm{Q,tot}}\rangle \simeq 37.5$, with a factor of 4.3 standard deviation. This value includes skipped questions, which represent roughly 20% of the total of all questions encountered.

Table 5.2: Quantiles of study volume indicators, shown for PHYSICS classes described in the text. Values are the 50% (median), 75%, and 95% quantiles of each quantity for a sample of students with non-zero PR study (see Table 5.1 for usage fractions) and available ACT/SAT math scores (93% of total enrollment).

Indicator	PHYSICS 140			PHYSICS 240		
	50%	75%	95%	50%	75%	95%
$N_{\mathrm{Q,tot}}$	50	103	389	25	36	296
N_{sess}	5	9	28	3	4	21
$N_{\mathrm{Q,mean}}$	9.5	12.4	26	8.3	10	24

We also examine two additional measures of practice: i) the number of PR sessions, N_{sess}, where a new session is triggered when a student logs on and responds to at least one question, either for the first time or after a period of inactivity since responding to the last question; and ii) the mean number of questions encountered per session, $N_{\mathrm{Q,mean}}$.

Over the study period, different approaches to midterm and final examinations were taken by instructors. In some terms three evening exams, spaced by roughly one month, were held, each with typically 20 questions. In other terms, biweekly 10-question "quizzes" were held. In some semesters these were held during class meetings while in others (particularly during the COVID-19 shutdown period) they were held in the evening. Final examinations were cumulative and typically consisted of 25 questions. All questions were multiple-choice (typically five possible responses), authored by the instructors for each exam. While some degree of similarity between exam questions

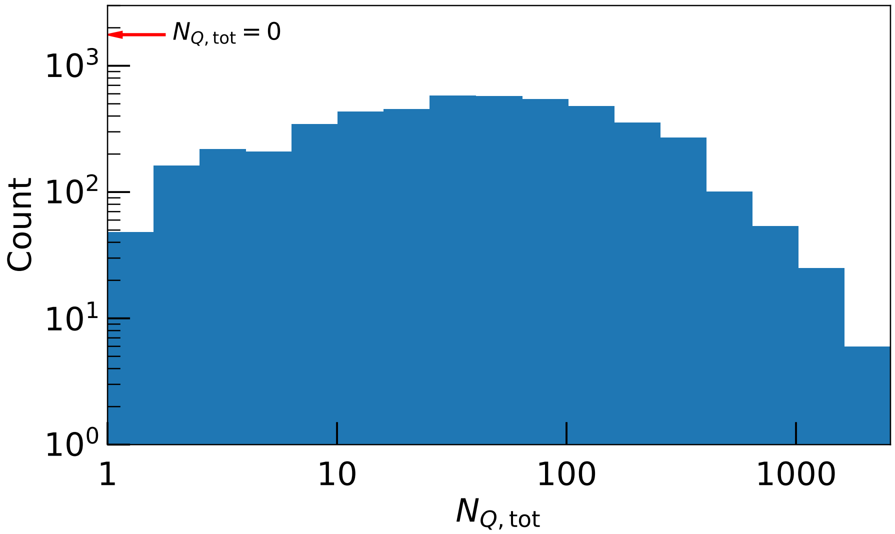


Figure 5.2: Histogram of total unique questions encountered per session, $N_{\mathrm{Q,tot}}$, for Physics 140 and 240 combined. Red arrow indicates the number of students who didn't use PR.

and PR questions is likely, direct re-use of PR questions is strongly disfavored by the departmental culture of testing students with new questions.[8] All examinations were proctored and timed, and across the study period student scores on these assessments contributed half or slightly more to their final letter grade. Overall weighted student scores map to letter grades in a manner defined at the beginning of the semester; there is no "curve" in the traditional sense of using ranked ordering to define grade boundaries.

With that context, the N_{sess} values in Table 5.2 reflect the fact that PR-engaged students typically use the service once per examination, but with quite a large spread. Combining both courses, we find log-mean number of sessions $\exp\langle \ln N_{\mathrm{sess}}\rangle \simeq 4.4$, with nearly a factor of three dispersion. Five percent of students had 25 or more sessions over the term, a frequency of roughly two sessions per week. Sessions typically average about 30 minutes in duration, during which students attempt roughly eight problems. Again, the dispersion is large (factor ~ 3), and five percent of students average 25 problems per session.

Our analysis using PR-based statistics is an incomplete proxy for total study effort by students. Students do use other methods of study, including doing homework problems from their online textbook service, using tutors, or by going to on-campus sites such as the Physics Help Room and Science Learning Center. While inherently only a partial picture of student study effort, the volume of PR activity is substantial enough to offer useful insight, as we show below (see Fig. 5.4).

[8] Further evidence against PR practice as "teaching to the test" comes from exam scores, which average $\sim 65\%$ with a dispersion of 15%.

Table 5.3: Demographic categories and sub-groups considered in this analysis.

Category	Sub-group	Description
Sex	Male/Female	Sex as catalogued in LARC
Parental education	≤HS/A,B/M,D	Highest degree earned: high school degree or less; associate's or bachelor's; master's or doctorate
Parents at home	Both/Single	Whether the student comes from a single-parent household
N/URM[9]	URM/NUR/ITL	ITL if student is international; else URM=underrepresented minority, NUR=non-URM
Income bin	1st/2nd/3rd/4th	Median income of high school ZIP in near-quartile groups, divided by values of \${50, 75, 100}k

Table 5.4: Fraction of course participants in each demographic subgroup for each physics course.

Course	Male/Female	≤HS / A,B / M,D	Single/Both	URM/NUR/ITL	1st / 2nd / 3rd / 4th
140	64%/36%	10%/29%/60%	15%/85%	15%/80%/5%	22%/31%/18%/22%
240	74%/26%	9%/27%/62%	13%/87%	12%/83%/5%	22%/32%/19%/21%

Furthermore, because these study indicators correlate with study habits in general, they may even be seen as proxies for studying of any kind, pointing towards effects beyond PR activity alone.

5.2.4 Demographic groups

We examine several demographic distinctions, delineated in Table 5.3. Table 5.4 gives the fractions for each group among the ACT/SAT student populations in the two physics courses of interest.

The University of Michigan's definition of underrepresented racial/ethnic minority (URM) is tied to nationality and requires some exposition. As defined in the Learning Analytics Data Architecture [LARC Lonn and Koester, 2019] Data Dictionary, students are considered international (ITL) if they are neither U.S. citizens nor U.S. permanent residents. The international student population is roughly two-thirds Asian, with the plurality of students from China (46%) and another 5% from the Republic of Korea. The rest of international students come to UM from over 100 countries, each individually accounting for < 4% of the remainder. Of the remaining domestic student population, students are considered underrepresented if they self-identify as i) Black or African American, ii) Hispanic, iii) Native American, or iv) Native Hawaiian or Other Pacific Islander. Otherwise, they are considered non-underrepresented minorities (NUR). In the terms considered here, the URM population was chiefly composed of Hispanic students (42%), students identifying as more than one ethnicity or race (30%), and Black students (27%).

5.2.5 ACT/SAT math score as a control condition

The standard deviation of grades achieved in these courses is large, roughly 0.8 on the standard 4.0 scale.[10] A particular student's grade is influenced by a host of factors; we do not seek to (nor could we) capture all these factors simultaneously. Instead, we opt to employ a proxy for mathematical proclivity using a student's score on the ACT or SAT mathematics test. Recent work [Simmons and Heckler, 2020] on a large sample of introductory physics students identifies ACT math score as a strong predictor of final grade. Of the 7417 students enrolled in this introductory physics sequence during the study period, only 498 have neither ACT nor SAT scores recorded, meaning that 93% of the total student sample have pre-college math test scores available.

In our analysis, we map ACT to SAT test scores using a 2018 concordance table from Compass Prep [Compass Education Group, 2018]. Though fractions of students with SAT and ACT scores is similar, between PHYSICS 140 and 240, 3% more students have SAT than ACT scores. We thus minimize imputation by converting scores from ACT to SAT. (Imputing in the opposite direction has negligible effects on results.) In cases where both scores were present, their average was taken. In cases where the same test was submitted multiple times, the highest score was used. Noting that the SAT minimum component score is 200 and maximum 800, we map SAT math scores to the unit interval by defining

$$T_i = \frac{\mathrm{SAT}_i - 200}{800 - 200}, \tag{5.2}$$

where SAT_i is the actual, imputed, or averaged SAT math score for the i^{th} student.

The distribution of T leans heavily toward high values in our datasets, especially so in the physics courses. The two physics courses considered in our study have similar distributions of T, with medians near 0.92, 10% & 90% quantile values near 0.80 & 1.0, respectively, and ~ 0.44 as their minimum value. A recent study on grade inflation [Evrard et al., 2021] at our university shows regular growth in ACT and SAT scores; rescaling SAT and ACT scores into the unit interval to match T, we see annual gains of 0.006 in scores for both tests. While this increase reflects a heightened selectivity of the institution, it is an order of magnitude smaller than the scatter in T and therefore has relatively little effect over our study period.

While differences in mean scores have been noted for several demographic sub-groups historically as well as in our sample (see Figure 5.5), we find that intrinsic scatter within each sub-group far exceeds the distance between group mean scores. There is no definitive consensus among researchers as to the *cause* of these differences in mean scores,[Jensen, 1980, Drasgow, 1987, Herrnstein and Murray, 1994, Jensen, 1998, Brown et al., 1999, Frey and Detterman, 2004, Koenig

[10] Our university uses a standard A–E/F letter grade scale that maps to a numerical point scale of 4.0 (A) to 0 (E/F); B=3.0, C=2.0, D=1.0. UM also supports ± gradations that count for 0.3 grade point deviations, so e.g. a B+ is 3.3 grade points and a B- is 2.7 grade points.

et al., 2008, Coyle and Pillow, 2008, Dorans, 2010, Soares et al., 2015, Geiser, 2020] particularly whether this is a bias in measurement (i.e., whether two individuals with equal capacity but different demographic membership have significantly different scores) or a purely relational bias (i.e., the tests properly measure capacity, but differences in mean group scores are caused by some secondary variable). A commensurate summary of the conflicting interpretations of demographic differences in scores is beyond the scope of this paper. Because of the significant predictive power of ACT and SAT tests in estimating college grades, we find utility in using the measure T as a baseline, and relegate investigation of its interpretation to future research.

5.2.6 Modelling mean GPE versus test scores

In this section, we describe how we condition grades on math test scores. To condition one variable on another is an attempt to remove its effects, in order to discern further trends. Here we subtract out mean grade points earned (GPE) at a given math test score T in order to mitigate the strongly confounding factor of differences in math skills on expected grade. For example, if two students of vastly differing math skills both don't study, they are unlikely to earn identical grades. By conditioning on T, we subtract out the mean expected grade, allowing us then to measure how divergence from those expectations corresponds with study habits.

We use KLLR to fit mean grade μ_{GPE} as a function of T, allowing us to detect any potential non-linearities in trends. Using all student T values, equation (5.1) yields $\sigma_{\mathrm{KLLR}} = 0.09$, which value we use for all course KLLR fits, shown in Figure 5.3. Though all five courses considered here are STEM courses, we find stark differences in the mean relationships between physics (concave up) versus non-physics courses (concave down). While students with higher math scores tend to score higher in all their courses, even students with above-average math scores of $T = 0.75$ tend to score lower grades in physics relative to the other STEM courses considered here.

For each course, we find that mean grade point earned as a function of T is well-fit for the majority of students by a quadratic:

$$\mu_{\mathrm{GPE}}(T) = a_0 + a_1(T - T_0) + a_2(T - T_0)^2 \tag{5.3}$$

(using pivot $T_0 = 0.9$, the physics median T value). Parameter fits for each course are listed in Table 5.5.

We note that when we measured fit parameters using only those students who did not use PR for study, we found good consistency in slopes and curvatures (hinting that study trends don't correlate strongly with T; see lower panel of Figure 5.5 and Figure 5.7), but intercepts a_0 were lower by 0.10 ± 0.04 grade points. This is not surprising when one considers that, overall, those who didn't study on PR in PHYSICS 140 had lower grades by 0.14 ± 0.04 points and in PHYSICS 240 by

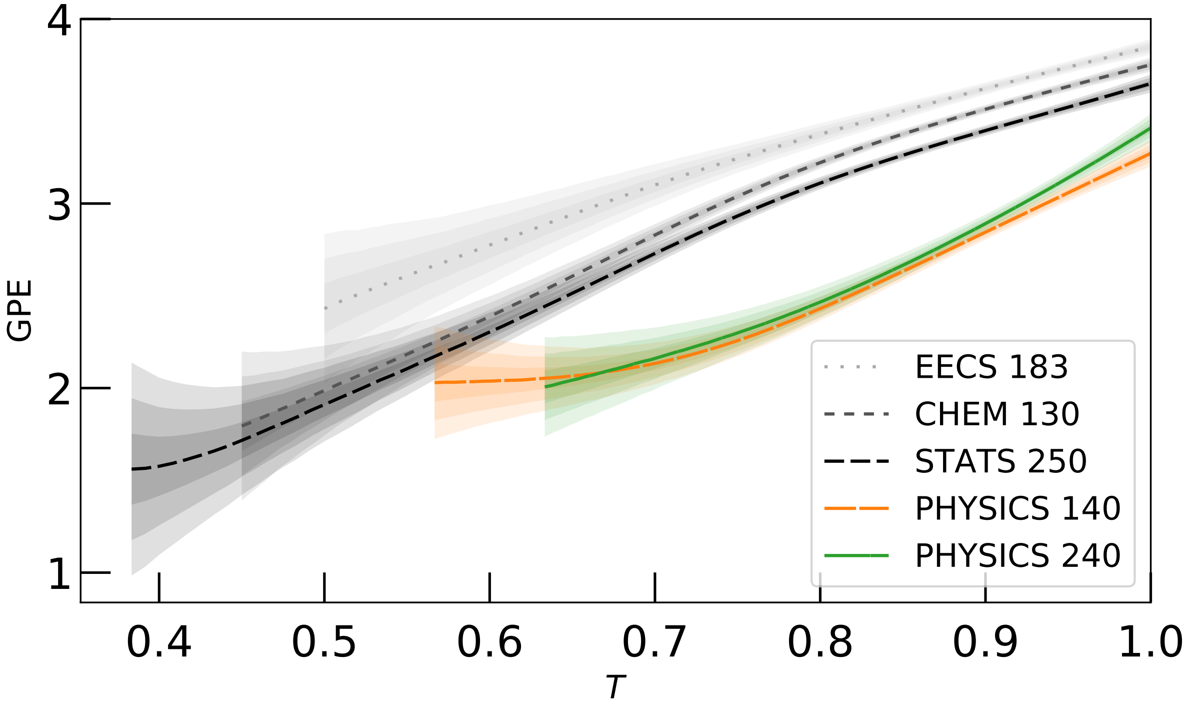


Figure 5.3: KLLR fits to mean grade points earned (GPE) as a function of normalized ACT/SAT math score (T), shown for multiple STEM courses (see legend). Shaded regions show $\pm 3\sigma$ uncertainties on the fit. The left-side T limits displayed are set by a limit of ten students.

0.10 ± 0.03 points. This vertical shift is precisely what this study seeks to quantify; by subtracting out $\mu_{\mathrm{GPE}}(T)$, one can then measure these divergences from expectations. These significant shifts hint already at the benefits of study.

For the two physics courses, we found no significant differences in slope nor curvature for any demographic distinction. Though we found several significant differences in vertical shift between sub-groups, as explained above, these shifts do not detract from our analysis, as they are precisely what we set out to investigate by subtracting out a $\mu_{\mathrm{GPE}}(T)$ baseline. See §5.3.2 for more details.

As is visually apparent from Figure 5.3, student grades in physics follow a different form compared to the other subject areas represented there. Compared to the non-physics courses, the two physics courses have intercepts at $T_0 = 0.9$ lower by ~ 0.7 grade points, the local slopes are nearly twice as large, and they have positive curvatures ($a_2 \sim +7$) whereas the other courses display negative curvatures ($a_2 \sim -3$). In essence, even students with the strong math abilities encounter a larger grade penalty in physics compared to these other subjects.

Table 5.5: Fit parameters of mean grade versus T. Fitted with quadratic functions $\mu_{\mathrm{GPE}}(T)$ (equation (5.3)) for each course, giving the mean trend of GPE (course grade points earned) as a function of T (concordant ACT / SAT math subscores), determined using all students with ACT or SAT scores recorded. Given for each class individually as well as the merged set of PHYSICS 140 + 240.

Course	a_0	a_1	a_2
PHYSICS 140	2.82 ± 0.02	4.3 ± 0.2	+5.3 ± 1.2
PHYSICS 240	2.82 ± 0.02	5.0 ± 0.2	+8.9 ± 1.7
MERGED	2.82 ± 0.01	4.6 ± 0.1	+6.7 ± 1.0
CHEM 130	3.53 ± 0.01	2.6 ± 0.1	−3.8 ± 0.5
EECS 183	3.63 ± 0.01	2.3 ± 0.2	−1.8 ± 0.7
STATS 250	3.39 ± 0.01	2.8 ± 0.1	−3.4 ± 0.5

5.3 Results

Due to their similar structure, we analyze the combined sample of PHYSICS 140 and 240 student behavior here, beginning with deviations from mean expected GPE at a given math test score T as a function of study volume. Because of the wide dynamic range and log-normal distribution of study volume (Figure 5.2), we employ $\log_{10}(N_{\mathrm{Q,tot}})$ as the independent variable. The use of a logarithmic measure of study enables a natural interpretation in terms of *multiplicative factors* of student effort. We then investigate differences in study volume and GPE among demographic sub-groups. The section concludes with remarks on using GPAO instead of T as a baseline condition for grade prediction.

5.3.1 Evidence of grade gains from practice study

With the fit to mean GPE as a function of math test score T in hand (see Figure 5.3), we are now poised to address the question: “For a given student, how do PR study habits relate to their final course grade?” We focus on the shift in grade earned by a given student from expectations at a given T:

$$\Delta_{\mathrm{GPE},i} \equiv \mathrm{GPE}_i - \mu_{\mathrm{GPE}}(T_i). \tag{5.4}$$

We look to measure how these mean shifts in student grade depend on some measure of study volume X, meaning we seek $\langle \Delta_{\mathrm{GPE}} \rangle (X)$.

Using KLLR, we measure mean final grade point deviation, equation (5.4), as a function of our primary study volume indicator, $X = \log_{10}(N_{\mathrm{Q,tot}})$. KLLR kernel width, as calculated from equation (5.1) is $\sigma_{\mathrm{KLLR}} = 0.56$ (a factor of 3.6 spread). We find a significant increase in mean grade points earned shown in the top panel of Figure 5.4. The mean behavior monotonically grows

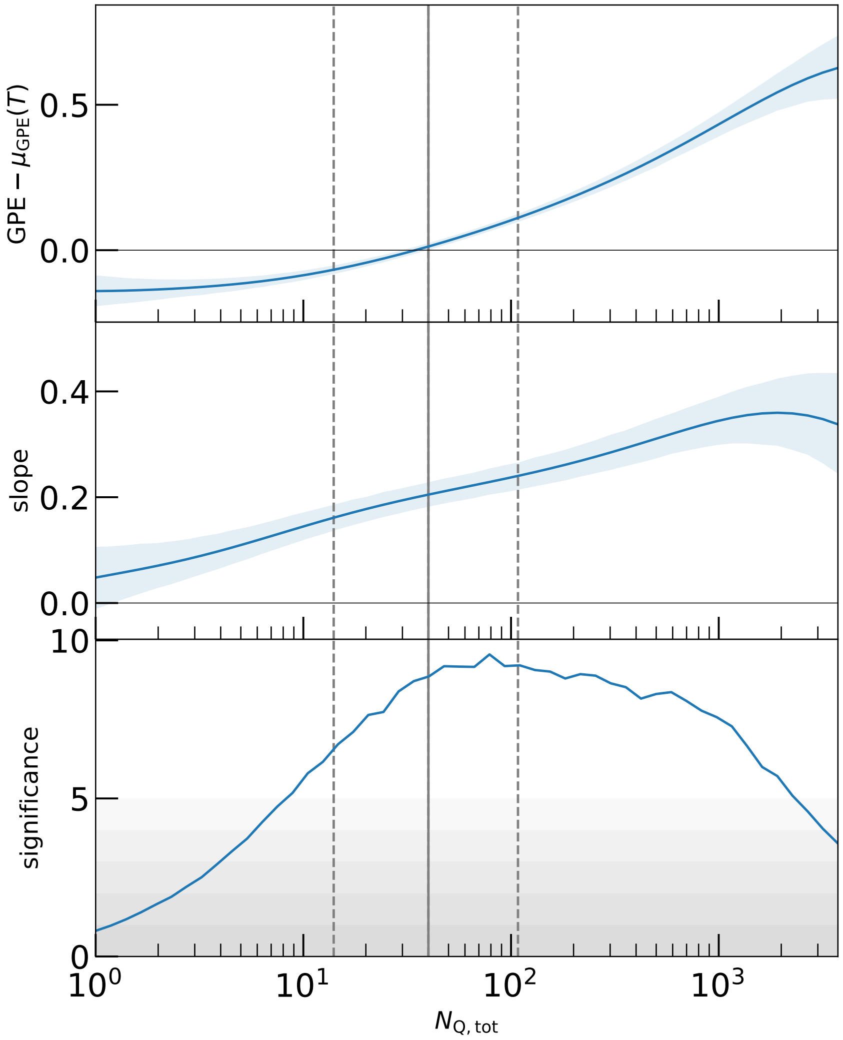


Figure 5.4: Grade gains as a function of practice effort, shown for the combined sample of PHYSICS 140 and 240 students. Study volume is measured by the logarithm of the total number of PR questions encountered over the term, $N_{\mathrm{Q,tot}}$. Vertical lines indicate quartile values of $N_{\mathrm{Q,tot}}$ for students who use PR; half practice between 14 and 108 problems, with a median of 40. Statistics are computed with KLLR using a kernel width of 0.56 in $\log_{10} N_{\mathrm{Q,tot}}$ (a factor of 3.6) and shaded regions are 1σ uncertainties from bootstrap resampling. **Upper panel:** Mean shift in grade points earned, $\langle \Delta_{\mathrm{GPE}} \rangle$, relative to that expected from pre-college math score, equation (5.4). **Middle:** Local slope of $\langle \Delta_{\mathrm{GPE}} \rangle$. **Lower:** Significance level of a non-zero slope. Note that students who didn't engage with PR are not included in this analysis.

from a value of −0.2 (a grade point penalty relative to the T-determined expectation) to +0.6 (a relative grade point benefit) as the total number of questions attempted over the term increases from one to over one thousand. The middle panel shows the local slope of the grade gain, which increases nearly linearly in $\log(N_{\mathrm{Q,tot}})$, and the lower panel shows the significance of that slope being non-zero.

The top panel shows the basic result that students who study more tend, on average, to do better in the course. The novelty of this unsurprising finding lies in both the *precise quantification* of mean grade gain as a function of term-aggregated study volume and its *large overall amplitude*. These findings provide evidence-based answers to questions such as “How much should I study to get a half grade point boost in my final grade?” (on average: approximately 660 questions over the term) or “If I study one question each school day, how much better am I likely do in the course than without study?” (on average: roughly 0.24 grade points improvement). These outcomes hold true *on average*; as discussed below, significant residual scatter in grade remains, reflecting a range of extraneous factors that influence student learning of physics.

The KLLR characterization of study gains are largely in agreement with a simple quadratic trend, as seen in the near-linearity of the slope (middle panel). Fitting the trend in Figure 5.4 around the median value of $N_{\mathrm{Q,tot}} = 40$ we find mean behavior

$$\langle \Delta_{\mathrm{GPE}} \rangle = (0.00 \pm 0.01) + (0.22 \pm 0.02)\, \nu + (0.08 \pm 0.02)\, \nu^2, \tag{5.5}$$

where $\nu \equiv \log_{10}\left(N_{\mathrm{Q,tot}}/40\right)$. *Grade gains from practice are thus quadratic in the log of number of questions attempted.* While doubling study from one question a day to two a day will benefit a student, doubling from two a day to four will, on average, yield even more incremental benefit to their final grade.

The lower panel of Figure 5.4 shows that the slope is $> 3\sigma$ significantly positive for $N_{\mathrm{Q,tot}} \geq 4$. The slope is roughly 0.1 at this point, but four problems over the course of a full term is a very minimal amount of practice, and grade earned at this level is indeed lower than the T-conditioned mean by 0.2 grade points. Gains in the top panel become more evident above ten questions, and the slope continues to grow, reaching 0.3 at $N_{\mathrm{Q,tot}} \simeq 300$. This study volume represents roughly 25 questions per week, or five questions per school day. The positive slope in grade gain continues up to the limit of our study volume data. In our sample, 34 students, or 0.5% of the population, attempted over 1000 questions within a semester, a volume that slightly oversamples the number of questions available and that corresponds to a rate of roughly 15 questions per school day.

Table 5.6 quantifies for each study volume indicator ($N_{\mathrm{Q,tot}}$, N_{sess}, and $N_{\mathrm{Q,mean}}$) the overall grade gain, $\max(\Delta_{\mathrm{GPE}})$, defined as the maximum difference in KLLR mean values across the study volume domain. Uncertainties in the difference are propagated from the respective KLLR bootstrap errors on the minimum and maximum values. The table also quantifies the fractional reduction in variance $\Delta\mathrm{Var}$ from the initial scatter in GPE σ_0 to the scatter after accounting for $\mu_{\mathrm{GPE}}(T)$

Table 5.6: Maximum study gains, max(ΔGPE), between high and low PR study volume (see text) for each study measure along with the fractional reduction in grade variance after accounting for both T and study volume. The latter is calculated as $\Delta\text{Var}_{0\to i} \equiv (\sigma_i{}^2 - \sigma_0{}^2)/\sigma_0{}^2$, where σ_0 is the scatter in GPE originally and σ_i is the scatter in grade point after removing trends in $\mu_{\text{GPE}}(T)$ ($i = 1$) and both that as well as $\langle\Delta_{\text{GPE}}\rangle(X)$ for study measure X ($i = 2$), e.g. $X = \log_{10} N_{\text{Q,tot}}$.

Indicator	$\max(\Delta_{\text{GPE}})$	$\Delta\text{Var}_{0\to 1}$	$\Delta\text{Var}_{0\to 2}$
$N_{\text{Q,tot}}$	$+0.77 \pm 0.12$	-0.16 ± 0.03	-0.19 ± 0.03
N_{sess}	$+0.61 \pm 0.19$	-0.16 ± 0.03	-0.18 ± 0.03
$N_{\text{Q,mean}}$	$+0.72 \pm 0.11$	-0.16 ± 0.03	-0.19 ± 0.03

($\Delta\text{Var}_{0\to 1}$) and after additionally accounting for $\langle\Delta_{\text{GPE}}\rangle(X)$ in study measure X ($\Delta\text{Var}_{0\to 2}$). The reduction in variance from the middle to end stage is then $\Delta\text{Var}_{1\to 2}$.

For each study indicator, students with high study volume tend to do significantly ($> 3\sigma$) better than those with low study volume. Across the three study indicators, we see similar gains (within 1σ of each other) between two-thirds and three-fourths of a full grade point. This result is consistent with moving from a B- to a B+ or from a B+ to an A. These results are also robust with varying kernel size: doubling or halving kernel width yields statistically identical results. Furthermore, the results are robust with varying student math ability. Dividing the student sample into terciles by T, we find similar overall grade gains: 0.60 ± 0.21, 0.73 ± 0.19, and 0.75 ± 0.12 for the lowest, intermediate, and highest T terciles, respectively, averaged across study indicators. *All* students benefit similarly from increased study volume.

Across each study indicator, we see similar reductions of variance, with $\sim 20\%$ of the scatter in student grades explained by the joint $\mu_{\text{GPE}}(T)$ and $\langle\Delta_{\text{GPE}}\rangle(X)$ trends. The trend of mean grades running with T accounts for the majority of this reduction, causing on its own a $16\% \pm 3\%$ reduction. The range of $\mu_{\text{GPE}}(T)$ is roughly 1.2 grade points, so the running of $\langle\Delta_{\text{GPE}}\rangle(X)$ with maximal running of roughly half the range has less capacity to reduce variance. Because the distribution of $N_{\text{Q,tot}}$ clusters strongly towards lower values, further limiting its ability to reduce variance. However, on sampling evenly in $\log N_{\text{Q,tot}}$, we find a significant variance reduction of $\Delta\text{Var}_{1\to 2} = (9 \pm 3)\%$, with study explaining roughly 10% of the remaining scatter, after accounting for $\mu_{\text{GPE}}(T)$ expectations. This suggests that our initial inability to measure significant reduction of variance on accounting for study trends is tied to the clustering of $N_{\text{Q,tot}}$ at relatively modest values.

Below, we incorporate both math test score and study volume as fit here into a model of student grade and investigate divergences from model expectations for several demographic subpopulations.

5.3.2 Demographic differences

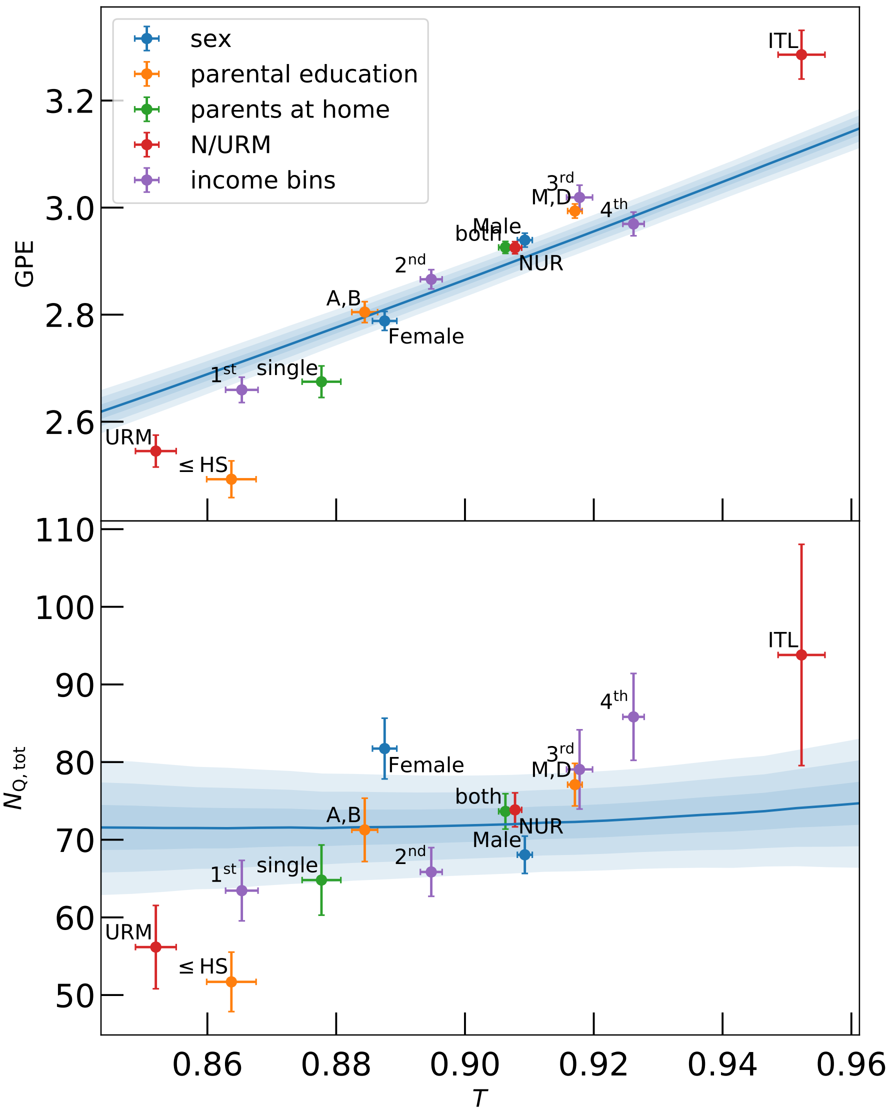


Figure 5.5: Mean GPE, $N_{\mathrm{Q,tot}}$, and T for the various demographic subpopulations indicated in the legend, with error bars showing standard deviation of the mean. KLLR fits of GPE and $N_{\mathrm{Q,tot}}$ as functions of T for the full student population, shown in blue with ±1, 2 and 3σ uncertainties, demonstrate starkly different trends. For the student population as a whole, grades strongly correlate with T while study volume does not.

On average, differing educational experiences exist across socioeconomic status, sex, race, ethnicity, home environment, and other factors [Pascarella et al., 2004, Delaney et al., 2011, McLanahan et al., 2013, Stroub and Richards, 2013, Owens et al., 2016]. As detailed in §5.2.4, we consider several of these demographic indicators (see Tables 5.3 & 5.4) to the extent and precision available to us; most indicators are self-reported. We investigate the degree to which demographic groups are stratified in the space of T, GPE, and $N_{\text{Q,tot}}$, then quantify how each sub-group's mean grade deviates from expectations.

As mentioned in §5.2.6, we first investigated whether $\mu_{\text{GPE}}(T)$ had any significant differences in form between demographic sub-groups. While there were significant differences in vertical intercept between sub-groups, there were no significant differences in slope nor curvature in the quadratic fitting of the courses. Compared to the global parameterization of Table 5.5, sub-population fits for curvature were consistent to $< 2\sigma$ and slopes were consistent to $< 3\sigma$. In contrast, there were several significant differences in offset a_0 for three groups, with the ≤HS group falling ~ 0.3 points below expectations while URM students and students of single-parent households both fell ~ 0.1 points below expectations. Once more, we emphasize that these shifts are precisely what this study seeks to quantify, asking what correlates with vertical shifts from mean expected grade at a given T. Now that we have shown that the fitting of $\mu_{\text{GPE}}(T)$ is robust to demographic sub-group, we turn our attention to how demographic sub-group deviation from expectations relate to study volume.

Figure 5.5 illustrates the relationship between mean measures for demographic groups, shown as points, and the overall trend for all students, shown as the line (with shaded regions indicating $\pm 3\sigma$ uncertainty on the KLLR fit). For this analysis, we measure mean $N_{\text{Q,tot}}$ values for all students for whom we have T scores, including those for whom $N_{\text{Q,tot}} = 0$. Note that the mean value, $\langle N_{\text{Q,tot}} \rangle \simeq 72$, differs from the median log value of 40 shown in Fig. 5.4 because the distribution of $N_{\text{Q,tot}}$ is close to log-normal and has substantial width. This level of effort corresponds to attempting roughly five questions per week during the term.

GPE differs starkly from $N_{\text{Q,tot}}$ in its trend with T. Mean grade earned (top panel) correlates strongly with math test score (0.38 ± 0.01), rising by half a grade point over just a 0.1 increase in T. In stark contrast, the volume of study by students (lower panel) is effectively flat (insignificantly correlated: 0.02 ± 0.01); low T students study very nearly as much as their high T counterparts. We return to this issue when considering GPAO as an alternate to T below.

Mean values of math test score T for demographic groups cover the 0.1 domain shown, with students whose parents have the lowest education level (≤HS) and underrepresented minority (URM) students possessing the lowest scores and international students (ITL) the highest. These groups deviate somewhat from the overall trend in grade, with ITL students lying above and URM & ≤HS below the population mean. Note that, because of the positive curvature in mean grade with T, mean values of sub-populations will tend to lie somewhat above the trend line.

While students across the spectrum of math test scores put in similar level of practice effort, some differences between demographic groups are apparent in the lower panel of Figure 5.5. Though deviations are generally less significant than in the upper panel, URM & ≤HS populations tend to rest below the trend whereas females, high income groups, and international students rest somewhat above.

Does Math Score and Study Volume Explain Demographic Grade Shifts?

The trends in the bottom panel of Figure 5.5 suggest an explanation for demographic grade deviations; vis. that the demographic shifts in study volume often have similar sign to those of Δ_{GPE}. However, these shifts sometimes misalign; for example, the highest income group tends to study more than average yet achieves only average grades. To address the degree to which study explains demographic deviations from mean trends, we employ a model that combines the expected mean grade conditioned on T combined with the grade shift as a function of study volume using the full student population (KLLR fits of Figures 5.3 & 5.4, largely equivalent to the parameterized fits of equations (5.3) & (5.5)).

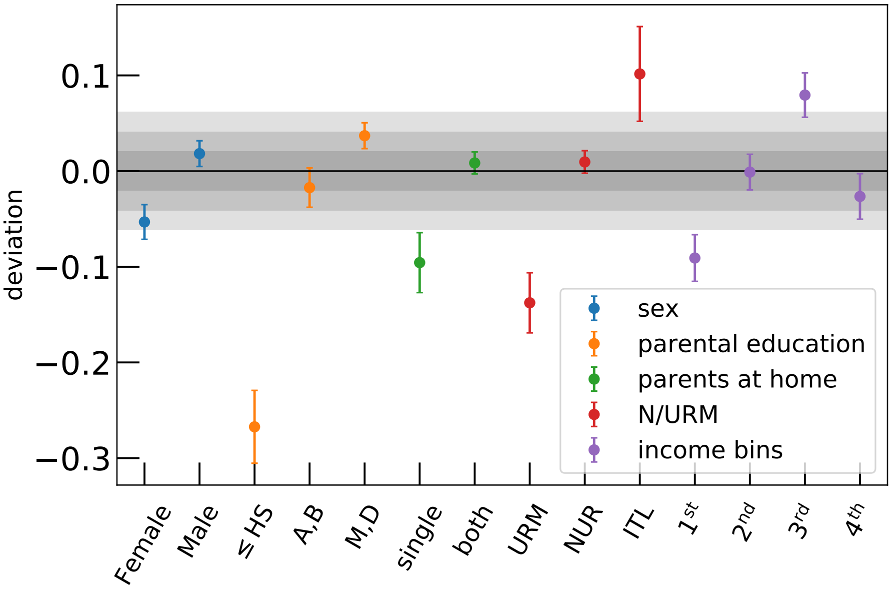


Figure 5.6: Mean GPE deviations of demographic groups from the overall trend, based on combining math score T and study volume $N_{\mathrm{Q,tot}}$ for the full student population. The shaded region around the null line shows $\pm 3\sigma$ uncertainties of the predictions based on combining T and $N_{\mathrm{Q,tot}}$. The error bar for each demographic group is the $\pm 1\sigma$ error in that group's mean GPE deviation from the mean model prediction.

Figure 5.6 shows the mean deviation from these expectations for each demographic subpopulation. The zero line and associated median $\pm 3\sigma$ uncertainties (median combined uncertainties of the KLLR fits to $\mu_{\mathrm{GPE}}(T)$ and $\langle \Delta_{\mathrm{GPE}} \rangle (\log_{10} N_{\mathrm{Q,tot}})$ for a median student) represent the expected value based on our model of the full student population. Most demographic sub-populations have mean deviations consistent with the model's expectations, but significant outliers are seen toward lower math scores. Students whose parents had a high school degree or less lie significantly below expectations, at -0.27 ± 0.04 grade points, as do URM students, at -0.14 ± 0.04 grade points. In addition, students from households with only a single parent present as well as students from high schools in the lowest income regions display a smaller, less significant deficit of -0.09 ± 0.04 grade points. Though the third income bin is slightly above expectations ($+0.08 \pm 0.03$), the fourth income bin is slightly below. All other sub-populations were $< 2\sigma$ deviant from expectations.

5.3.3 Comparison to GPAO

In this section, we investigate results when using GPAO as a baseline for comparison, rather than T. First, we explain some limitations of GPAO (with more issues outlined in Appendix J), Second, for PHYSICS 140 and 240 we show the $\mu_{\text{GPE}}(\text{GPAO})$ plot analogous to Figure 5.3. Third, we show for the two courses combined a gains of study table, analogous to Table 5.6.

5.3.3.1 GPAO correlation with study

A core issue with using GPAO as a baseline for comparison (as is done, for example, with the "grade anomaly" GPE − GPAO of Weaverdyck et al. [2020]) is its correlation with study habits. Averaged over log or non-logged versions of all three study volume indicators used in this study (a total of six possible metrics), correlations between study and GPAO are statistically significant, 0.132 ± 0.013, whereas correlations between study and T are insignificant, only 0.013 ± 0.013. Regardless of study metric, GPAO consistently shows stronger and more significant correlations to study than T does. Because we seek to measure how study volume influences final grade, this coupling makes GPAO relatively untenable compared to T for this purpose.

Figure 5.7 shows the interplay between PR study volume, GPAO, and composite test score T for PHYSICS 140 and 240 combined. KLLR fits to study volume as a function of GPAO show clear rising behavior for each individual T bin, while the means (displayed as points with error bars) show no significant trend in study behavior with T. The striation of trend lines shows that among students with the same GPAO, those with lower T scores tended to study more on PR. In contrast to the lack of correlation between T and study, at *fixed* GPAO there is a significant trend of decreasing PR study volume for students with higher math scores, T.

Though students with high T scores tend to have higher grades and though students who studied more tended to have higher grades, there was no significant correlation between T and study volume. This seeming inconsistency is resolved in the anticorrelation between study volume and T *at fixed GPAO*. If we take PR study volume as a proxy for typical student study habits, then this could perhaps be interpreted to reveal that students with higher T scores tended to not need to study as

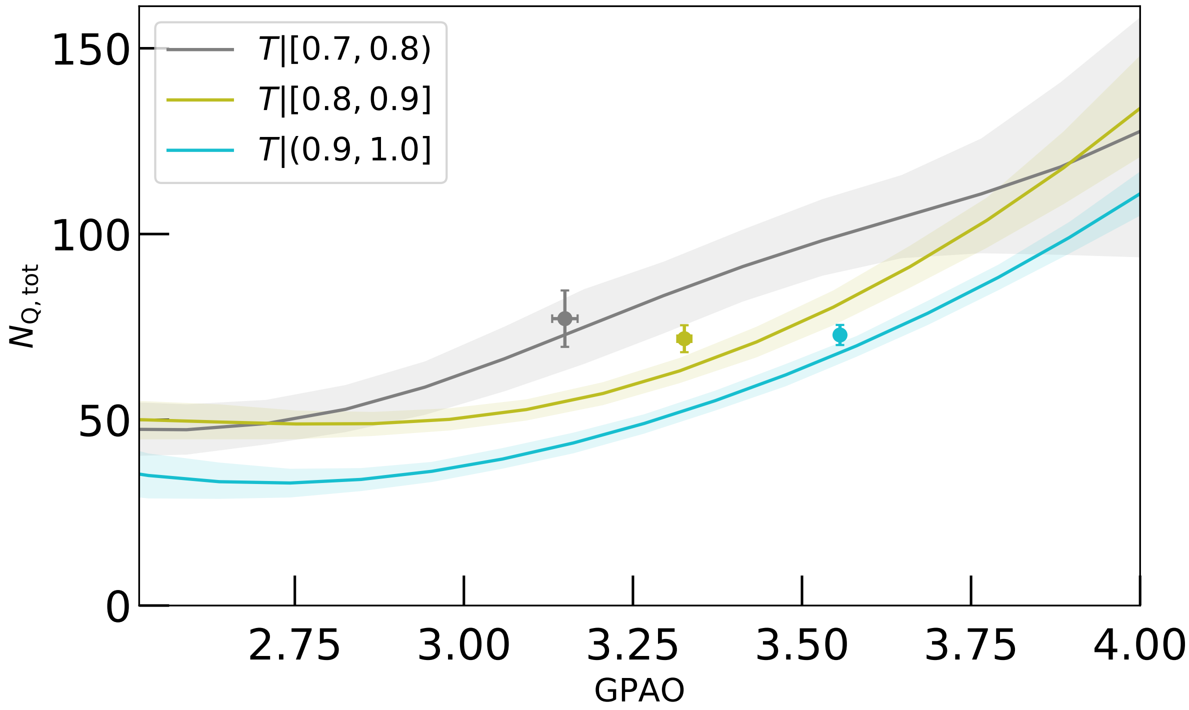


Figure 5.7: Study volume as a function of GPAO, shown for students in bins of T as indicated by the legend. Mean values are KLLR-derived using a kernel width 0.40 and 1σ uncertainties are shown as shaded regions. Population means for each T group are shown as points with $\pm1\sigma$ error bars. Significant study trends with GPAO exist within each math score range.

much to receive the same high grades as students with lower T scores.

Because study volume correlates to GPAO, its use would at least partially undermine our attempt to measure the benefits of study. Due to its correlation, any study gain measured using GPAO as baseline instead of T as baseline would show reduced study gains at fixed GPAO. We proceed now to quantify this reduction in measured gains.

5.3.3.2 Fitting GPE as a function of GPAO

We begin by fitting student grades earned in each physics class as a function of their end-of-term grade point average in all other courses (GPAO). As in §5.2.6, we use a KLLR fitting to discern the general trend from a relatively agnostic viewpoint.

Figure 5.8 shows the KLLR fits of mean GPE as a function of GPAO for both physics courses individually. As with $\mu_{\mathrm{GPE}}(T)$, we see significant positive curvature. The bulk of the population has GPE < GPAO (diagonal line), meaning that these physics courses tend to be the lower grades on their transcripts, especially so for students with low GPAO already. For example, a student with

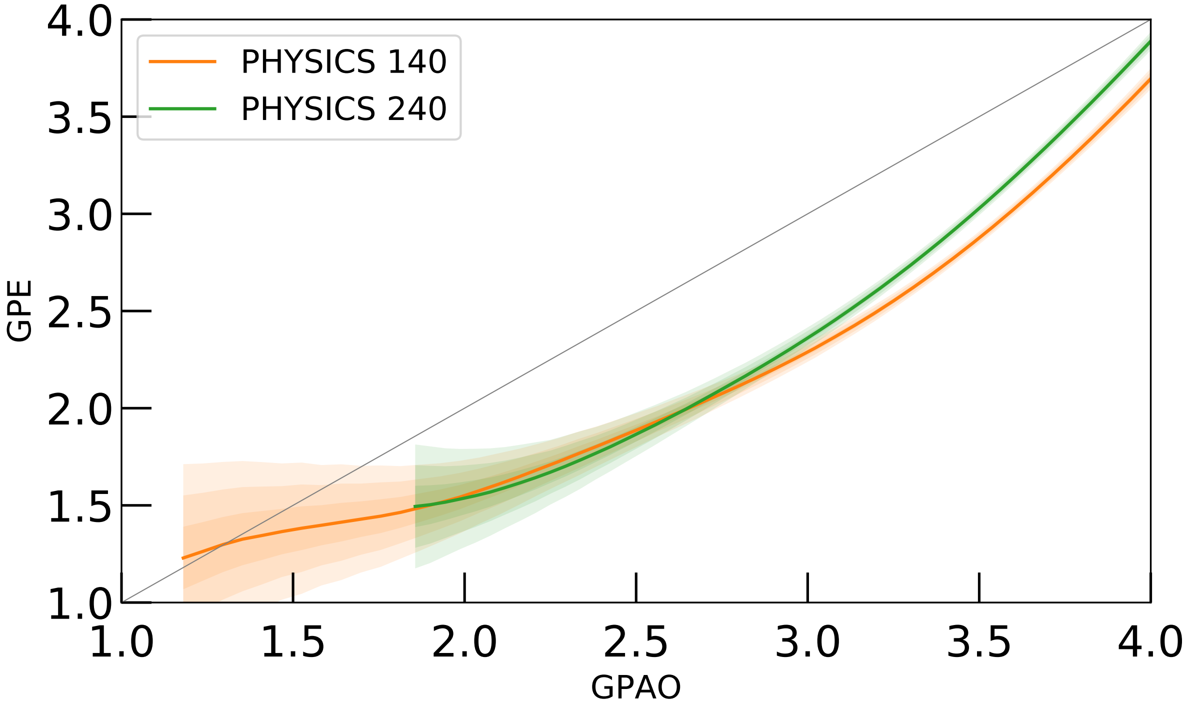


Figure 5.8: GPE vs GPAO: as Figure 5.3, but fitting mean GPE for the two physics courses as a function of GPAO rather than T. Grey line indicates equality, where a student's grade earned in physics would equal their average grades earned from other courses. The GPAO distribution was similar between courses, with means around 3.4, standard deviation of ±0.5, and GPAO $\gtrsim 2.45$ for 95% of students.

otherwise straight As (GPAO = 4.0) would likely get an A- while a student with otherwise straight Bs (GPAO = 3.0) would likely get a C+.

5.3.3.3 Deviations from GPAO mean due to study volume

After calculating a deviation from expected grade at a given GPAO, $\Delta = \text{GPE} - \mu_{\text{GPE}}(\text{GPAO})$, we investigate whether a trend exists in this deviation as a function of PR study volume, using KLLR fitting, analogous to that done in Figure 5.4. We present gains of study in Table 5.7, analogous to Table 5.6. Those with highest study volume tended to do better than expected by 0.28 ± 0.07 grade points (averaging over the three study indicators) as compared to those with low study volume.

This is less than half the grade gain measured in Table 5.6, where T was used as a baseline for comparison. The stark difference in grade gains measured suggests that the significant correlation between GPAO and study volume (see §5.3.3.1) has subsumed a large portion of the grade gains measured. That is, because GPAO reflects in part an individual's study volume, subtracting it out

Table 5.7: Using GPAO as a baseline to measure grade gains: as Table 5.6, but measuring gains and variance reduction with GPAO as a baseline for grade expectations (rather than T). Variance decreases $\Delta\text{Var}_{1\to2}$ are all insignificant.

Indicator	$\max(\Delta_{\text{GPE}})$	$\Delta\text{Var}_{0\to1}$	$\Delta\text{Var}_{0\to2}$
$N_{\text{Q,tot}}$	$+0.302 \pm 0.080$	-0.467 ± 0.010	-0.469 ± 0.009
N_{sess}	$+0.207 \pm 0.059$	-0.467 ± 0.010	-0.468 ± 0.010
$N_{\text{Q,mean}}$	$+0.335 \pm 0.053$	-0.467 ± 0.010	-0.471 ± 0.009

in the initial baseline grade estimate $\mu_{\text{GPE}}(\text{GPAO})$ removes a significant fraction of the measurable grade gains due to study. Despite GPAO predicting grades better than T (reducing variance by 47% rather than only 19%), GPAO has less utility in quantifying the grade benefits of study, washing out over half of the measurable grade benefit.

5.4 Discussion

Our findings help demonstrate to educators and learners the benefits of practice study in introductory physics courses. While this general finding is neither original nor surprising, its precise quantification is novel, and the overall magnitude in grade gain for students classified by pre-college math ability is large.

The low-stakes formative assessment of PR matches well to the summative midterm and final examinations of the two physics courses considered here. After all, these PR courses consist of nearly 1000 problems used on past examinations in each course. Given the median response time per problem of roughly 90 seconds [Weaverdyck et al., 2020], we do not think that memorization *per se* plays an important role. As mentioned earlier, exams consist of entirely novel questions (though certainly with some structural and topical overlap), so literal repetition of PR questions is quite unlikely. Rather than memorization, other factors such as the similarity of problem construction and structure, the mix of quantitative and qualitative questions, and particular "tricky" problem styles involving multiple physics concepts are likely to be similar between old and new exams.

Our findings with respect to demographic characteristics motivate more study of how better to support physics learning for first-generation (our ≤HS category) and underrepresented minority students (our URM category). Students from single-parent and low-income households also earn lower grades than expected (Fig. 5.5). These categories overlap, inviting future work to understand how students with multiple of these identities perform [Saw et al., 2018].

Our comparison to GPAO study trends reveals far more grade benefit measured when conditioning on T than on GPAO. Because GPAO significantly correlates with study volume (while T does not), subtracting out a GPAO baseline from grades subtracts out the effects of study behaviors to some extent. While GPAO may be the most accurate predictor of GPE, it is a flawed baseline for the purposes of our analysis (see also Appendix J) as it washes away the majority the the study benefit trend measured in §5.3.1. We thus find utility in using T as a baseline for student comparison, rather than GPAO.

The utility of T could in part be due to its significant correlation with general mental ability g [Frey and Detterman, 2004, Koenig et al., 2008, Coyle and Pillow, 2008, Sackett et al., 2008]. As g is largely intrinsic (polygenic, yet $\gtrsim 50\%$ heritable and $\gtrsim 90\%$ consistent with age; see Jensen [1998], Neisser et al. [1996], Bouchard [2013], Panizzon et al. [2014]) to each student, it is largely orthogonal to a student's external, personal choice of study habits (dependent more so on personality than on T; see §5.3.3.1). In contrast, GPAO represents a mix of both internal ability and external choice, and thus subsumes a portion of the study signal we attempt to measure in this study.

Another potential reason for the strong utility of T could lie in the similarity of assessment styles. Both standardized tests and the assessment methods used in many large introductory STEM courses take the form of high-stakes multiple-choice questions in a timed setting.

Regardless of interpretation, this difference in grade gains (measured with GPAO vs. T baselines) warrants further investigation and discussion. As more colleges move to "test-optional" admissions in the future, we may in the process be limiting the potential of future investigations to quantify student learning gains.

5.5 Chapter Conclusions

Using a large sample of practice study from the Problem Roulette service, we find that student final grades scale quadratically with the logarithm of the term-aggregated number of questions encountered, with an overall gain of nearly 0.8 grade points. By comparison, the largest demographic deviation we find is $\lesssim 0.3$ grade points, so with roughly a question per day on average, this deficit can be overcome purely through study.

We summarize our findings in the following advice to students and teachers. Our advice to students is:

> *Do a problem for every time you brush your teeth. Attempting at least one or two problems each school day of the term will likely lift your grade by one-quarter point. Quadratic gains with log study implies that while doubling study volume from one to two questions per day benefits you, doubling again from two to four questions per day benefits your grade even more!*

Our advice to teachers is:

> *Encouraging students to do at least one problem a day puts them above the median study habit; on average, this will land them above their expected grade. Study can benefit students by up to three-quarters of a letter grade. Differences between demographic groups, after accounting for T and study, were* $\lesssim 0.3$ *grade points; first-generation students were the most significant outlier.*

This study only scratches the surface of PR data available. Besides only working with physics courses, our model was relatively simple, treating *T* and GPE separately and using other demographic, academic, and study-related variables only tangentially. A more sophisticated model such as multi-level modeling or simultaneous fitting of many variables at once (such as with the machine learning tool SHAP—SHapley Additive exPlanations, see Lundberg and Lee [2017]) could help disentangle which factors are more or less causal (though it could not determine absolute causality). Future analyses should also treat nationality, ethnicity, and race more carefully, rather than using the broad categories of nationality and binary URM status. We also would like to investigate the effects of study session length—early results suggest that working for more than about 50 minutes tends to yield minimal gains. Finally, we wish to compare findings between all courses, observing which trends persist cross-subject.

We should never lose sight that the purpose of this work is to help individual students. Teachers and students alike can benefit from understanding how *T* and study affect their grades. Teaching is a handshake, requiring earnest participation of both parties for best results. (We can't fall into the trap of subscribing to either a student deficit model or a teacher deficit model alone; both parties need to improve their habits and grow as individuals.) As students improve their scholastic habits and teachers improve course structure and learn how to best reach struggling individuals, we can grow together and improve the education system.

Acknowledgements and Data Availability

WKB was partially funded by an Academic Innovation Fund award from the UM Center for Academic Innovation (CAI). The data were obtained from CAI and the UM Learning Analytics Data Architecture (LARC) database.

CHAPTER 6

Conclusions

A PhD is like a delta function of knowledge. If general education is a wide Gaussian, then a bachelor's degree in physics was like adding a narrower-width Gaussian on top. As I specialized from physics to cosmology to galaxy clusters to galaxy populations to modeling photometry of those populations to doing so with GMMs, the width of the metaphorical Gaussian of knowledge continued to narrow, approaching an infinitely narrow delta function. I believe it's correct to say that I'm the world expert in my sub-(sub-sub-etc.)field: using GMMs to select the RS and BC via photometry; I don't know of anyone on Earth who has spent more time thinking about it than I have. I have discovered many things regarding that practice, like good ways to account for uncertainty (by using error-corrected GMMs), how much GMM fit parameters depend on redshift or stellar mass, the importance of including bootstrap uncertainties when calculating BIC, and so forth.

I have also contributed to the field of learning analytics by publishing our study on PR, which 1) precisely quantified grade gains with incremental practice study volume and 2) showed the residual differences in student grades by demographic subgroup after accounting for T and study. Grade quantification showed both the power and limit of time spent on PR: students with high study volume scored three-quarters of a letter grade higher than those without study logged on PR. This both shows the significant benefits of study, but also can serve as a precaution against telling a C-student "just study more", as study on average gains students less than a letter grade. The demographic residual grades provide useful information regarding who needs extra help. Though on average there were no significant differences between men and women, there *were* significantly lower grades on average from first-generation students and (less significant but still notably) average lower grades in: underrepresented minority students, students with only one parent at home, and in students from low-income areas. More so than any other category, first-generation students need extra assistance, as at a given math proclivity and study volume, they still tend to earn lower grades than students whose parents attended college. More research is needed to investigate what precisely causes these differences, so we instructors can best aid struggling students.

I now shift focus towards the future, suggesting a course for research in these fields.

6.1 Future research using Red Dragon

RD is a hammer poised to hit some nails. In addition to doing more thorough analyses of deep fields, several other projects are ripe for analysis.

6.1.1 Delving deeper into space

The results of Chapter 4 focused on DES main bands ($griz$) at lower redshifts (using mass-complete samples), leaving much work to be done with this sample.

First and foremost, the extended photometry of u band and the VIRCAM bands JHK_s were only used in passing; features in these bands ought to be investigated. For example, the extended photometry would provide better constraints on rest-frame spectra (see Figure 4.13), allowing for a wider view of average RS and BC spectral features. These would help sort out whether observed features were purely due to deformities or asymmetries in filters or whether the features are redshift-dependent or whether features are stellar-mass dependent. This will nurture a more complete view of the two populations.

Secondly, with RS & BC parameterizations in hand, we can measure fits to luminosity function ϕ evolution over time for both RS and BC. In particular, we can fit the faint-end slope α and characteristic magnitude M^* over redshift (along with normalization ϕ^*, though those are more sample-dependent). A fit to these parameters can then yield an analytic expression for red fraction f_R as a function of redshift and magnitude for the entire sample. (However, this would be somewhat limited in the COSMOS dataset, especially at high masses, due to the effects of cosmic variance.)

Thirdly, while the lion's share of Chapter 4 focused on low-redshift results, I look forward to characterizing higher redshifts. Red Dragon fits two galaxy populations smoothly out to high redshifts ($z \sim 4$), though, as discussed earlier, this may be somewhat contaminated by a noise term, as the size of the quiescent population becomes negligible towards higher redshifts. A good first step towards a higher-redshift characterization would be to update from the COSMOS2015 dataset to the COSMOS2020 dataset, and add in supernovae fields from DES (which, similar to the COSMOS field, have superior photometry to the DES wide field survey). With increased numerical support at higher redshifts—vis. an increased population of quiescent galaxies—Red Dragon will be better able to characterize the RS out to high redshifts, expanding its capacity to quantify GMM parameters.

6.1.2 Color dependence on stellar mass & local density

Nearly two decades ago, Baldry et al. [2004] and Balogh et al. [2004] measured dependence of RS and BC parameterization (red fraction, mean colors, and scatters) on magnitude (a proxy of stellar mass) and local overdensity (measured by projected local densities Σ_5, computed using the distance to the fifth-nearest neighbor). Photometry was K-corrected to $z = 0$. It's about time for a modern update, including not only more colors than one, but also inter-color correlations and redshift evolution.

How do fit parameters $\boldsymbol{\theta}_{\mathrm{RS}}$ and $\boldsymbol{\theta}_{\mathrm{BC}}$ depend on galactic stellar mass $\mu_\star$, local density δ, redshift z, and perhaps other[1] parameters? If nothing else, a measurement of red fraction as it depends on stellar mass, redshift, and environment would be wonderful to have, as such measurements only exist for certain sub-samples of galaxies; a *universal* fit could provide stellar insights into galaxy evolution!

Similarly, it would be wonderful if we could investigate Baldry's dual linear splice fits [Baldry et al., 2004], confirming or updating the fitting functions, especially for extreme stellar masses or high redshifts. If a universal fit exists for mean colors and scatters (along with red fraction and correlations), those fits could be hardcoded into a RS/BC fitting code. Such a characterization of galaxy populations would allow for more detailed fits, ultimately more capable of identifying the RS and BC or predicting galactic redshift. This would provide fits for new colors and provide fits for covariances for the first time as functions of stellar mass.

We could then dive into more astrophysical phenomena. High-metallicity galaxies will tend towards a particular direction in (multi-)color space; how, precisely, would this and other properties (e.g. age, dustiness) respond? Where do they lie in color space, relative to mean RS/BC values? This could then perhaps be reversed, e.g. with principle components analysis (PCA), such that astrophysical properties could be estimated from the galaxy's position in color space. Such analysis may not only improve RS/BC characterization, but could also tease out novel components, be they starburst galaxies, mid-quenching galaxies, or other galactic populations.

6.1.3 Using RD to find clusters and select cluster members

The Red Dragon algorithm could also be used to find clusters and select cluster members. In short, it would act much like the redMaPPer cluster finder algorithm, though now fitting for BC as well as RS. Adding in an entirely new population of galaxies could vastly improve both cluster identification (especially at $z \gtrsim 1$, where due to the Butcher–Oemler effect, the fraction of quiescent galaxies diminishes towards vanishingly small values) and reduce contamination for cluster members, ultimately providing more reliable mass proxies.

Rather than remain agnostic as to RS/BC membership as with WaZP [Aguena et al., 2021], RD would purposefully employ both, using characterizations of both populations to estimate a galaxy's redshift. So long as the mean colors never overlap in multi-dimensional color space, this would

[1] For example, geometry—proximity to filaments or clusters—may have different effects from local density; cluster velocity dispersion or cluster mass may also play a distinct role.

have flawless accuracy if all galaxies were in the center of the RS or BC color distributions at all redshifts. While galaxies certainly scatter off of mean colors at every redshift, the populations are often distinct enough from each other and across redshift that, given a galaxy's position in multi-color space, we not only can estimate a galaxy's RS/BC membership status but we can also estimate its approximate redshift.

In a field of view, given a trained dragon, we can then estimate redshift for all galaxies by measuring membership likelihoods across all redshifts (rather than at the estimated redshift, as is used when creating a dragon initially) and finding the maximum (while using its spread as an uncertainty). This will then yield two redshift estimates: one assuming it is a RS member and another assuming it is a BC member—the preferred redshift would then be that with the higher likelihood. This then gives us a redshift estimate $z_{\rm RD}$ for each galaxy in some field of view, along with uncertainties σ_z and RS/BC membership estimation $P_{\rm red}$. This assumes that the galaxy is at the *center* of the RS/BC distribution, so it may be somewhat biased for galaxies with extreme dustiness, age, or metallicity (which drive displacement from mean colors). However, even with nonzero scatters in RS and BC, reasonably good photo-z values can be estimated.

We can then, for some spectroscopic sample, compare accuracies of a RD-defined redshift $z_{\rm RD}$ to its spectroscopic redshift $z_{\rm spec}$, determining how RD photo-z estimates compare to other photo-z estimators. Preliminary results show that RD photo-z errors $e_z \equiv (z_{\rm RD} - z_{\rm spec})/(1 + z_{\rm spec})$ yield mean values of $\langle e_z \rangle = .005$, signifying precise characterization, consistent with accuracies of WaZP. Using the interquartile range method of Schmidt et al. [2020, roughly equivalent to the scatter of e_z, excluding outliers], we find $\sigma_{\rm IQR} = .03$ for RD-estimated redshifts, as compared to the $O(.02)$ values scored by the dozen photo-z algorithms considered (albeit for a different dataset; more testing is needed to rigorously quantify RD photo-z capabilities). We can then calculate $z_{\rm RD}$ for all galaxies in a wider field of view.

Next, using some viewing window of the sky (hard cut or Gaussian-weighted, dynamic with redshift), we find a sky map of stacked redshift estimates, downweighted by their uncertainties. This could be done by adding log likelihood distributions $\mathcal{L}_\alpha(z)$ (for each component α of RS/BC), resulting in a summed likelihood for galaxies being at any queried redshift at each point on the sky. This then would give a sky map of likely redshifts; points with high likelihood would then likely correspond to galaxy cluster locations. If there are 200 galaxies in a small region which all share a common redshift estimate, then there is likely a galaxy cluster at that location. Again, like redMaPPer, this finds clusters, but with the added information of the BC aiding an otherwise RS-exclusive method.

Once a likely cluster has been identified, its members can be found by running RD at that redshift on all galaxies in the field of view. If galaxies are likely members of the RS or BC at that redshift, then they are likely cluster members; otherwise if their membership likelihoods at that redshift are low for both, they are likely foreground or background galaxies. Resulting contamination could be compared to other photo-z methods on a subset of spec-z identified clusters.

A good first step towards quantifying the results of this methodology would be to test it on simulations. The CosmoDC2 synthetic sky catalogue [Korytov et al., 2019] contains not only photometry, SFR, $\mu_\star$ values (along with other galactic information, like central black hole mass), but also truth labels about host halos, central galaxy status, and RS membership. This could thus quantitatively score how well a Red Dragon cluster finder predicts cluster memberships for each galaxy. This could potentially improve cluster classification, improving optical mass proxies and thereby cosmological and astrophysical analyses of galaxy clusters.

6.2 Future research with Problem Roulette

Our modeling of student grades was conditioned on only math proclivity T and study volume $N_{\mathrm{Q,tot}}$, but far more sophisticated methods are available. While causality can only truly be teased out with experimentation (which is questionable to implement on an active student population), machine learning techniques like SHAP [SHapley Additive exPlanations, defined in Lundberg and Lee, 2017] can tease out relative (though not *absolute*) causal impact of various variables, determining which factors best predict final grade.

Preliminary investigation using SHAP for relative importance of demographic indicators shows that population of the student's HS zip code was an even better predictor of grade than parental education, and it was also a better predictor than all of {income, sex, N/URM status, and single parent status} combined (see definitions of Table 5.3). This surprising result could potentially relate to differences in resources available to school districts in rural, suburban, and urban areas. Results such as these can help educators focus on the populations in most need of assistance, helping each student improve learning outcomes.

There are also several cross-demographic trends we would like to investigate. For example, a few preliminary results yield surprising trends: international students tended to actually score lower grades with increasing parental education; the fourth income quartile (> $100k) consistently did worse than the third income quartile; international students performed superior than average except for the sub-group which were additionally from single-parent households. These trends and more are waiting to be investigated, to parse out causality and identify which student populations struggle most, and why.

Though not currently implemented in PR, **discrimination index** could be added to aid question curation [Ermie, 2016]. Discrimination index d is calculated for a particular question on some assessment by first dividing overall grades into low and high groups (typically the top and bottom 27% are used). The index is then the fraction of high-score students who got the question correct minus the fraction of low-score students who got the question correct.

$$d \equiv f_{\text{high, correct}} - f_{\text{low, correct}} \tag{6.1}$$

This then ranges $d|[-1, +1]$ for each question, where -1 means only low-scoring students answered correctly and $+1$ means only high-scoring students answered correctly. Scores $d < .2$ are considered non-discriminating whereas scores $d > .3$ are considered highly discriminating. If high-scoring students consistently do poorly on a certain question, especially if they do worse than average, then that question is perhaps poorly worded or a trick question, not properly assessing student performance. Discrimination index thus quantifies the utility of a question: how well it *discriminates* between students with high levels of comprehension and those without. By quantitatively scoring question utility, I will continually curate my courses, improving my tests (be they quizzes or exams) so they can be of greatest utility to both teacher and student alike.

The Age of AI

Large language models (LLMs) like ChatGPT, Bard, Perplexity, and others have opened a new era for the human race, moving from the *information* age into the *intelligence* age. We are still in the "Wild West" of handling AI in the classroom. As I've chatted with various physics departments in my job interviews, they've shared with me their hopes, sense of unknown, wariness, and wishes for the future of LLMs and other AI software in the classroom.

I have already planned several ways to use LLMs in the classroom, which I discussed in a U–M Teach-Out on ChatGPT [University of Michigan, 2023]. Of top priority is to teach students the *fallibility* of LLMs; secondarily is to teach students the *power* of LLMs. Working through problems with LLMs as copilot or analyzing how LLMs solve problems (and especially when they *fail* to) can deepen students' understanding of physics, helping them to better organize their own thinking and catch errors in their own reasoning. Moving forward, teaching and assessments must fundamentally change in order to help students make the largest positive use of the tools at their disposal.

Like any tool, LLMs have potential for good as well as for harm. ("A scalpel can be used to either heal a wound or to create one.") A quote from *Dune* is particularly relevant here:

> "Once, men turned their thinking over to machines in the hope that this would set them free. But that only permitted other men with machines to enslave them."
> – Frank Herbert

As we use ChatGPT and other tools in life, we must ensure that the tools don't enslave us; that is, we shouldn't *give up* our thinking to ChatGPT or other tools, but rather, we should only use it to *aide* our thinking—to think more effectively and more efficiently, so we can think *more*.

6.3 Final Thoughts

A scholar once told me that they believed most of modern technology wouldn't exist without the stars to motivate humans towards nobler causes. We are truly blessed to live on a planet with an atmosphere transparent at night, in a darker region of our Milky Way galaxy (far from brilliant star clusters or nebulae, which could overshadow or outshine our view of the skies), in a darker region of our local supercluster, *Lanieakea* (allowing us to see a larger fraction of our local structure, less obscured than it would be if the Milky Way were in the core of the Virgo Cluster). How fortunate we are to live early enough in the universe to be able to detect the CMB (and early enough to still see distant galaxies, before Hubble expansion rips them away from our view) yet late enough in the universe to be able to observe the matter and dark energy equality epoch! And how grateful *I* am to have been born in an era of CCD telescopes and handheld computers with GHz clock speeds! Without these inventions, my research would be impossible.

The "DIKW pyramid" shows a process of **data** being contextualized into meaningful **information** given understanding and insight to develop **knowledge** which is then applied soundly as **wisdom**. What a fantastic age we live in, where "educational data mining" (EDM) can be a common phrase or acronym thrown around—to live in an era where data is so abundant, and tools so freely available to analyze and learn from data, drawing out information and knowledge. But the jump from knowledge to wisdom is far from inevitable. It's easy to think, especially when one is intelligent or knowledgeable, that wisdom naturally follows from knowledge. But this hubris serves no one. Wisdom is ratified by its observed consequences, not by data analysis. As we continue to seek wisdom in how to best apply knowledge, may we all move with caution in a world which sometimes has more artificial intelligence than true wisdom.

With more tools at our disposal than ever before, we have great potential to learn about scales as large as our universe and as small as our selves; The sky is no limit.

Ad Astra!

APPENDIX A

Density Characterization

Density characterized using same vetting as in table 2.1. To account for cosmic / local variance, I used a fit of the comoving number density rather than the local number density directly. Fitting in log density space ensured positive number densities while fitting relative to the log of the expansion factor $a = 1/(1+z)$ better matches the universe's evolution rate. We find a fit of

$$\bar{\rho}(z) \sim (0.00429 \text{ cMpc}^{-3})\,(1+z)^{0.820} \exp\left\{.532\,[\ln(1+z)]^2\right\} \tag{A.1}$$

to the redshift-evolving mean number density of galaxies brighter than $0.2\,L_*$ in Buzzard (v2.0.0).

Deviations from the black line suggest LSS over- or under-densities at those redshifts. In the local, low-redshift universe, the overdensity near $z = 0.1$ indicates LSS near the simulated observer significant enough to deviate from the homogeneity observed at larger volumes. (Low redshift observations suffer more from the statistical errors of cosmic variance than higher redshifts.)

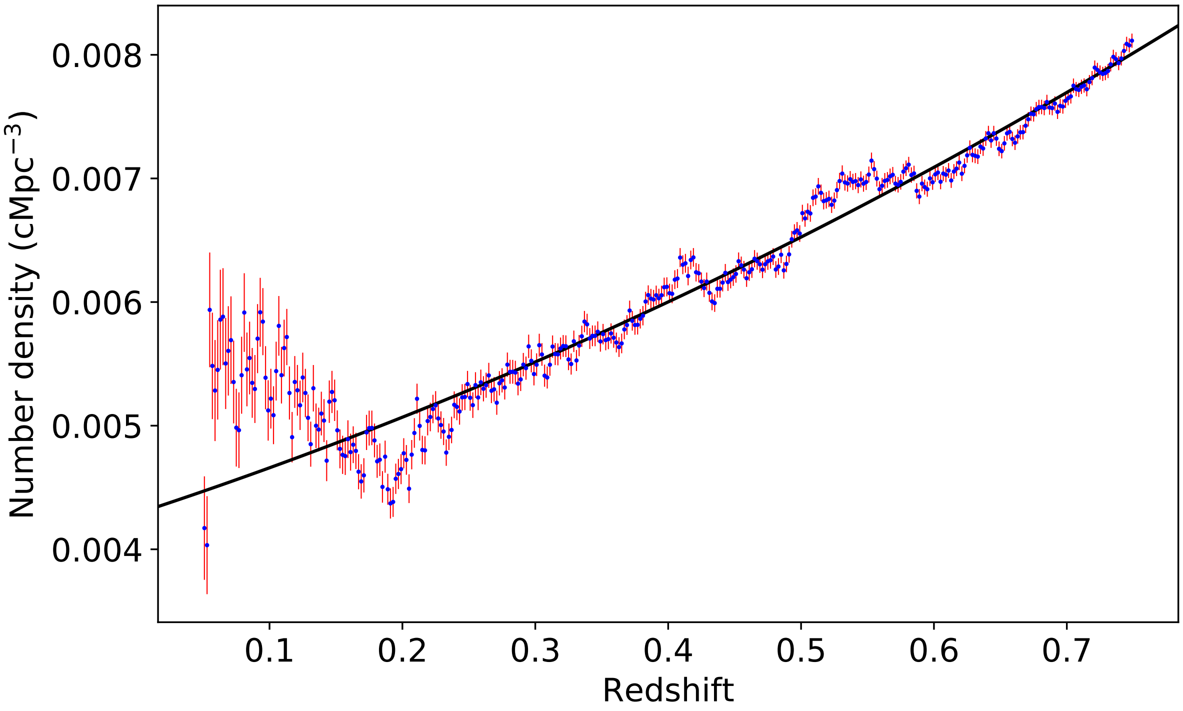


Figure A.1: Number density of $L > 0.2L_*$ galaxies in Buzzard. Power law fit (black line; see equation A.1) weighted by Poisson uncertainty (red; five sigma).

APPENDIX B

Radial profiles extended

See figures B.1 and B.2 below for binning in mass and redshift simultaneously. These show minimal redshift evolution, so combining all into the single frame of figure 2.13 loses little information while gaining statistical power.

Figure B.3 shows visually the distortions in profile shapes at high and low masses, visually depicting differences from NFW expectations.

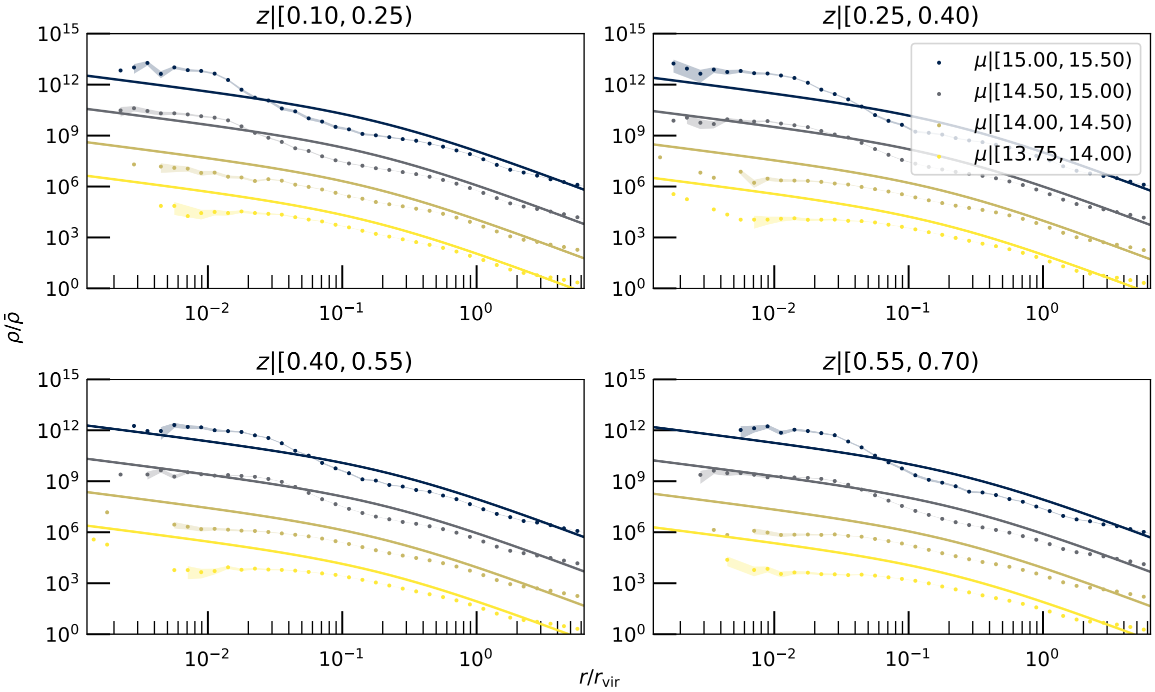


Figure B.1: Radial profiles of galaxies in clusters in contrast with NFW expectations.

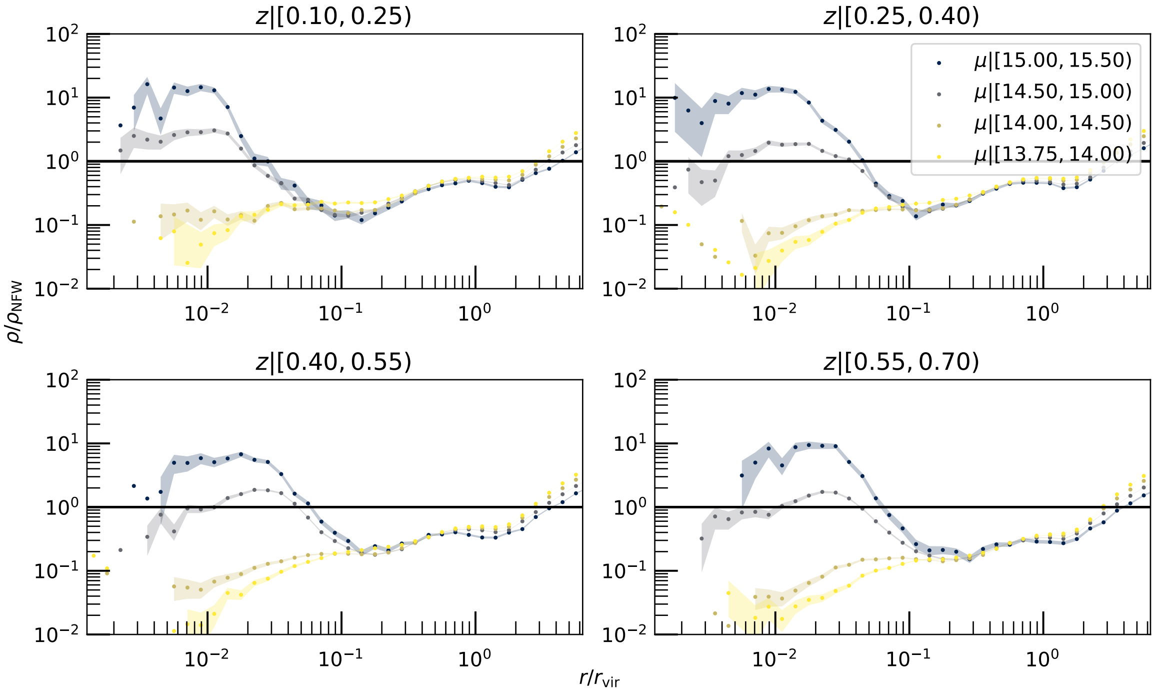


Figure B.2: Radial profiles of galaxies in clusters, given relative to NFW expectations.

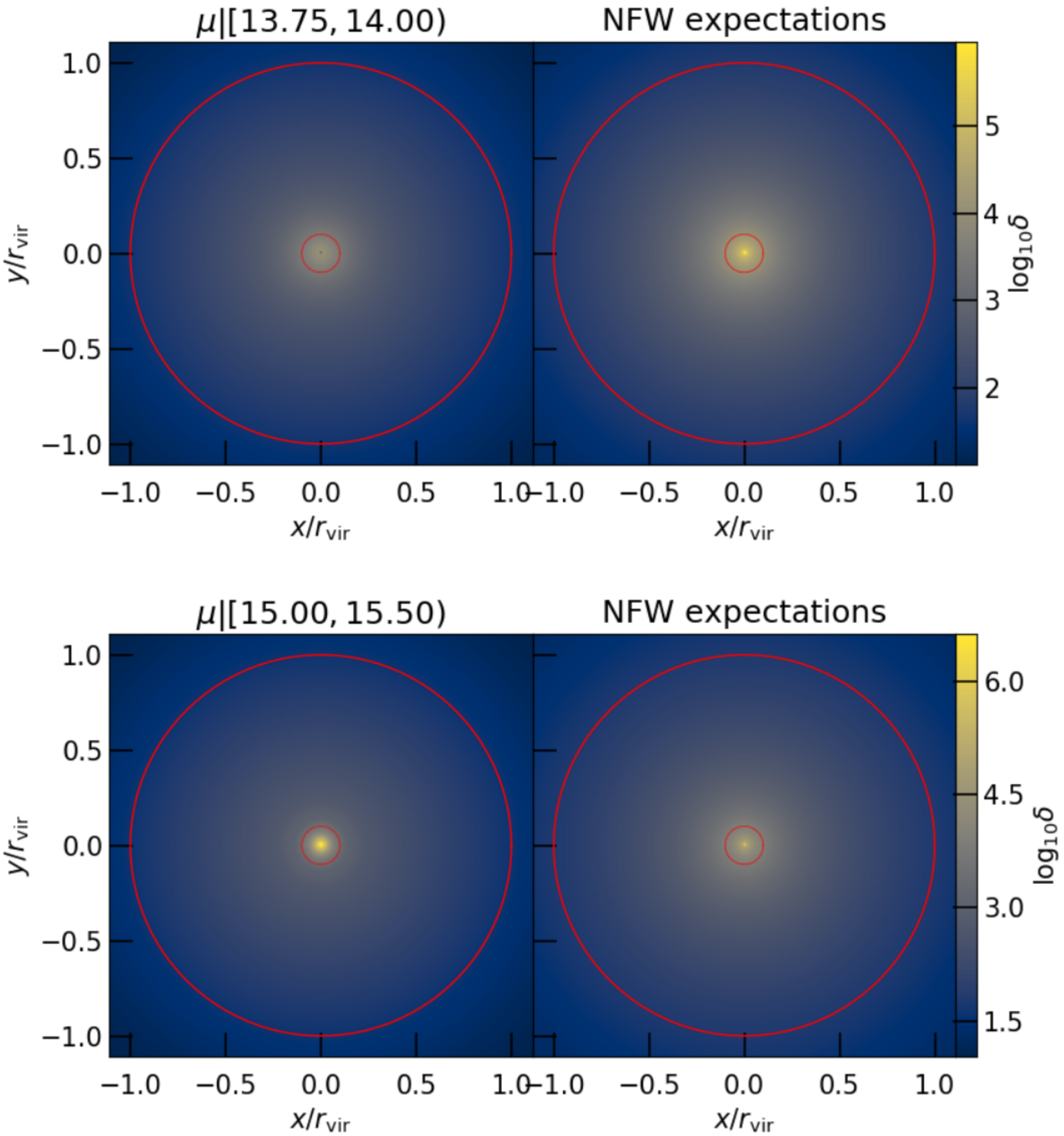


Figure B.3: Visualized 2D radial profiles, exemplifying differences between Buzzard (left) and observations (right), for a low-mass (top) and a high-mass (bottom) halo. Red circle indicates $r_{\rm vir}$ while the faint inner circles indicate 1/10 and 1/100 $r_{\rm vir}$.

APPENDIX C

Flipped Space

Here we present the inverse HOD relation to that of §2.3.1, i.e. log mass μ conditioned on richness λ. We use power law fit:

$$\langle \mu | \lambda \rangle = \alpha + \beta \log_{10}(\lambda / \lambda_p) \pm \sigma \tag{C.1}$$

(where $\lambda_p = 40$; pivot richnesses typically span $\lambda_p | [20, 70]$ in the literature). As in §2.3.1, I parameterize temporal evolution with $\zeta \equiv \ln \frac{1+z}{1+z_p}$, using $z_p = .4$ as a pivot redshift.

Figure C.1 shows the redshift evolution of the HOD in this flipped space. Our fit finds log mass normalization $\alpha = 14.82 \pm 0.01$, slope $\beta = (1.049 \pm 0.016) - (0.71 \pm 0.13)\,\zeta$, and intrinsic scatter $\sigma = 0.147 \pm 0.0021$. Though α generally includes redshift evolution, we find no evidence of evolution for α nor σ. In contrast, β shows $> 6\sigma$ evidence for evolution: a significant shallowing of the slope with redshift.

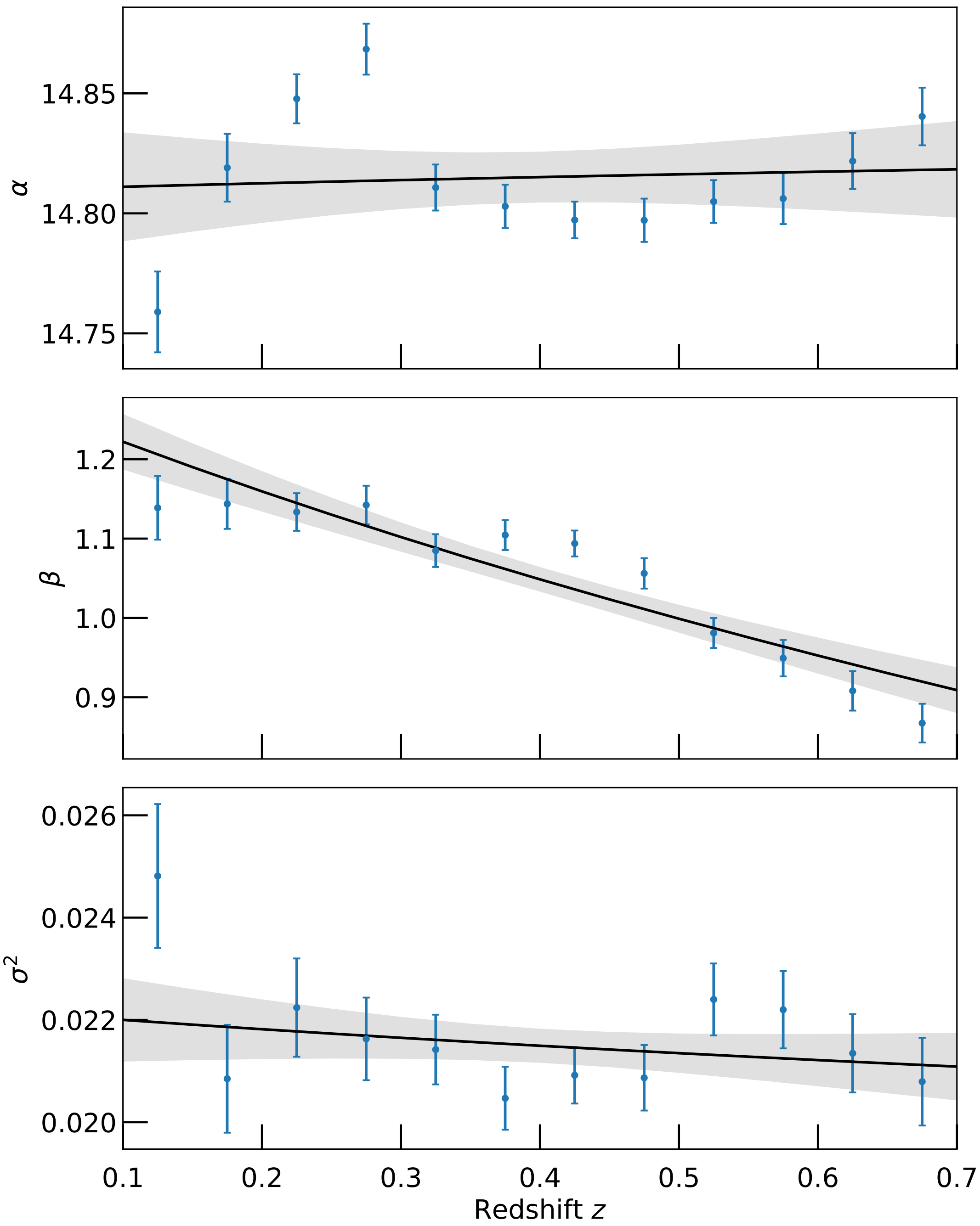


Figure C.1: Redshift evolution of mass–richness relation, given by equation C.1. Fit lines show mean and scatter in the mean as measured by linmix, fit in $\log((1 + z)/(1 + .4))$ space. Only the slope β shows significant evolution with redshift (> 6 sigma evidence, as compared to α and σ^2 which have < 1 sigma evidence for evolution).

APPENDIX D

Varying threshold of truth label

Because the distribution of sSFR is skew-lognormal rather than bimodal, hard cuts separating quenched and star-forming galaxies—such as that of equation (3.9)—are not robust. That is, shifts in the cut parameterization can result in relatively large shifts in truth label. (In contrast, if the distribution were bimodal, with a large valley between components, then small shifts to the dividing line would affect relatively few galaxies' truth labels.)

Here we show the effect of shifting the sSFR limit on the balanced accuracy of the SDSS/low-z sample, effectively recreating the Red Dragon points of Figure 3.4, extending the plot for various truth labels. Though balanced accuracy covers a range of about 10% in this plot, it's reasonable to expect equation (3.9) to be accurate within ±0.25 dex, which then corresponds to only a $\sim 3\%$ shift in accuracy between either extreme. Furthermore, we see here that *relative* characterization is somewhat consistent, with each of the values being within a few percent of each other across the board. This means that in all plots using a sSFR cut, one should read results as *relative* to each other (rather than as absolute measurements of accuracy).

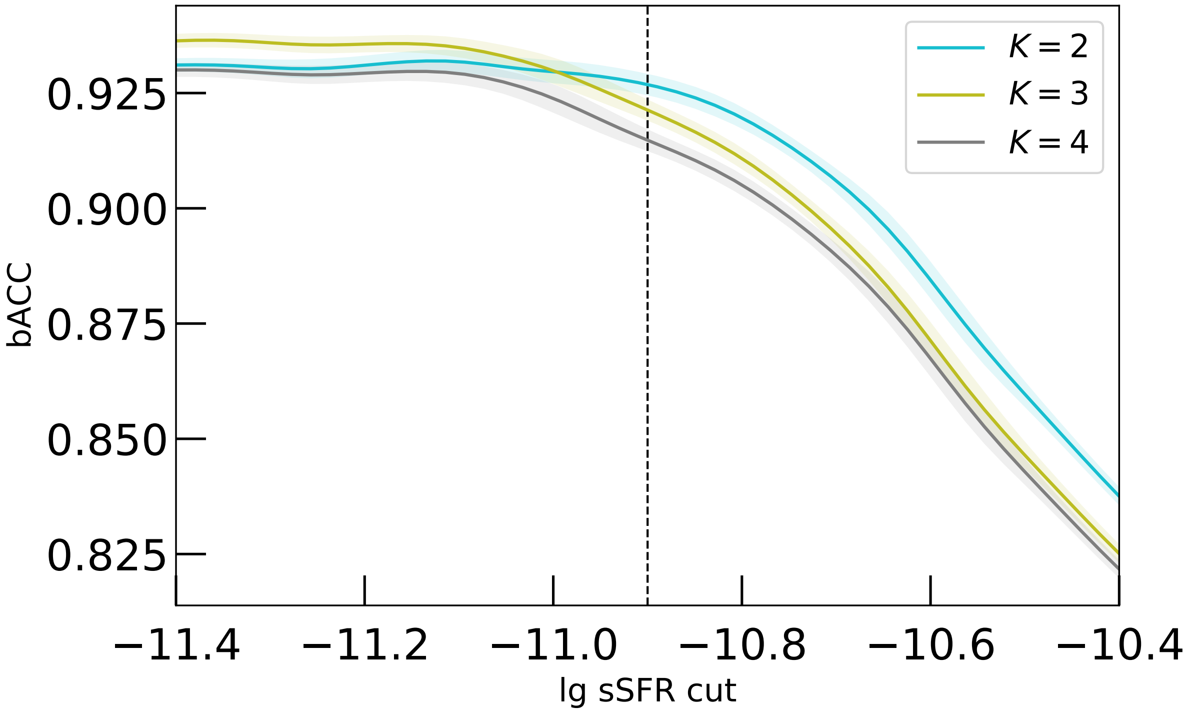


Figure D.1: Balanced accuracy shift due to changing sSFR truth label (the limit shown in equation (3.9); dashed line) for various Red Dragon component counts K. Ran on bootstrap resamplings of the SDSS/low-z sample; results then smoothed with KLLR.

APPENDIX E

Magnitude trends in Buzzard

Here we investigate trends of magnitude running as found in Buzzard. Though running of parameters θ with magnitude is statistically significant (see §3.5.4), slope of mean color relative to the scatter of the populations was relatively small (§E.1), so differences in selection were relatively minimal (§E.2). Magnitude trends had the largest effect for dimmer galaxies, resulting in $\lesssim$ 10% differences in selection.

E.1 Shift of mean color relative to RS width

Though the RS may have a statistically significant shift in mean color, it may not significantly change selection of the RS by Gaussian mixture if that running is small compared to the width of the RS. To measure this, we use the metric ς to quantify slope of the RS relative to its scatter:

$$\varsigma \equiv \frac{d\langle \mu_a \rangle / dM_b}{\sigma_a} \tag{E.1}$$

(for color c_a and magnitude M_b; the magnitude may be any—our three fits use M_r, M_i, and M_z).

This measure of **relative drift** quantifies the significance of magnitude differences. For a sample of $0.2\,L_*$ limited galaxies with luminosities following a Schechter function (with $\alpha|[-1.5, -1]$), $\gtrsim$ 99.9% of its galaxies fall in a magnitude spread of $\Delta\text{mag} < 4$. Using the $\pm 2\sigma_{\text{RS}}$ definition of RS width from Hao et al. [2009], we can then take 4σ as the distance needed to move such that the RS radically shifts with magnitude. These two factors of four then cancel, leaving ς as **a unitless metric for significance of RS running**. If $|\varsigma| < 1$, the magnitude variation of the RS mean color is completely within the typical scatter of the RS, but if $|\varsigma| > 1$, then the RS mean color moves beyond the typical scatter of the RS.

We measured ς values from SDSS data Baldry et al. [2004], the Buzzard flock, and a redMaPPer Rykoff et al. [2014] fit of the DES Y3 RS. Each of the three datasets had typical values of $|\varsigma_{\text{RS}}| \lesssim 0.5$ (focusing at a given redshift on the appropriate primary color from table 3.2), implying that in nearly every case, the running of the RS mean color was insignificant compared to its width. This implies that it takes many magnitudes to significantly shift the mean RS color, relative to its intrinsic scatter. Therefore, characterizing the RS in a magnitude-ignorant way will still properly select galaxies.

E.2 Similarity in selection

Though including magnitude running of GMM parameters would better represent the photometric population, the extra dimensionality of running with magnitude increases the number of parameters needed for fitting without drastically altering selection.

Though the current version of Red Dragon doesn't include magnitude running, there are a few workarounds to model it explicitly. 1) You can slice your data by magnitude and run Red Dragon to characterize a bright vs dim sample of your galaxies, then interpolate between (or simply bin using) the two fits to estimate P_{red} for individual galaxies. 2) If working with a relatively thin redshift slice, you can set in the code `Z=m_i`, i.e. let Red Dragon measure and interpolate across magnitude rather than redshift. We use this latter method to measure similarity in selection between magnitude-running and redshift-running versions of Red Dragon.

At several thin redshift slices (for $z|[.1, .7]$), we measured P_{red} differences between the color-only and magnitude-running dragons. In binary selection, these two dragons resulted in about 7% opposing characterization (i.e. 7% of galaxies were characterized as RS instead of BC or vice versa) with a balanced accuracy of about 94% (between magnitude-running dragons and redshift-running dragons; not an accuracy of selecting the quenched population). Looking at $\Delta P = (P_{\text{red,mag}} - P_{\text{red},z})$, we found that though $\sim$ 10% of galaxies had $|\Delta P| \gtrsim 0.25$, less than $\sim$ 3% of galaxies had $|\Delta P| \gtrsim 0.50$ (so few galaxies had significantly different characterization). While the effect of running with magnitude is notable, it can be ignored without much loss.

Magnitude matters most when you're uncertain of redshift, but the current version of Red Dragon is designed for clusters, where you have very good redshift estimates. For the current iteration of Red Dragon, the algorithm remains magnitude-ignorant, solely using colors to distinguish the RS from the BC.

APPENDIX F

Band Characterization

Table F.1 characterizes the photometric bandpass filters used in this analysis, including central wavelengths, filter widths, and entry/exit wavelengths for the 4000 Å break. Note that the exit and entry wavelengths don't usually coincide between bands, since filters may overlap (e.g. in the case of $r \longrightarrow i$) or have gaps between them (e.g. the VIRCAM filters). These entry and exit redshifts can be shifted for any other rest-frame wavelength λ (rather than 4000 Å):

$$Z_\lambda = (Z_{4\mathrm{k}} + 1)\frac{4000\ \text{Å}}{\lambda} - 1 \tag{F.1}$$

So, for example, the Balmer break at $\lambda_\mathrm{B} = 3647.05$ Å would have shifted transition redshifts, such that $Z_\mathrm{B} \approx 1.097(Z_{4\mathrm{k}} + 1) - 1 = 1.097Z_{4\mathrm{k}} + .097$ and the Lyman limit at $\lambda_\mathrm{L} = 911.763$ Å would have $Z_\mathrm{L} \approx 4.4(Z_{4\mathrm{k}} + 1) - 1$. Thus Balmer transition redshifts are roughly 10% larger than 4000 Å break redshifts. Note that the Lyman limit enters g band at $Z_\mathrm{L} \approx 3.37$, just beyond the maximum redshifts considered in this paper.

Transition wavelengths between filters are taken as the geometric midpoint between edges, i.e.:

$$\lambda_\mathrm{tr} \equiv \sqrt{\left(\lambda_{\mathrm{c},a} + \frac{1}{2}\Delta\lambda_a\right)\left(\lambda_{\mathrm{c},b} - \frac{1}{2}\Delta\lambda_b\right)} \tag{F.2}$$

for bands a and b (using the geometric mean rather than the arithmetic tends to make a $< 1/1000$ difference, but it makes more sense to think in terms of log wavelengths, rather than linear, so geo mean makes more sense). This yields the transition wavelengths listed in Table F.2.

Table F.1: Entry and exit redshifts of the 4kÅ break: Redshifts at which a 4000 Å ("4k") rest-frame wavelength source will enter ("EN") and exit ("EX") visible and NIR bands from DECam (*ugriz*) and VIRCAM (JHK_s). Note that central wavelengths λ_c and widths $\Delta\lambda$ (FWHM) belie the asymmetries and rounded edges of each filter's transmission, so crossing redshifts are only approximate. Also note that the central wavelength for *u* band is shorter than even the Balmer break, so the entry redshifts are negative.

band	λ_c (Å)	$\Delta\lambda$ (Å)	$Z_{4k,EN}$	$Z_{4k,EX}$
u	3552	885	-.22	0
g	4730	1503	-.01	.371
r	6415	1487	.418	.790
i	7835	1470	.776	1.14
z	9260	1520	1.12	1.50
J	12 523	1725	1.92	2.35
H	16 451	2915	2.75	3.48
K_s	21 467	3090	3.98	4.75

Table F.2: Transition wavelengths: locations at which a monochromatic signal moves from being picked up by one band more than another.

bands	λ_{tr} (Å)
$u \longrightarrow g$	3987
$g \longrightarrow r$	5576
$r \longrightarrow r$	7131
$i \longrightarrow z$	8534
$z \longrightarrow J$	10,800
$J \longrightarrow H$	14,200
$H \longrightarrow K_s$	18,900

APPENDIX G

Other color fits to main dragon

Here we detail the remaining color parameterization of the $griz$ dragon, including the remaining mean colors, scatters, and correlations not discussed in the main text.

We note here that Figure G.1 shows the RD fit to the RS in $g - r$ has a ~ 0.5 mag divergence from RM near $z = 1$. However, the RM-measured slope of mean color with respect to luminosity is significantly larger than this offset. Therefore, relative to Figures 4.5 & G.2, and taking into account divergence from fit relative to slope offset, the fit is still in fair agreement. Also worth noting: this separation is roughly the size of the measured RS scatter at that redshift (see Figure G.3), so to 2σ, the lines are in fine agreement.

We note that Figure G.4 shows RD measuring a significantly larger RS scatter than RM, by roughly a factor of three in places (most often closer to a factor of two). This could perhaps be related to a difference in data quality, in estimated uncertainties in $i - z$ color between the DES Y3 sample as compared to the COSMOS2015 sample. For example, if the DES Y3 catalogue overestimated uncertainties in z band, or if the COSMOS2015 catalogue underestimated the same, then this could explain the difference in measured scatter.

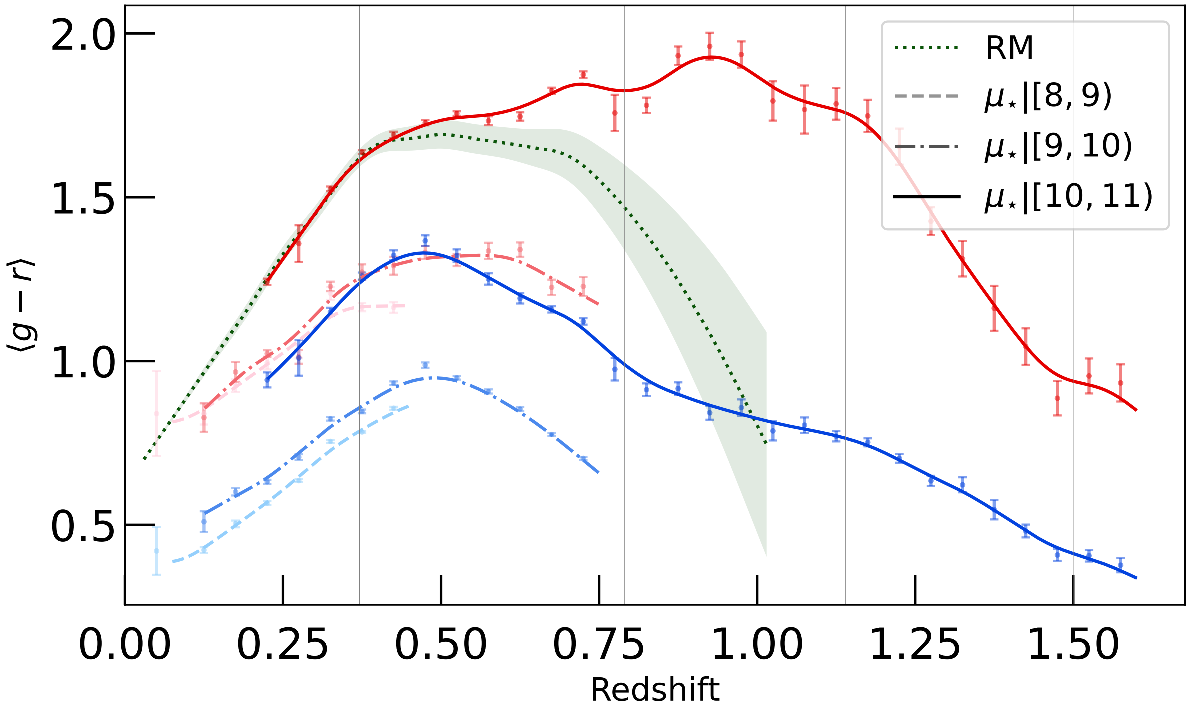


Figure G.1: As Figure 4.5, but for $g - r$.

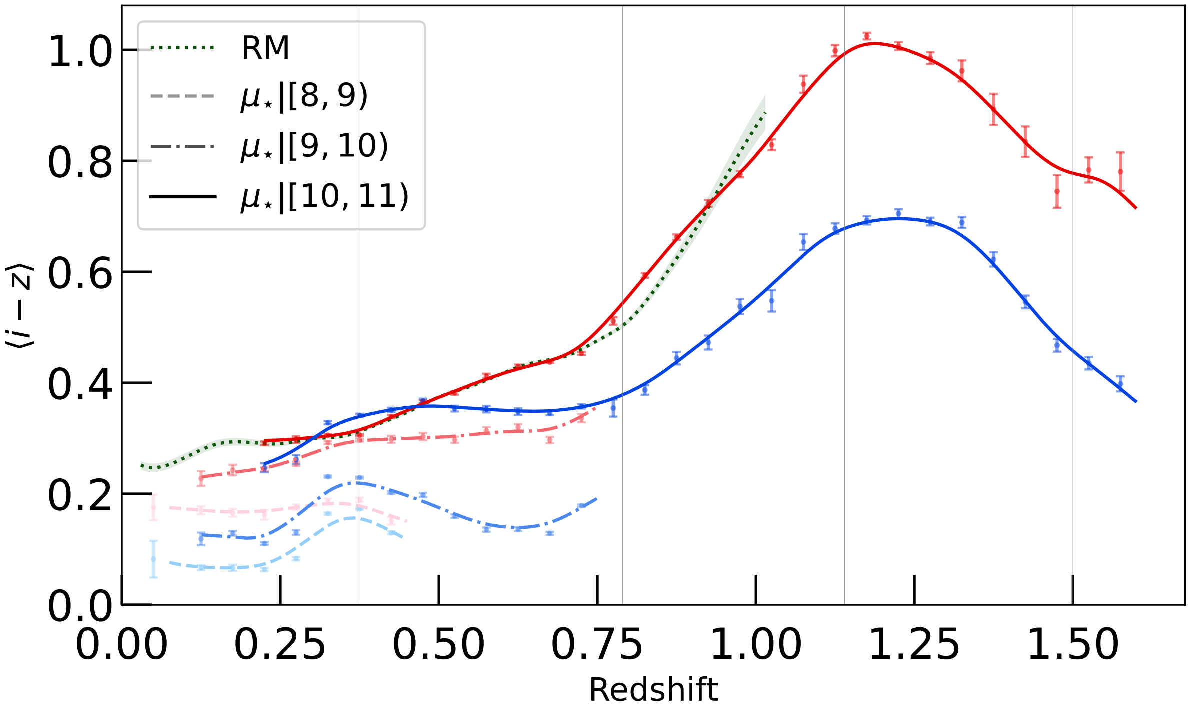


Figure G.2: As Figure 4.5, but for $i - z$.

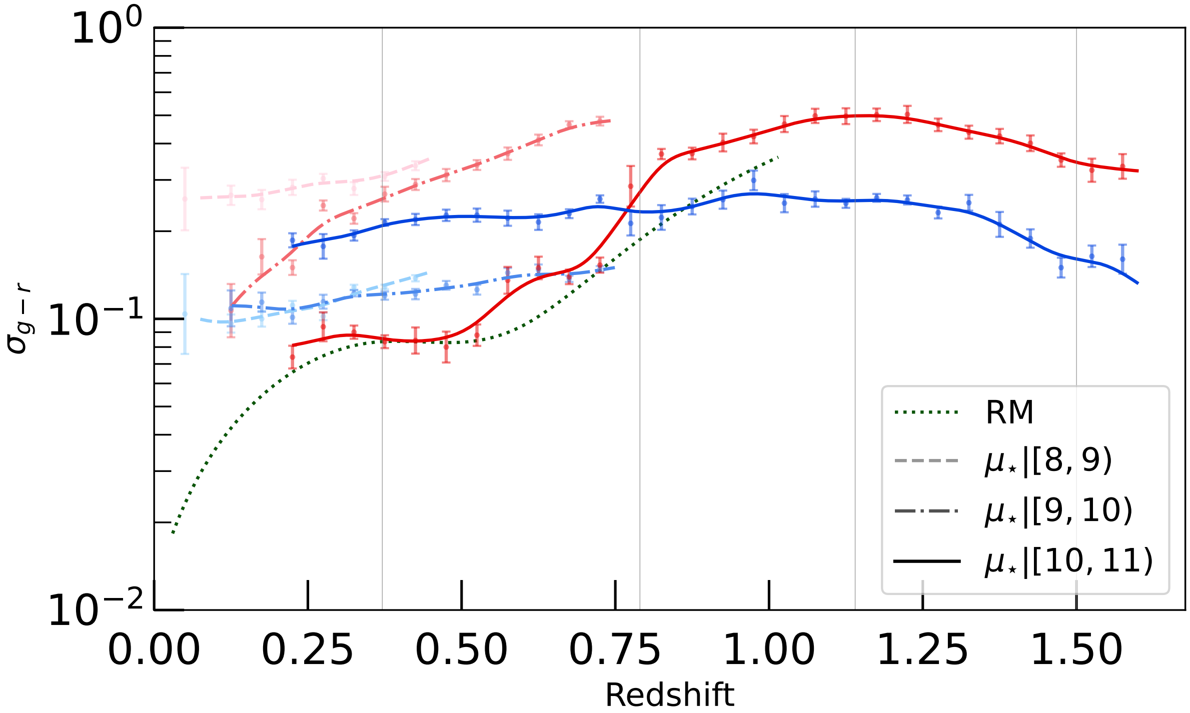


Figure G.3: As Figure 4.6, but for $g - r$.

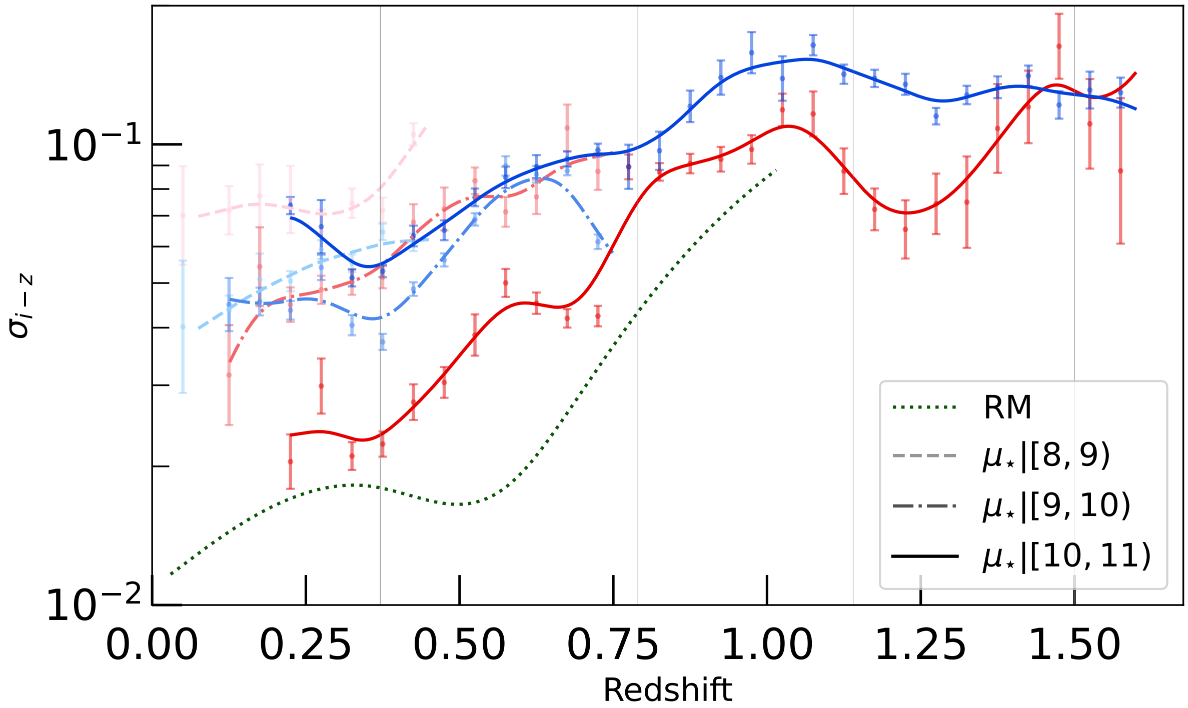


Figure G.4: As Figure 4.6, but for $i - z$.

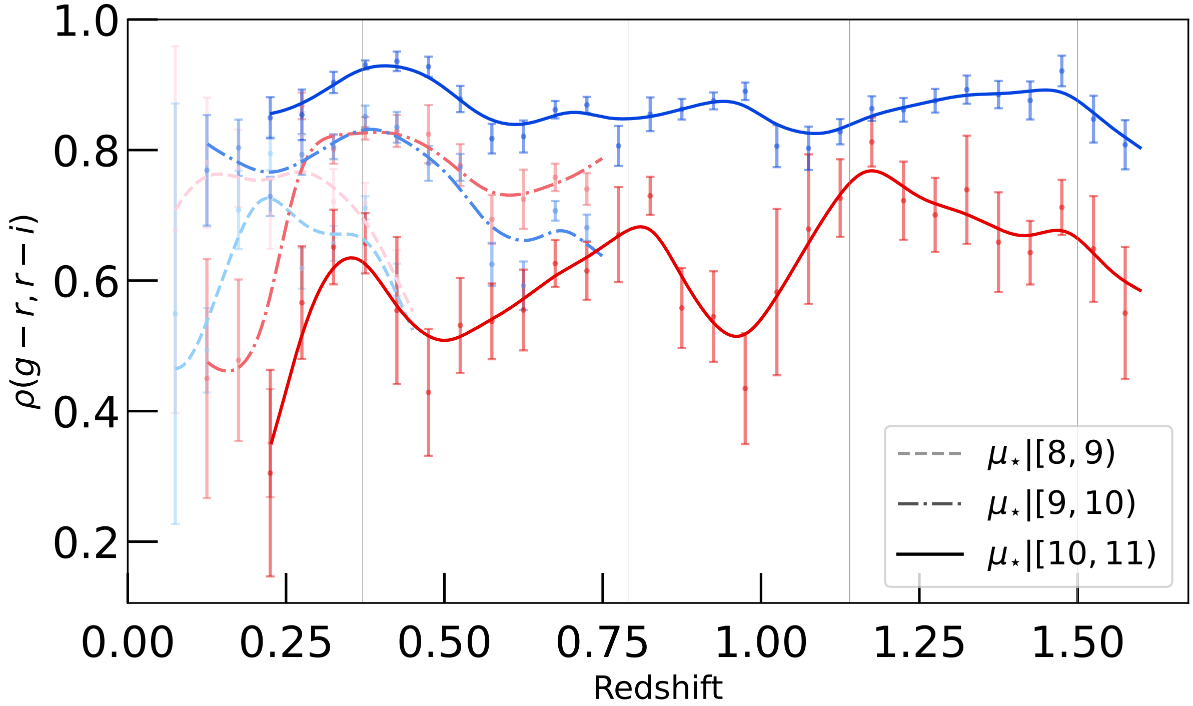


Figure G.5: As Figure 4.7, but for $\rho(g-r, r-i)$.

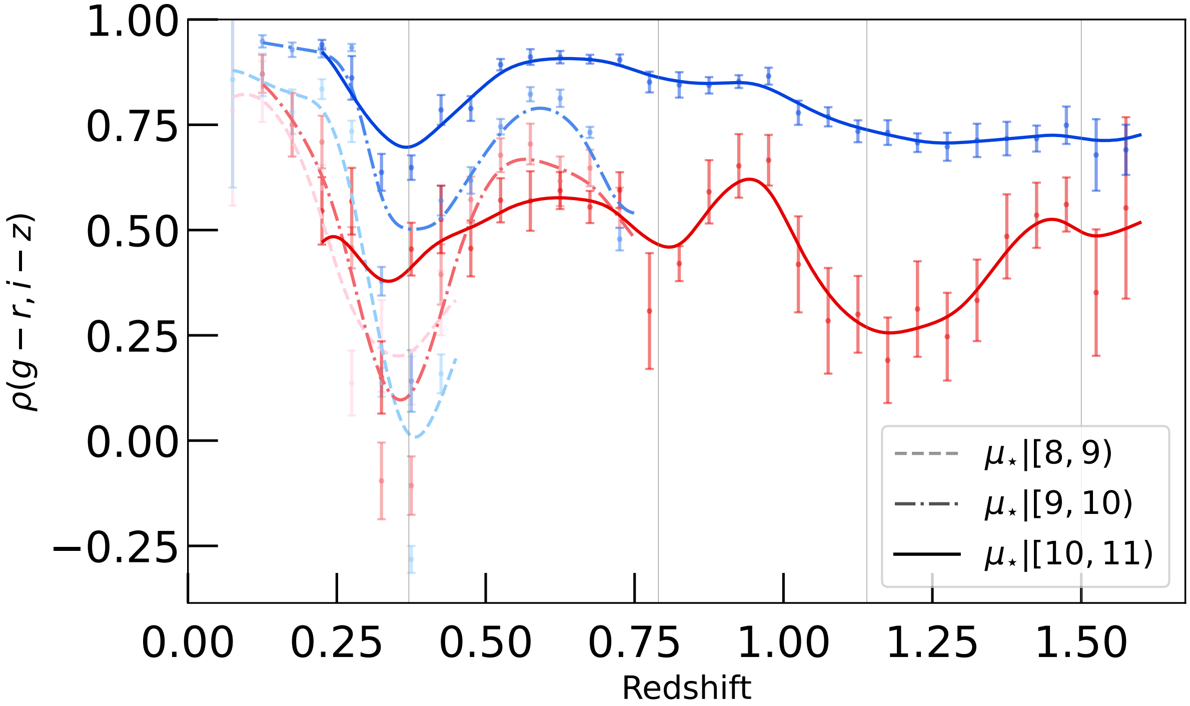


Figure G.6: As Figure 4.7, but for $\rho(g-r, i-z)$.

APPENDIX H

Permissible thickness of mass and redshift bins

What step size in redshift or galactic stellar mass causes a significant drift in RD fit parameters $\vec{\theta}$?

H.1 Color slope method

One straightforward way to quantify this is to compare how much the RS mean color drifts in comparison to its scatter. Using redshift as an example (though one can mirror the analysis using stellar mass), if a step Δz in redshift makes the mean color shift by more than its scatter, we must consider decreasing redshift bin width in order to not miss significant spectral features. In particular, as $\pm 2\sigma_{\rm RS}$ has historically been used to select the RS, if the mean color drifts by $4\sigma_{\rm RS}$, it would be completely excluded from selection, so we use $4\sigma_{\rm RS}$ as an upper limit. We can then demand $\Delta z \cdot d\vec{c}_{\rm RS}/dz < 4 \cdot \sigma_{\rm RS}$, so

$$\Delta z < \frac{4\sigma_{\rm RS}}{d\vec{c}_{\rm RS}/dz} = \Delta z_{\rm max}. \tag{H.1}$$

This then estimates a maximum z-bin width, beyond which colors vary more within the bin than they scatter at a fixed z. A more conservative bin width would be half this value, but there would be little justification to move bin width below a quarter this value (unless, as will be discussed in §H.2, other parameters drift significantly quicker than mean color does compared to its scatter).

From our three main fits here, we find $\Delta z_{\rm max}$ tends to be lowest (most strict) at lower redshifts and at higher stellar masses. In particular, we measure a minimum of $\Delta z_{\rm max} \sim .12$, implying that redshift bins should not exceed this width if the RS is to be properly characterized (at low redshifts and high stellar masses). Both the increased RS scatter and the decreased RS mean color slope with redshift at lower stellar masses lead to higher values of $\Delta z_{\rm max}$; we find values of .25 and .75 for the middle and lowest mass sample respectively. (This analysis can also be performed for the BC; we find strictest values of $\Delta z_{\rm max}$ for each sample were consistent with $\sim .3$, indicating a looser requirement for the highest-mass sample but a stricter requirement for the lowest-mass sample).

This analysis can also be performed with stellar mass-running dragons, using $d\vec{c}/d\mu_\star$. Preliminary results find strictest values of $\Delta\mu_{\star,\mathrm{max}} \sim 1$, implying that within a stellar mass decade, the RS and BC mean colors tend to not experience a drift in mean colors greater than four times the RS scatter. However, parameters besides mean color may experience significant drift with redshift or stellar mass, suggesting stricter bin width requirements.

H.2 Parameter curvature method

A more conservative approach would use bootstrap uncertainties of fit parameters and curvature of fit parameters to determine desired bin width.

Generally speaking, if an unknown function $f(x)$ is linear in a given domain, then only two points (at any distance from each other) are needed to characterize the curve, with no constraint on bin width; in contrast, if the function has significant curvature, this constrains bin size. In particular, one might hope to have sufficiently sampled space so as to properly characterize all curvatures in the function, ensuring that all peaks are identified. If all points sampled of the function have some uncertainty, then one could demand **insignificant curvature between neighboring points**. This would distinguish outliers from significant peaks.

In particular, using finite differencing, the curvature $f_{xx,i}$ of a function at point x_i is estiamted by

$$f_{xx}(x_i) \approx \frac{f_{i-1} - 2f_i + f_{i+1}}{h^2} \tag{H.2}$$

(where h is bin width). Demanding no significant curvature between points would then imply

$$\frac{f_{xx}}{\sigma_{f_{xx}}} \lesssim 1. \tag{H.3}$$

We can then estimate an optimal bin width using this constraint. Because $f_{xx} \propto h^{-2}$, we find that to make point-to-point curvature insignificant, we must use bin width

$$h_{\mathrm{new}} = h_{\mathrm{old}} \cdot \sqrt{\frac{\sigma_{f_{xx}}}{f_{xx}}}. \tag{H.4}$$

Thus, larger relative uncertainties on curvature $\sigma_{f_{xx}}/f_{xx}$ allow for larger bin widths while more significant curvatures $f_{xx}/\sigma_{f_{xx}}$ demand narrower bins. Using h_{new} thus ensures that all point-to-point trends are consistent to 1σ with linear, attesting that the function is well-characterized, without missing any significant peaks of $f(x)$.

Applying equation (H.4) to our parameters $\vec{\theta}$ then gives us a strict estimate on bin width, setting a minimum reasonable bin width for a given dragon. Rather than strictly adhere to equation H.3, one could reasonably allow for significance of curvature up to $f_{xx}/\sigma_{f_{xx}} \lesssim 5$, only excluding undeniably significant curvatures. Restating these bounds in terms of a given parameter θ, we then have

$$h_{\mathrm{old}} \cdot \sqrt{\frac{\sigma_{\theta_{xx}}}{\theta_{xx}}} \lesssim h_{\mathrm{new}} \lesssim h_{\mathrm{old}} \cdot \sqrt{\frac{5\sigma_{\theta_{xx}}}{\theta_{xx}}}. \tag{H.5}$$

Using the left constraint for component weights, mean colors, log variances, and correlations, we find that across parameters and components, the typical desired redshift bin width h_{new} falls around $\Delta z = .04$ and a minimum across parameters towards $\Delta z = .013$, leading to an upper constraint of $\Delta z \lesssim .065$ as a strict value of a maximum permissible bin width for redshift. Our study's bin width of $\Delta z = .05$ is thus permissible, albeit near the wide side (as is desired, for the sake of increasing RS number counts in each bin, improving statistical power).

Constraints on bin width are often most stringent due to sharp evolution of colors, but depending on sample, other elements of $\vec{\theta}$ demand narrow bin widths: weight, correlation, or even scatter (in decreasing frequency) can also issue the strictest calls for narrow bin widths.

The same methodology can be used for magnitude-running dragons. Analyzing one such dragon, we find preliminary results of a typical desired bin width of $\Delta\mu_\star \sim .35$ and a minimum desired bin width of $\Delta\mu_\star > .1$ (implying an upper limit of $\Delta\mu_\star \lesssim .5$ for negligible parameter evolution within a mass bin). This implies that using bins any thinner than a tenth of a mass decade is unreasonable for the purposes of GMM characterization of galaxy colors. This suggests that using a full decade for a stellar mass bin width is too lenient, as it may miss significant parameter evolution.

H.3 Suggestions

In summary, we find no evidence of needing bin width thinner than $\Delta z < .01$ for redshift and $\Delta\mu_\star < .1$ for stellar mass. Future modeling of galaxy populations in photometric space should ensure their models use bin widths $\Delta z \lesssim .05$ and $\Delta\mu_\star \lesssim .5$; analyses otherwise risk missing significant evolution of parameters within their bins.

We end with a warning that if mass bins are made too wide, then high-μ BC galaxies are likely to be characterized as RS galaxies. As this population tends to have lower star formation rates and is likely to be in the middle of mass quenching, this characterization may not be entirely wrong, though it is perhaps incorrect.

APPENDIX I

Problem Roulette structure

After logging in and selecting a course, students can choose from various modes of study: Individual, Group, or Practice Exam (student-generated or faculty-generated as formal practice exams). Problem Roulette also collects questions from topics that students struggle with most, which students can study from.

In the Individual or Group study modes, students select which topics to study (e.g., “Vector Algebra”, “Relative Motion”), whether or not to use a timer for the session, and how many questions to pull (defaults: 10, 25, all). They are then presented with an exam-like set of questions one at a time. After the session, they are presented with a review of the questions attempted with correct answers indicated, along with an overall accuracy score and the session duration.

Instructors can view course analytics for a given term. At the instructor dashboard, they see student activity (e.g., which days saw more PR use), student accuracy versus questions answered for each topic, the accuracy of answers by topic, and the accuracy of individual questions answered. This information can reveal which topics or types of questions students tend to struggle with most or least.

APPENDIX J

Additional issues with GPAO

Though GPAO correlates more with course grade earned (GPE) than any other indicator we investigated, it has a several issues. The most crucial issue of its correlation with study is detailed in §5.3.3; here we delineate several additional issues with using GPAO as a baseline for GPE comparison (vs. T as a baseline, as used in this paper), that is, using grade anomaly GPE − GPAO instead of the metric GPE − $\mu_{\text{GPE}}(T)$ used herein.

J.1 Relation to personality traits

Personality traits correlate significantly with grades,Köseoglu [2016] mediated through study habits.Aluja and Blanch [2004] Because GPAO depends on study and some demographic populations have different mean study behaviors than others, significant differences exist between some different demographic sub-populations. For example, male students tend to study less than female students, and they tend to have a lower GPAO than female students by roughly 0.07 to 0.14 points, despite having significantly higher T scores on average.Richardson et al. [2012] This means that in a class where each student got identical grades, the grade anomaly for males and females could still differ for causes *extrinsic* to the course in question.

J.2 Differing course selection

GPAO depends on course selection of individual students, but course selection differs drastically by demographics, and not all course loads are equal. If all students took similar courses regardless of demographic group, then GPAO would be a fair measure of "anomaly"—how differently the course was graded compared to expectations from other courses. However, different groups tend to take different courses, leading to systemic shifts between different demographic sub-groups. If a student tends to take easier courses, they could have a 4.0 GPAO, then this would allow the student more free time to study on PR. However, a similar signal comes from a student who, despite taking incredibly challenging courses, works very hard and still has a 4.0 GPAO, yet then has very little free time to study on PR. While this exemplifies why GPAO has power in predicting GPE, it also shows that interpretation of GPAO is clouded.

J.3 Deletion of campus-wide trends

Grade anomaly GPE – GPAO wipes out campus-wide biases. Expressing this numerically: if every course in the school consistently awarded lower grades by a bias b to a given group g versus the rest of the population p, then the grade anomaly of that group would be

$$\begin{aligned}(\text{grade anomaly})_g &= \text{GPE}_g - \text{GPAO}_g \\ &= (\text{GPE}_p - b) - (\text{GPAO}_p - b) \\ &= \text{GPE}_p - \text{GPAO}_p = (\text{grade anomaly})_p,\end{aligned} \tag{J.1}$$

so the group then has identical grade anomaly to the rest of the population. This then erases the effects of campus-wide biases.

BIBLIOGRAPHY

T. M. C. Abbott, F. B. Abdalla, S. Allam, A. Amara, J. Annis, J. Asorey, S. Avila, O. Ballester, M. Banerji, W. Barkhouse, and et al. The dark energy survey: Data release 1. *The Astrophysical Journal Supplement Series*, 239(2):18, Nov 2018. ISSN 0067-0049. doi: 10.3847/1538-4365/aae9f0. URL https://doi.org/10.3847%2F1538-4365%2Faae9f0.

T. M. C. Abbott, M. Aguena, A. Alarcon, S. Allam, S. Allen, J. Annis, S. Avila, D. Bacon, K. Bechtol, A. Bermeo, G. M. Bernstein, E. Bertin, S. Bhargava, S. Bocquet, D. Brooks, D. Brout, E. Buckley-Geer, D. L. Burke, A. Carnero Rosell, M. Carrasco Kind, J. Carretero, F. J. Castander, R. Cawthon, C. Chang, X. Chen, A. Choi, M. Costanzi, M. Crocce, L. N. da Costa, T. M. Davis, J. De Vicente, J. DeRose, S. Desai, H. T. Diehl, J. P. Dietrich, S. Dodelson, P. Doel, A. Drlica-Wagner, K. Eckert, T. F. Eifler, J. Elvin-Poole, J. Estrada, S. Everett, A. E. Evrard, A. Farahi, I. Ferrero, B. Flaugher, P. Fosalba, J. Frieman, J. García-Bellido, M. Gatti, E. Gaztanaga, D. W. Gerdes, T. Giannantonio, P. Giles, S. Grandis, D. Gruen, R. A. Gruendl, J. Gschwend, G. Gutierrez, W. G. Hartley, S. R. Hinton, D. L. Hollowood, K. Honscheid, B. Hoyle, D. Huterer, D. J. James, M. Jarvis, T. Jeltema, M. W. G. Johnson, M. D. Johnson, S. Kent, E. Krause, R. Kron, K. Kuehn, N. Kuropatkin, O. Lahav, T. S. Li, C. Lidman, M. Lima, H. Lin, N. MacCrann, M. A. G. Maia, A. Mantz, J. L. Marshall, P. Martini, J. Mayers, P. Melchior, J. Mena-Fernández, F. Menanteau, R. Miquel, J. J. Mohr, R. C. Nichol, B. Nord, R. L. C. Ogando, A. Palmese, F. Paz-Chinchón, A. A. Plazas, J. Prat, M. M. Rau, A. K. Romer, A. Roodman, P. Rooney, E. Rozo, E. S. Rykoff, M. Sako, S. Samuroff, C. Sánchez, E. Sanchez, A. Saro, V. Scarpine, M. Schubnell, D. Scolnic, S. Serrano, I. Sevilla-Noarbe, E. Sheldon, J. Allyn. Smith, M. Smith, E. Suchyta, M. E. C. Swanson, G. Tarle, D. Thomas, C. To, M. A. Troxel, D. L. Tucker, T. N. Varga, A. von der Linden, A. R. Walker, R. H. Wechsler, J. Weller, R. D. Wilkinson, H. Wu, B. Yanny, Y. Zhang, Z. Zhang, and J. Zuntz. Dark energy survey year 1 results: Cosmological constraints from cluster abundances and weak lensing. *Physical Review D*, 102(2):023509, Jul 2020. doi: 10.1103/PhysRevD.102.023509. URL https://link.aps.org/doi/10.1103/PhysRevD.102.023509.

T. M. C. Abbott, M. Aguena, A. Alarcon, S. Allam, O. Alves, A. Amon, F. Andrade-Oliveira, J. Annis, S. Avila, D. Bacon, E. Baxter, K. Bechtol, M. R. Becker, G. M. Bernstein, S. Bhargava, S. Birrer, J. Blazek, A. Brandao-Souza, S. L. Bridle, D. Brooks, E. Buckley-Geer, D. L. Burke, H. Camacho, A. Campos, A. Carnero Rosell, M. Carrasco Kind, J. Carretero, F. J. Castander, R. Cawthon, C. Chang, A. Chen, R. Chen, A. Choi, C. Conselice, J. Cordero, M. Costanzi, M. Crocce, L. N. da Costa, M. E. da Silva Pereira, C. Davis, T. M. Davis, J. De Vicente, J. DeRose, S. Desai, E. Di Valentino, H. T. Diehl, J. P. Dietrich, S. Dodelson, P. Doel, C. Doux, A. Drlica-Wagner, K. Eckert, T. F. Eifler, F. Elsner, J. Elvin-Poole, S. Everett, A. E. Evrard,

X. Fang, A. Farahi, E. Fernandez, I. Ferrero, A. Ferté, P. Fosalba, O. Friedrich, J. Frieman, J. García-Bellido, M. Gatti, E. Gaztanaga, D. W. Gerdes, T. Giannantonio, G. Giannini, D. Gruen, R. A. Gruendl, J. Gschwend, G. Gutierrez, I. Harrison, W. G. Hartley, K. Herner, S. R. Hinton, D. L. Hollowood, K. Honscheid, B. Hoyle, E. M. Huff, D. Huterer, B. Jain, D. J. James, M. Jarvis, N. Jeffrey, T. Jeltema, A. Kovacs, E. Krause, R. Kron, K. Kuehn, N. Kuropatkin, O. Lahav, P. F. Leget, P. Lemos, A. R. Liddle, C. Lidman, M. Lima, H. Lin, N. MacCrann, M. A. G. Maia, J. L. Marshall, P. Martini, J. McCullough, P. Melchior, J. Mena-Fernández, F. Menanteau, R. Miquel, J. J. Mohr, R. Morgan, J. Muir, J. Myles, S. Nadathur, A. Navarro-Alsina, R. C. Nichol, R. L. C. Ogando, Y. Omori, A. Palmese, S. Pandey, Y. Park, F. Paz-Chinchón, D. Petravick, A. Pieres, A. A. Plazas Malagón, A. Porredon, J. Prat, M. Raveri, M. Rodriguez-Monroy, R. P. Rollins, A. K. Romer, A. Roodman, R. Rosenfeld, A. J. Ross, E. S. Rykoff, S. Samuroff, C. Sánchez, E. Sanchez, J. Sanchez, D. Sanchez Cid, V. Scarpine, M. Schubnell, D. Scolnic, L. F. Secco, S. Serrano, I. Sevilla-Noarbe, E. Sheldon, T. Shin, M. Smith, M. Soares-Santos, E. Suchyta, M. E. C. Swanson, M. Tabbutt, G. Tarle, D. Thomas, C. To, A. Troja, M. A. Troxel, D. L. Tucker, I. Tutusaus, T. N. Varga, A. R. Walker, N. Weaverdyck, R. Wechsler, J. Weller, B. Yanny, B. Yin, Y. Zhang, J. Zuntz, and DES Collaboration. Dark Energy Survey Year 3 results: Cosmological constraints from galaxy clustering and weak lensing. *Physical Review D*, 105(2):023520, January 2022. doi: 10.1103/PhysRevD.105.023520.

Susmita Adhikari, Tae-hyeon Shin, Bhuvnesh Jain, Matt Hilton, Eric Baxter, Chihway Chang, Risa H. Wechsler, Nick Battaglia, J. Richard Bond, Sebastian Bocquet, and et al. Probing galaxy evolution in massive clusters using act and des: splashback as a cosmic clock. *arXiv:2008.11663 [astro-ph]*, Aug 2020. URL http://arxiv.org/abs/2008.11663. arXiv: 2008.11663.

N. Aghanim, Y. Akrami, M. Ashdown, J. Aumont, C. Baccigalupi, M. Ballardini, A. J. Banday, R. B. Barreiro, N. Bartolo, S. Basak, R. Battye, K. Benabed, J.-P. Bernard, M. Bersanelli, P. Bielewicz, J. J. Bock, J. R. Bond, J. Borrill, F. R. Bouchet, F. Boulanger, M. Bucher, C. Burigana, R. C. Butler, E. Calabrese, J.-F. Cardoso, J. Carron, A. Challinor, H. C. Chiang, J. Chluba, L. P. L. Colombo, C. Combet, D. Contreras, B. P. Crill, F. Cuttaia, P. de Bernardis, G. de Zotti, J. Delabrouille, J.-M. Delouis, E. Di Valentino, J. M. Diego, O. Doré, M. Douspis, A. Ducout, X. Dupac, S. Dusini, G. Efstathiou, F. Elsner, T. A. Enßlin, H. K. Eriksen, Y. Fantaye, M. Farhang, J. Fergusson, R. Fernandez-Cobos, F. Finelli, F. Forastieri, M. Frailis, A. A. Fraisse, E. Franceschi, A. Frolov, S. Galeotta, S. Galli, K. Ganga, R. T. Génova-Santos, M. Gerbino, T. Ghosh, J. González-Nuevo, K. M. Górski, S. Gratton, A. Gruppuso, J. E. Gudmundsson, J. Hamann, W. Handley, F. K. Hansen, D. Herranz, S. R. Hildebrandt, E. Hivon, Z. Huang, A. H. Jaffe, W. C. Jones, A. Karakci, E. Keihänen, R. Keskitalo, K. Kiiveri, J. Kim, T. S. Kisner, L. Knox, N. Krachmalnicoff, M. Kunz, H. Kurki-Suonio, G. Lagache, J.-M. Lamarre, A. Lasenby, M. Lattanzi, C. R. Lawrence, M. Le Jeune, P. Lemos, J. Lesgourgues, F. Levrier, A. Lewis, M. Liguori, P. B. Lilje, M. Lilley, V. Lindholm, M. López-Caniego, P. M. Lubin, Y.-Z. Ma, J. F. Macías-Pérez, G. Maggio, D. Maino, N. Mandolesi, A. Mangilli, A. Marcos-Caballero, M. Maris, P. G. Martin, M. Martinelli, E. Martínez-González, S. Matarrese, N. Mauri, J. D. McEwen, P. R. Meinhold, A. Melchiorri, A. Mennella, M. Migliaccio, M. Millea, S. Mitra, M.-A. Miville-Deschênes, D. Molinari, L. Montier, G. Morgante, A. Moss, P. Natoli, H. U. Nørgaard-Nielsen, L. Pagano, D. Paoletti, B. Partridge, G. Patanchon, H. V. Peiris, F. Perrotta, V. Pettorino, F. Piacentini, L. Polastri, G. Polenta, J.-L. Puget, J. P. Rachen, M. Reinecke, M. Remazeilles,

A. Renzi, G. Rocha, C. Rosset, G. Roudier, J. A. Rubiño-Martín, B. Ruiz-Granados, L. Salvati, M. Sandri, M. Savelainen, D. Scott, E. P. S. Shellard, C. Sirignano, G. Sirri, L. D. Spencer, R. Sunyaev, A.-S. Suur-Uski, J. A. Tauber, D. Tavagnacco, M. Tenti, L. Toffolatti, M. Tomasi, T. Trombetti, L. Valenziano, J. Valiviita, B. Van Tent, L. Vibert, P. Vielva, F. Villa, N. Vittorio, B. D. Wandelt, I. K. Wehus, M. White, S. D. M. White, A. Zacchei, and A. Zonca. Planck 2018 results - vi. cosmological parameters. *Astronomy and Astrophysics*, 641:A6, Sep 2020. ISSN 0004-6361, 1432-0746. doi: 10.1051/0004-6361/201833910.

M Aguena, C Benoist, L N da Costa, R L C Ogando, J Gschwend, H B Sampaio-Santos, M Lima, M A G Maia, S Allam, S Avila, D Bacon, E Bertin, S Bhargava, D Brooks, A Carnero Rosell, M Carrasco Kind, J Carretero, M Costanzi, J De Vicente, S Desai, H T Diehl, P Doel, S Everett, A E Evrard, I Ferrero, A Ferté, B Flaugher, P Fosalba, J Frieman, J García-Bellido, P Giles, R A Gruendl, G Gutierrez, S R Hinton, D L Hollowood, K Honscheid, D J James, T Jeltema, K Kuehn, N Kuropatkin, O Lahav, P Melchior, R Miquel, R Morgan, A Palmese, F Paz-Chinchón, A A Plazas, A K Romer, E Sanchez, B Santiago, M Schubnell, S Serrano, I Sevilla-Noarbe, M Smith, M Soares-Santos, E Suchyta, G Tarle, C To, D L Tucker, and R D Wilkinson. The wazp galaxy cluster sample of the dark energy survey year 1. *Monthly Notices of the Royal Astronomical Society*, 502(3):4435–4456, Apr 2021. ISSN 0035-8711. doi: 10.1093/mnras/stab264.

Lana Al-Shawwa, AA Abulaban, A Merdad, S Baghlaf, A Algethami, J Abu-shanab, and A Balkhoyor. Differences in studying habits between male and female medical students of king abdulaziz university (kau), jeddah, saudi arabia. *Egyptian Dental Journal*, 60(2):1687–1693, 2014.

Steven W. Allen, August E. Evrard, and Adam B. Mantz. Cosmological parameters from observations of galaxy clusters. *Annual Review of Astronomy and Astrophysics*, 49(1):409–470, Sep 2011. ISSN 0066-4146, 1545-4282. doi: 10.1146/annurev-astro-081710-102514. URL http://arxiv.org/abs/1103.4829. arXiv: 1103.4829.

Anton Aluja and Angel Blanch. Socialized personality, scholastic aptitudes, study habits, and academic achievement: Exploring the link. *European Journal of Psychological Assessment*, 20 (3):157–165, Jan 2004. ISSN 1015-5759. doi: 10.1027/1015-5759.20.3.157. URL https://econtent.hogrefe.com/doi/10.1027/1015-5759.20.3.157.

Saad S. Alzahrani, Yoon Soo Park, and Ara Tekian. Study habits and academic achievement among medical students: A comparison between male and female subjects. *Medical Teacher*, 40(sup1):S1–S9, Jul 2018. ISSN 0142-159X. doi: 10.1080/0142159X.2018.1464650. URL https://doi.org/10.1080/0142159X.2018.1464650.

A. Amon, D. Gruen, M. A. Troxel, N. MacCrann, S. Dodelson, A. Choi, C. Doux, L. F. Secco, S. Samuroff, E. Krause, J. Cordero, J. Myles, J. DeRose, R. H. Wechsler, M. Gatti, A. Navarro-Alsina, G. M. Bernstein, B. Jain, J. Blazek, A. Alarcon, A. Ferté, P. Lemos, M. Raveri, A. Campos, J. Prat, C. Sánchez, M. Jarvis, O. Alves, F. Andrade-Oliveira, E. Baxter, K. Bechtol, M. R. Becker, S. L. Bridle, H. Camacho, A. Carnero Rosell, M. Carrasco Kind, R. Cawthon, C. Chang, R. Chen, P. Chintalapati, M. Crocce, C. Davis, H. T. Diehl, A. Drlica-Wagner, K. Eckert, T. F. Eifler, J. Elvin-Poole, S. Everett, X. Fang, P. Fosalba, O. Friedrich, E. Gaztanaga, G. Giannini, R. A. Gruendl, I. Harrison, W. G. Hartley, K. Herner, H. Huang, E. M. Huff, D. Huterer, N. Kuropatkin,

P. Leget, A. R. Liddle, J. McCullough, J. Muir, S. Pandey, Y. Park, A. Porredon, A. Refregier, R. P. Rollins, A. Roodman, R. Rosenfeld, A. J. Ross, E. S. Rykoff, J. Sanchez, I. Sevilla-Noarbe, E. Sheldon, T. Shin, A. Troja, I. Tutusaus, I. Tutusaus, T. N. Varga, N. Weaverdyck, B. Yanny, B. Yin, Y. Zhang, J. Zuntz, M. Aguena, S. Allam, J. Annis, D. Bacon, E. Bertin, S. Bhargava, D. Brooks, E. Buckley-Geer, D. L. Burke, J. Carretero, M. Costanzi, L. N. da Costa, M. E. S. Pereira, J. De Vicente, S. Desai, J. P. Dietrich, P. Doel, I. Ferrero, B. Flaugher, J. Frieman, J. García-Bellido, E. Gaztanaga, D. W. Gerdes, T. Giannantonio, J. Gschwend, G. Gutierrez, S. R. Hinton, D. L. Hollowood, K. Honscheid, B. Hoyle, D. J. James, R. Kron, K. Kuehn, O. Lahav, M. Lima, H. Lin, M. A. G. Maia, J. L. Marshall, P. Martini, P. Melchior, F. Menanteau, R. Miquel, J. J. Mohr, R. Morgan, R. L. C. Ogando, A. Palmese, F. Paz-Chinchón, D. Petravick, A. Pieres, A. K. Romer, E. Sanchez, V. Scarpine, M. Schubnell, S. Serrano, M. Smith, M. Soares-Santos, G. Tarle, D. Thomas, C. To, and J. Weller. Dark energy survey year 3 results: Cosmology from cosmic shear and robustness to data calibration. *Physical Review D*, 105(2):023514, Jan 2022. doi: 10.1103/PhysRevD.105.023514. URL https://link.aps.org/doi/10.1103/PhysRevD.105.023514.

Dhayaa Anbajagane, August E. Evrard, Arya Farahi, David J. Barnes, Klaus Dolag, Ian G. McCarthy, Dylan Nelson, and Annalisa Pillepich. Stellar property statistics of massive halos from cosmological hydrodynamics simulations: Common kernel shapes. *Monthly Notices of the Royal Astronomical Society*, 495(1):686–704, Jun 2020. ISSN 0035-8711, 1365-2966. doi: 10.1093/mnras/staa1147. URL http://arxiv.org/abs/2001.02283. arXiv: 2001.02283.

Stephane Arnouts, Lauro Moscardini, Eros Vanzella, Stefano Colombi, Stefano Cristiani, Adriano Fontana, Emanuele Giallongo, Sabino Matarrese, and P Saracco. Measuring the redshift evolution of clustering: the hubble deep field south. *Monthly Notices of the Royal Astronomical Society*, 329(2):355–366, 2002.

Felipe Avila, Armando Bernui, Rafael C Nunes, Edilson de Carvalho, and Camila P Novaes. The homogeneity scale and the growth rate of cosmic structures. *Monthly Notices of the Royal Astronomical Society*, 509(2):2994–3003, Jan 2022. ISSN 0035-8711. doi: 10.1093/mnras/stab3122. URL https://doi.org/10.1093/mnras/stab3122.

Ivan K. Baldry, Karl Glazebrook, Jon Brinkmann, Željko Ivezić, Robert H. Lupton, Robert C. Nichol, and Alexander S. Szalay. Quantifying the bimodal color-magnitude distribution of galaxies. *The Astrophysical Journal*, 600(2):681–694, jan 2004. doi: 10.1086/380092. URL https://doi.org/10.1086/380092.

Michael L. Balogh, Simon L. Morris, H. K. C. Yee, R. G. Carlberg, and Erica Ellingson. Differential galaxy evolution in cluster and field galaxies at $z \approx 0.3$. *The Astrophysical Journal*, 527(1):54, Dec 1999. ISSN 0004-637X. doi: 10.1086/308056. URL https://dx.doi.org/10.1086/308056.

Michael L. Balogh, Ivan K. Baldry, Robert Nichol, Chris Miller, Richard Bower, and Karl Glazebrook. The bimodal galaxy color distribution: Dependence on luminosity and environment. *The Astrophysical Journal*, 615(2):L101, Sep 2004. ISSN 0004-637X. doi: 10.1086/426079. URL https://iopscience.iop.org/article/10.1086/426079/meta.

Eric F. Bell, Christian Wolf, Klaus Meisenheimer, Hans-Walter Rix, Andrea Borch, Simon Dye, Martina Kleinheinrich, Lutz Wisotzki, and Daniel H. McIntosh. Nearly 5000 distant early-type galaxies in combo-17: A red sequence and its evolution since z 1. *The Astrophysical Journal*, 608(2):752, Jun 2004. ISSN 0004-637X. doi: 10.1086/420778. URL https://iopscience.iop.org/article/10.1086/420778/meta.

Suman Bhattacharya, Salman Habib, Katrin Heitmann, and Alexey Vikhlinin. Dark matter halo profiles of massive clusters: Theory versus observations. *The Astrophysical Journal*, 766:32, Mar 2013. doi: 10.1088/0004-637X/766/1/32. URL http://adsabs.harvard.edu/abs/2013ApJ...766...32B.

William K Black and August Evrard. Red dragon: a redshift-evolving gaussian mixture model for galaxies. *Monthly Notices of the Royal Astronomical Society*, 516(1):1170–1182, Oct 2022. ISSN 0035-8711. doi: 10.1093/mnras/stac2052. URL https://doi.org/10.1093/mnras/stac2052.

Michael R. Blanton. Galaxies in sdss and deep2: A quiet life on the blue sequence? *The Astrophysical Journal*, 648(1):268, Sep 2006. ISSN 0004-637X. doi: 10.1086/505628.

Michael R. Blanton and John Moustakas. Physical properties and environments of nearby galaxies. *Annual Review of Astronomy and Astrophysics*, 47(1):159–210, 2009. doi: 10.1146/annurev-astro-082708-101734. URL https://doi.org/10.1146/annurev-astro-082708-101734.

L. E. Bleem, S. Bocquet, B. Stalder, M. D. Gladders, P. A. R. Ade, S. W. Allen, A. J. Anderson, J. Annis, M. L. N. Ashby, J. E. Austermann, and et al. The sptpol extended cluster survey. *The Astrophysical Journal Supplement Series*, 247:25, Mar 2020. doi: 10.3847/1538-4365/ab6993. URL http://adsabs.harvard.edu/abs/2020ApJS..247...25B.

Lex Borghans, Bart H. H. Golsteyn, James J. Heckman, and John Eric Humphries. What grades and achievement tests measure. *Proceedings of the National Academy of Sciences*, 113(47): 13354–13359, Nov 2016. doi: 10.1073/pnas.1601135113. URL https://www.pnas.org/doi/full/10.1073/pnas.1601135113.

Alessandro Boselli, Matteo Fossati, and Ming Sun. Ram pressure stripping in high-density environments. *A&A Rev.*, 30(1):3, December 2022. doi: 10.1007/s00159-022-00140-3.

Thomas J. Bouchard. The Wilson Effect: The Increase in Heritability of IQ With Age. *Twin Research and Human Genetics*, 16(5):923–930, Oct 2013. ISSN 1832-4274, 1839-2628. doi: 10.1017/thg.2013.54. URL https://www.cambridge.org/core/journals/twin-research-and-human-genetics/article/wilson-effect-the-increase-in-heritability-of-iq-with-age/FF406CC4CF286D78AF72C9E7EF9B5E3F.

R. G. Bower, J. R. Lucey, and R. S. Ellis. Precision photometry of early type galaxies in the coma and virgo clusters - a test of the universality of the colour / magnitude relation - part two - analysis. *Monthly Notices of the Royal Astronomical Society*, 254:601, Feb 1992. ISSN 0035-8711.

doi: 10.1093/mnras/254.4.601. URL http://adsabs.harvard.edu/abs/1992MNRAS.254..601B.

Derek C. Briggs. The effect of admissions test preparation: Evidence from nels:88. *CHANCE*, 14(1):10–18, Jan 2001. ISSN 0933-2480. doi: 10.1080/09332480.2001.10542245. URL https://doi.org/10.1080/09332480.2001.10542245.

Derek C. Briggs. Preparation for college admission exams. 2009 nacac discussion paper. *National Association for College Admission Counseling*, 2009. URL https://eric.ed.gov/?id=ED505529. ERIC Number: ED505529.

Robert T. Brown, Cecil R. Reynolds, and Jean S. Whitaker. Bias in mental testing since bias in mental testing. *School Psychology Quarterly*, 14:208–238, 1999. ISSN 1939-1560. doi: 10.1037/h0089007.

G. Bruzual and S. Charlot. Stellar population synthesis at the resolution of 2003. *Monthly Notices of the Royal Astronomical Society*, 344(4):1000–1028, Oct 2003. ISSN 0035-8711. doi: 10.1046/j.1365-8711.2003.06897.x. URL https://doi.org/10.1046/j.1365-8711.2003.06897.x.

G. Bruzual A. Spectral evolution of galaxies. i. early-type systems. *The Astrophysical Journal*, 273:105–127, Oct 1983. ISSN 0004-637X. doi: 10.1086/161352. URL https://ui.adsabs.harvard.edu/abs/1983ApJ...273..105B. ADS Bibcode: 1983ApJ...273..105B.

Michael T. Busha and Risa H. Wechsler. Making mock galaxy catalogs with addgals. *43rd Rencontres de Moriond on Cosmology*, page 227–230, 2008. URL http://inspirehep.net/record/1517667/files/1422658_227-230.pdf.

H. Butcher and Jr. Oemler, A. The evolution of galaxies in clusters. i - isit photometry of c1 0024+1654 and 3c 295. *The Astrophysical Journal*, 219:18–30, Jan 1978. ISSN 0004-637X. doi: 10.1086/155751. URL http://adsabs.harvard.edu/abs/1978ApJ...219...18B.

K. I. Caputi, O. Ilbert, C. Laigle, H. J. McCracken, O. Le Fèvre, J. Fynbo, B. Milvang-Jensen, P. Capak, M. Salvato, and Y. Taniguchi. Spitzer bright, ultravista faint sources in cosmos: The contribution to the overall population of massive galaxies at z = 3–7. *The Astrophysical Journal*, 810(1):73, Sep 2015. ISSN 0004-637X. doi: 10.1088/0004-637X/810/1/73. URL https://dx.doi.org/10.1088/0004-637X/810/1/73.

J. Carretero, F. J. Castander, E. Gaztañaga, M. Crocce, and P. Fosalba. An algorithm to build mock galaxy catalogues using mice simulations. *Monthly Notices of the Royal Astronomical Society*, 447:646–670, Feb 2015. ISSN 0035-8711. doi: 10.1093/mnras/stu2402. URL http://adsabs.harvard.edu/abs/2015MNRAS.447..646C.

Daniel Ceverino and Anatoly Klypin. The role of stellar feedback in the formation of galaxies. *The Astrophysical Journal*, 695(1):292, Mar 2009. ISSN 0004-637X. doi: 10.1088/0004-637X/695/1/292. URL https://dx.doi.org/10.1088/0004-637X/695/1/292.

Benjamin P. Chapman, Paul R. Duberstein, Silvia Sörensen, and Jeffrey M. Lyness. Gender differences in five factor model personality traits in an elderly cohort. *Personality and Individual Differences*, 43(6):1594–1603, Oct 2007. ISSN 0191-8869. doi: 10.1016/j.paid.2007.04.028. URL https://www.sciencedirect.com/science/article/pii/S0191886907001663.

GLADYS Charles-Ogan. Gender influences on study habits of mathematics students' achievement. *International Journal of Academic Research and Reflection*, 3(7):24–28, 2015.

Maria Christopoulou, Agathi Lakioti, Christos Pezirkianidis, Eirini Karakasidou, and Anastassios Stalikas. The role of grit in education: A systematic review. *Psychology*, 9(1515):2951–2971, Dec 2018. doi: 10.4236/psych.2018.915171.

William S Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, 74(368):829–836, 1979.

William S Cleveland and Clive Loader. Smoothing by local regression: Principles and methods. In *Statistical theory and computational aspects of smoothing*, pages 10–49. Springer, 1996.

Andrew Collette, James Tocknell, Thomas A Caswell, Darren Dale, Ulrik Kofoed Pedersen, Aleksandar Jelenak, Andrea Bedini, Martin Raspaud, Jialin Lei, Laurence Hole, Simgon Gregor Ebner, Smutch, Matthew Zwier, Antony Lee, Matthew Brett, Joseph Kleinhenz, Jonah Bernhard, and John Tyree. h5py. *GitHub*, Mar 2017. doi: 10.5281/zenodo.400660. URL https://doi.org/10.5281/zenodo.594310.

Compass Education Group. *Comparing SAT and ACT Scores: Official New Concordance*. CollegeBoard, 2018. URL https://www.compassprep.com/concordance-and-conversion-sat-and-act-scores/.

A. J. Connolly, A. S. Szalay, Mark Dickinson, M. U. SubbaRao, and R. J. Brunner. The Evolution of the Global Star Formation History as Measured from the Hubble Deep Field. *ApJ Letters*, 486 (1):L11–L14, September 1997. doi: 10.1086/310829.

Charlie Conroy and James E. Gunn. The propagation of uncertainties in stellar population synthesis modeling. iii. model calibration, comparison, and evaluation. *The Astrophysical Journal*, 712 (2):833–857, Mar 2010. ISSN 0004-637X. doi: 10.1088/0004-637X/712/2/833. URL https://doi.org/10.1088/0004-637x/712/2/833.

Charlie Conroy, James E. Gunn, and Martin White. The propagation of uncertainties in stellar population synthesis modeling. i. the relevance of uncertain aspects of stellar evolution and the initial mass function to the derived physical properties of galaxies. *The Astrophysical Journal*, 699(1):486–506, Jun 2009. ISSN 0004-637X. doi: 10.1088/0004-637X/699/1/486. URL https://doi.org/10.1088/0004-637x/699/1/486.

Asantha Cooray and Ravi Sheth. Halo models of large scale structure. *Physics Reports*, 372(1): 1, Dec 2002. doi: 10.1016/S0370-1573(02)00276-4. URL https://ui.adsabs.harvard.edu/abs/2002PhR...372....1C/abstract.

Paul T. Costa, Antonio Terracciano, and Robert R. McCrae. Gender differences in personality traits across cultures: robust and surprising findings. *Journal of Personality and Social Psychology*, 81(2):322–331, Aug 2001. ISSN 0022-3514. doi: 10.1037/0022-3514.81.2.322.

M. Costanzi, E. Rozo, E. S. Rykoff, A. Farahi, T. Jeltema, A. E. Evrard, A. Mantz, D. Gruen, R. Mandelbaum, J. DeRose, and et al. Modelling projection effects in optically selected cluster catalogues. *Monthly Notices of the Royal Astronomical Society*, 482(1):490–505, Jan 2019. ISSN 0035-8711. doi: 10.1093/mnras/sty2665. URL https://academic.oup.com/mnras/article/482/1/490/5114581.

Thomas R. Coyle and David R. Pillow. SAT and ACT predict college GPA after removing g. *Intelligence*, 36(6):719–729, Nov 2008. ISSN 0160-2896. doi: 10.1016/j.intell.2008.05.001. URL https://www.sciencedirect.com/science/article/pii/S0160289608000603.

Marcus Credé, Michael C. Tynan, and Peter D. Harms. Much ado about grit: A meta-analytic synthesis of the grit literature. *Journal of Personality and Social Psychology*, 113(3):492–511, 2017. ISSN 1939-1315. doi: 10.1037/pspp0000102.

Darren J. Croton, Volker Springel, Simon D. M. White, G. De Lucia, C. S. Frenk, L. Gao, A. Jenkins, G. Kauffmann, J. F. Navarro, and N. Yoshida. The many lives of active galactic nuclei: cooling flows, black holes and the luminosities and colours of galaxies. *MNRAS*, 365(1):11–28, January 2006. doi: 10.1111/j.1365-2966.2005.09675.x.

Tara Dacunha, Matthew Belyakov, Susmita Adhikari, Tae-hyeon Shin, Samuel Goldstein, and Bhuvnesh Jain. Connecting galaxy evolution in clusters with their radial profiles and phase space distribution: results from the illustristng hydrodynamical simulations. *Monthly Notices of the Royal Astronomical Society*, page stac392, Feb 2022. ISSN 0035-8711. doi: 10.1093/mnras/stac392. URL https://doi.org/10.1093/mnras/stac392.

G. B. Dalton, M. Caldwell, A. K. Ward, M. S. Whalley, G. Woodhouse, R. L. Edeson, P. Clark, S. M. Beard, A. M. Gallie, S. P. Todd, J. M. D. Strachan, N. N. Bezawada, W. J. Sutherland, and J. P. Emerson. The vista infrared camera. In *Ground-based and Airborne Instrumentation for Astronomy*, volume 6269, page 314–323. SPIE, Jun 2006. doi: 10.1117/12.670018. URL https://www.spiedigitallibrary.org/conference-proceedings-of-spie/6269/62690X/The-VISTA-infrared-camera/10.1117/12.670018.full.

L J M Davies, J E Thorne, S Bellstedt, M Bravo, A S G Robotham, S P Driver, R H W Cook, L Cortese, J D'Silva, M W Grootes, B W Holwerda, A M Hopkins, M J Jarvis, C Lidman, S Phillipps, and M Siudek. Deep Extragalactic VIsible Legacy Survey (DEVILS): evolution of the σSFR–$M_\star$ relation and implications for self-regulated star formation. *Monthly Notices of the Royal Astronomical Society*, 509(3):4392–4410, 11 2021. ISSN 0035-8711. doi: 10.1093/mnras/stab3145. URL https://doi.org/10.1093/mnras/stab3145.

Liam Delaney, Colm Harmon, and Cathy Redmond. Parental education, grade attainment and earnings expectations among university students. *Economics of Education Review*, 30(6): 1136–1152, Dec 2011. ISSN 0272-7757. doi: 10.1016/j.econedurev.2011.04.004. URL https://www.sciencedirect.com/science/article/pii/S0272775711000598.

J. DeRose, R. H. Wechsler, M. R. Becker, E. S. Rykoff, S. Pandey, N. MacCrann, A. Amon, J. Myles, E. Krause, D. Gruen, and et al. Dark energy survey year 3 results: cosmology from combined galaxy clustering and lensing – validation on cosmological simulations. *arXiv e-prints*, page arXiv:2105.13547, May 2021. URL https://ui.adsabs.harvard.edu/abs/2021arXiv210513547D.

J. DeRose, R. H. Wechsler, M. R. Becker, E. S. Rykoff, S. Pandey, N. MacCrann, A. Amon, J. Myles, E. Krause, D. Gruen, B. Jain, M. A. Troxel, J. Prat, A. Alarcon, C. Sánchez, J. Blazek, M. Crocce, G. Giannini, M. Gatti, G. M. Bernstein, J. Zuntz, S. Dodelson, X. Fang, O. Friedrich, L. F. Secco, J. Elvin-Poole, A. Porredon, S. Everett, A. Choi, I. Harrison, J. Cordero, M. Rodriguez-Monroy, J. McCullough, R. Cawthon, A. Chen, O. Alves, F. Andrade-Oliveira, K. Bechtol, H. Camacho, A. Campos, A. Carnero Rosell, M. Carrasco Kind, H. T. Diehl, A. Drlica-Wagner, K. Eckert, T. F. Eifler, R. A. Gruendl, W. G. Hartley, H. Huang, E. M. Huff, N. Kuropatkin, M. Raveri, R. Rosenfeld, A. J. Ross, J. Sanchez, I. Sevilla-Noarbe, E. Sheldon, B. Yanny, B. Yin, Y. Zhang, P. Fosalba, M. Aguena, S. Allam, J. Annis, S. Avila, D. Bacon, S. Bhargava, D. Brooks, E. Buckley-Geer, D. L. Burke, J. Carretero, F. J. Castander, C. Chang, M. Costanzi, L. N. da Costa, M. E. S. Pereira, J. De Vicente, S. Desai, J. P. Dietrich, P. Doel, A. E. Evrard, I. Ferrero, A. Ferté, B. Flaugher, J. Frieman, J. García-Bellido, E. Gaztanaga, T. Giannantonio, J. Gschwend, G. Gutierrez, S. R. Hinton, D. L. Hollowood, K. Honscheid, D. Huterer, D. J. James, K. Kuehn, O. Lahav, M. Lima, M. A. G. Maia, J. L. Marshall, P. Melchior, F. Menanteau, R. Miquel, J. J. Mohr, R. Morgan, A. Palmese, F. Paz-Chinchón, D. Petravick, A. Pieres, A. A. Plazas Malagón, E. Sanchez, V. Scarpine, S. Serrano, M. Smith, M. Soares-Santos, E. Suchyta, G. Tarle, D. Thomas, C. To, T. N. Varga, and DES Collaboration. Dark Energy Survey Year 3 results: Cosmology from combined galaxy clustering and lensing validation on cosmological simulations. *Physical Review D*, 105(12):123520, June 2022. doi: 10.1103/PhysRevD.105.123520.

Joseph DeRose, Risa H. Wechsler, Matthew R. Becker, Michael T. Busha, Eli S. Rykoff, Niall MacCrann, Brandon Erickson, August E. Evrard, Andrey Kravtsov, Daniel Gruen, and et al. The buzzard flock: Dark energy survey synthetic sky catalogs. *arXiv:1901.02401 [astro-ph]*, Jan 2019. URL http://arxiv.org/abs/1901.02401.

Benedikt Diemer. Colossus: A python toolkit for cosmology, large-scale structure, and dark matter halos. *The Astrophysical Journal Supplement Series*, 239:35, Dec 2018. doi: 10.3847/1538-4365/aaee8c. URL http://adsabs.harvard.edu/abs/2018ApJS..239...35D.

Benedikt Diemer and Andrey V. Kravtsov. Dependence of the outer density profiles of halos on their mass accretion rate. *The Astrophysical Journal*, 789(1):1, Jun 2014. ISSN 0004-637X. doi: 10.1088/0004-637X/789/1/1. URL https://doi.org/10.1088/0004-637x/789/1/1.

Mamoru Doi, Masayuki Tanaka, Masataka Fukugita, James E. Gunn, Naoki Yasuda, Željko Ivezić, Jon Brinkmann, Ernst de Haars, S. J. Kleinman, Jurek Krzesinski, and et al. Photometric response functions of the sloan digital sky survey imager. *The Astronomical Journal*, 139: 1628–1648, Apr 2010. ISSN 0004-6256. doi: 10.1088/0004-6256/139/4/1628. URL https://ui.adsabs.harvard.edu/abs/2010AJ....139.1628D.

Megan Donahue and G. Mark Voit. Baryon cycles in the biggest galaxies. *Physics Reports*, 973: 1–109, August 2022. doi: 10.1016/j.physrep.2022.04.005.

Martina Donnari, Annalisa Pillepich, Dylan Nelson, Federico Marinacci, Mark Vogelsberger, and Lars Hernquist. Quenched fractions in the illustristng simulations: comparison with observations and other theoretical models. *arXiv:2008.00004 [astro-ph]*, Jul 2020. URL http://arxiv.org/abs/2008.00004. arXiv: 2008.00004.

Neil Dorans. Further comment: Freedle's table 2: Fact or fiction? *Harvard Educational Review*, 74(1):62–73, Jun 2010. ISSN 0017-8055. doi: 10.17763/haer.74.1.8729105044552127. URL https://doi.org/10.17763/haer.74.1.8729105044552127.

Fritz Drasgow. Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology*, 72:19–29, 1987. ISSN 1939-1854. doi: 10.1037/0021-9010.72.1.19.

Angela L Duckworth, Christopher Peterson, Michael D Matthews, and Dennis R Kelly. Grit: perseverance and passion for long-term goals. *Journal of personality and social psychology*, 92 (6):1087, 2007.

Angela Lee Duckworth and Patrick D Quinn. Development and validation of the short grit scale (grit–s). *Journal of personality assessment*, 91(2):166–174, 2009.

Angela Lee Duckworth, Teri A. Kirby, Eli Tsukayama, Heather Berstein, and K. Anders Ericsson. Deliberate practice spells success: Why grittier competitors triumph at the national spelling bee. *Social Psychological and Personality Science*, 2(2):174–181, Mar 2011. ISSN 1948-5506. doi: 10.1177/1948550610385872. URL https://doi.org/10.1177/1948550610385872.

James S. Dunlop. Observing the first galaxies. *Astrophysics and Space Science Library*, page 223–292, Sep 2012. ISSN 0067-0057. doi: 10.1007/978-3-642-32362-1_5. URL https://ned.ipac.caltech.edu/level5/Sept14/Dunlop/Dunlop3.html.

John Dunlosky, Katherine A. Rawson, Elizabeth J. Marsh, Mitchell J. Nathan, and Daniel T. Willingham. Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, 14(1):4–58, 2013. doi: 10.1177/1529100612453266. URL https://doi.org/10.1177/1529100612453266.

Stephen Eales, Pieter de Vis, Matthew W. L. Smith, Kiran Appah, Laure Ciesla, Chris Duffield, and Simon Schofield. The galaxy end sequence. *Monthly Notices of the Royal Astronomical Society*, 465(3):3125–3133, Mar 2017. ISSN 0035-8711. doi: 10.1093/mnras/stw2875. URL https://doi.org/10.1093/mnras/stw2875.

Stephen A Eales, Maarten Baes, Nathan Bourne, Malcolm Bremer, Michael J I Brown, Christopher Clark, David Clements, Pieter de Vis, Simon Driver, Loretta Dunne, and et al. The causes of the red sequence, the blue cloud, the green valley, and the green mountain. *Monthly Notices of the Royal Astronomical Society*, 481(1):1183–1194, Nov 2018. ISSN 0035-8711. doi: 10.1093/mnras/sty2220. URL https://doi.org/10.1093/mnras/sty2220.

J. Emerson, A. McPherson, and W. Sutherland. Visible and infrared survey telescope for astronomy: Progress report. *The Messenger*, 126:41–42, Dec 2006. ISSN 0722-6691. URL https://ui.adsabs.harvard.edu/abs/2006Msngr.126...41E. ADS Bibcode: 2006Msngr.126...41E.

Eric Ermie. Psychometrics 101: Know what your assessment data is telling you, May 2016. URL https://www.slideshare.net/examsoft/psychometrics-101-know-what-your-assessment-data-is-telling-you.

Euclid Collaboration. Euclid preparation - iii. galaxy cluster detection in the wide photometric survey, performance and algorithm selection. *A&A*, 627, Jul 2019. ISSN 0004-6361, 1432-0746. doi: 10.1051/0004-6361/201935088. URL https://www.aanda.org/articles/aa/abs/2019/07/aa35088-19/aa35088-19.html.

August Evrard, Kyle Schulz, and Caitlin Hayward. How did you get that A? Selectivity's role in rising undergraduate grades at a large public university. In *LAK21: 11th International Learning Analytics and Knowledge Conference*, LAK21, page 565–571, New York, NY, USA, Apr 2021. Association for Computing Machinery. ISBN 978-1-4503-8935-8. doi: 10.1145/3448139.3448199. URL https://doi.org/10.1145/3448139.3448199.

August E. Evrard, Pablo Arnault, Dragan Huterer, and Arya Farahi. A model for multiproperty galaxy cluster statistics. *Monthly Notices of the Royal Astronomical Society*, 441(4):3562–3569, Jul 2014. ISSN 0035-8711. doi: 10.1093/mnras/stu784. URL https://academic.oup.com/mnras/article/441/4/3562/1217975.

August E. Evrard, Michael Mills, David Winn, Kathryn Jones, Jared Tritz, and Timothy A. McKay. Problem roulette: Studying introductory physics in the cloud. *American Journal of Physics*, 83 (1):76–84, 2015. doi: 10.1119/1.4894061. URL https://doi.org/10.1119/1.4894061.

G. Fabbiano. X-rays from normal galaxies. *Annu. Rev. Astron. Astrophys*, 27:87–138, January 1989. doi: 10.1146/annurev.aa.27.090189.000511.

Arya Farahi, August E. Evrard, Ian McCarthy, David J. Barnes, and Scott T. Kay. Localized massive halo properties in bahamas and macsis simulations: scalings, lognormality, and covariance. *Monthly Notices of the Royal Astronomical Society*, 478(2):2618–2632, Aug 2018. ISSN 0035-8711. doi: 10.1093/mnras/sty1179. URL https://academic.oup.com/mnras/article/478/2/2618/4993329.

Arya Farahi, Dhayaa Anbajagane, and August E. Evrard. KLLR: A scale-dependent, multivariate model class for regression analysis. *The Astrophysical Journal*, 931(2):166, Jun 2022. ISSN 0004-637X. doi: 10.3847/1538-4357/ac6ac7. URL https://doi.org/10.3847/1538-4357/ac6ac7.

Arya Farahi, Dhayaa Anbajagane, and August E. Evrard. KLLR: A Scale-dependent, Multivariate Model Class for Regression Analysis. *ApJ*, 931(2):166, June 2022. doi: 10.3847/1538-4357/ac6ac7.

Robert Feldmann. Are star formation rates of galaxies bimodal? *Monthly Notices of the Royal Astronomical Society: Letters*, 470(1):L59–L63, Sep 2017. ISSN 1745-3925. doi: 10.1093/mnrasl/slx073. URL https://doi.org/10.1093/mnrasl/slx073.

Laura Ferrarese and David Merritt. A fundamental relation between supermassive black holes and their host galaxies. *The Astrophysical Journal*, 539(1):L9, Aug 2000. ISSN 0004-637X. doi: 10.1086/312838. URL https://dx.doi.org/10.1086/312838.

B. Flaugher, H. T. Diehl, K. Honscheid, T. M. C. Abbott, O. Alvarez, R. Angstadt, J. T. Annis, M. Antonik, O. Ballester, L. Beaufore, G. M. Bernstein, R. A. Bernstein, B. Bigelow, M. Bonati, D. Boprie, D. Brooks, E. J. Buckley-Geer, J. Campa, L. Cardiel-Sas, F. J. Castander, J. Castilla, H. Cease, J. M. Cela-Ruiz, S. Chappa, E. Chi, C. Cooper, L. N. da Costa, E. Dede, G. Derylo, D. L. DePoy, J. de Vicente, P. Doel, A. Drlica-Wagner, J. Eiting, A. E. Elliott, J. Emes, J. Estrada, A. Fausti Neto, D. A. Finley, R. Flores, J. Frieman, D. Gerdes, M. D. Gladders, B. Gregory, G. R. Gutierrez, J. Hao, S. E. Holland, S. Holm, D. Huffman, C. Jackson, D. J. James, M. Jonas, A. Karcher, I. Karliner, S. Kent, R. Kessler, M. Kozlovsky, R. G. Kron, D. Kubik, K. Kuehn, S. Kuhlmann, K. Kuk, O. Lahav, A. Lathrop, J. Lee, M. E. Levi, P. Lewis, T. S. Li, I. Mandrichenko, J. L. Marshall, G. Martinez, K. W. Merritt, R. Miquel, F. Muñoz, E. H. Neilsen, R. C. Nichol, B. Nord, R. Ogando, J. Olsen, N. Palaio, K. Patton, J. Peoples, A. A. Plazas, J. Rauch, K. Reil, J.-P. Rheault, N. A. Roe, H. Rogers, A. Roodman, E. Sanchez, V. Scarpine, R. H. Schindler, R. Schmidt, R. Schmitt, M. Schubnell, K. Schultz, P. Schurter, L. Scott, S. Serrano, T. M. Shaw, R. C. Smith, M. Soares-Santos, A. Stefanik, W. Stuermer, E. Suchyta, A. Sypniewski, G. Tarle, J. Thaler, R. Tighe, C. Tran, D. Tucker, A. R. Walker, G. Wang, M. Watson, C. Weaverdyck, W. Wester, R. Woods, B. Yanny, and (The DES Collaboration). The dark energy camera. *The Astronomical Journal*, 150(5):150, Oct 2015. ISSN 1538-3881. doi: 10.1088/0004-6256/150/5/150. URL https://dx.doi.org/10.1088/0004-6256/150/5/150.

James M. Fraser, Anneke L. Timan, Kelly Miller, Jason E. Dowd, Laura Tucker, and Eric Mazur. Teaching and physics education research: bridging the gap. *Reports on Progress in Physics*, 77(3):032401, Mar 2014. ISSN 0034-4885. doi: 10.1088/0034-4885/77/3/032401. URL https://dx.doi.org/10.1088/0034-4885/77/3/032401.

Meredith C. Frey and Douglas K. Detterman. Scholastic assessment or g?: The relationship between the scholastic assessment test and general cognitive ability. *Psychological Science*, 15(6):373–378, Jun 2004. ISSN 0956-7976. doi: 10.1111/j.0956-7976.2004.00687.x. URL https://doi.org/10.1111/j.0956-7976.2004.00687.x.

Joshua A. Frieman, Michael S. Turner, and Dragan Huterer. Dark energy and the accelerating universe. *Annual Review of Astronomy and Astrophysics*, 46(1):385–432, 2008. doi: 10.1146/annurev.astro.46.060407.145243. URL https://doi.org/10.1146/annurev.astro.46.060407.145243.

Saul Geiser. SAT/ACT scores, high-school GPA, and the problem of omitted variable bias: Why the UC taskforce's findings are spurious. *Research & Occasional Paper Series: CSHE.1.2020*, Mar 2020. URL https://eric.ed.gov/?id=ED606658. ERIC Number: ED606658.

Michael D. Gladders and H. K. C. Yee. A new method for galaxy cluster detection. i. the algorithm. *The Astronomical Journal*, 120(4):2148–2162, Oct 2000. ISSN 1538-3881. doi: 10.1086/301557. URL https://doi.org/10.1086/301557.

Michael D. Gladders, Omar López-Cruz, H. K. C. Yee, and Tadayuki Kodama. The slope of the cluster elliptical red sequence: A probe of cluster evolution. *The Astrophysical Journal*, 501(2):571, Jul 1998. ISSN 0004-637X. doi: 10.1086/305858. URL https://iopscience.iop.org/article/10.1086/305858/meta.

R. Gobat, E. Daddi, M. Onodera, A. Finoguenov, A. Renzini, N. Arimoto, R. Bouwens, M. Brusa, R.-R. Chary, A. Cimatti, and et al. A mature cluster with x-ray emission at z = 2.07. *Astronomy & Astrophysics*, 526:A133, Feb 2011. ISSN 0004-6361, 1432-0746. doi: 10.1051/0004-6361/201016084. URL https://www.aanda.org/articles/aa/abs/2011/02/aa16084-10/aa16084-10.html.

Emmet Golden-Marx, E. L. Blanton, R. Paterno-Mahler, M. Brodwin, M. L. N. Ashby, E. Moravec, L. Shen, B. C. Lemaux, L. M. Lubin, R. R. Gal, and et al. The high-redshift clusters occupied by bent radio agn (cobra) survey: Radio source properties. *The Astrophysical Journal*, 907(2): 65, Jan 2021. ISSN 0004-637X. doi: 10.3847/1538-4357/abcd96. URL https://doi.org/10.3847/1538-4357/abcd96.

Andrew Grodner and Nicholas G. Rupp. The role of homework in student learning outcomes: Evidence from a field experiment. *The Journal of Economic Education*, 44(2):93–109, 2013. doi: 10.1080/00220485.2013.770334. URL https://doi.org/10.1080/00220485.2013.770334.

Krisandra Guidry. Delivery versus time devoted to assignments: The effect on course performance. *Journal of Instructional Pedagogies*, 19, 2017.

Carl J Hansen, Steven D Kawaler, and Virginia Trimble. *Stellar interiors: physical principles, structure, and evolution*. Springer Science & Business Media, 2012.

Sarah M. Hansen, Timothy A. McKay, Risa H. Wechsler, James Annis, Erin Scott Sheldon, and Amy Kimball. Measurement of galaxy cluster sizes, radial profiles, and luminosity functions from sdss photometric data. *The Astrophysical Journal*, 633(1):122–137, Nov 2005. ISSN 0004-637X. doi: 10.1086/444554. URL https://doi.org/10.1086/444554.

Sarah M. Hansen, Erin S. Sheldon, Risa H. Wechsler, and Benjamin P. Koester. The galaxy content of sdss clusters and groups. *The Astrophysical Journal*, 699(2):1333–1353, Jun 2009. ISSN 0004-637X. doi: 10.1088/0004-637X/699/2/1333. URL https://doi.org/10.1088%2F0004-637x%2F699%2F2%2F1333.

Jiangang Hao, Benjamin P. Koester, Timothy A. Mckay, Eli S. Rykoff, Eduardo Rozo, August Evrard, James Annis, Matthew Becker, Michael Busha, David Gerdes, and et al. Precision measurements of the cluster red sequence using an error-corrected gaussian mixture model. *The Astrophysical Journal*, 702(1):745–758, Aug 2009. ISSN 0004-637X. doi: 10.1088/0004-637X/702/1/745.

W G Hartley, A Choi, A Amon, R A Gruendl, E Sheldon, I Harrison, G M Bernstein, I Sevilla-Noarbe, B Yanny, K Eckert, H T Diehl, A Alarcon, M Banerji, K Bechtol, R Buchs, S Cantu, C Conselice, J Cordero, C Davis, T M Davis, S Dodelson, A Drlica-Wagner, S Everett, A Ferté, D Gruen, K Honscheid, M Jarvis, M D Johnson, N Kokron, N MacCrann, J Myles, A B Pace, A Palmese, F Paz-Chinchón, M E S Pereira, A A Plazas, J Prat, M Rodriguez-Monroy, E S Rykoff, S Samuroff, C Sánchez, L F Secco, F Tarsitano, A Tong, M A Troxel, Z Vasquez, K Wang, C Zhou, T M C Abbott, M Aguena, S Allam, J Annis, D Bacon, E Bertin, S Bhargava, D Brooks, D L Burke, A Carnero Rosell, M Carrasco Kind, J Carretero, F J Castander, M Costanzi, M Crocce, L N da Costa, J De Vicente, J DeRose, S Desai, J P Dietrich, T F Eifler, J Elvin-Poole, I Ferrero, B Flaugher, P Fosalba, J García-Bellido, E Gaztanaga, D W Gerdes, J Gschwend, G Gutierrez, S R Hinton, D L Hollowood, D Huterer, D J James, S Kent, E Krause, K Kuehn, N Kuropatkin, O Lahav, H Lin, M A G Maia, M March, J L Marshall, P Martini, P Melchior, F Menanteau, R Miquel, J J Mohr, R Morgan, E Neilsen, R L C Ogando, S Pandey, A K Romer, A Roodman, M Sako, E Sanchez, V Scarpine, S Serrano, M Smith, M Soares-Santos, E Suchyta, M E C Swanson, G Tarle, D Thomas, C To, T N Varga, A R Walker, W Wester, R D Wilkinson, J Zuntz, and (DES Collaboration). Dark energy survey year 3 results: Deep field optical + near-infrared images and catalogue. *Monthly Notices of the Royal Astronomical Society*, 509(3):3547–3579, Jan 2022. ISSN 0035-8711. doi: 10.1093/mnras/stab3055. URL https://doi.org/10.1093/mnras/stab3055.

Marissa K. Hartwig and Eric D. Malain. Do students space their course study? those who do earn higher grades. *Learning and Instruction*, 77:101538, Feb 2022. ISSN 0959-4752. doi: 10.1016/j.learninstruc.2021.101538. URL https://www.sciencedirect.com/science/article/pii/S0959475221000979.

G. Hasinger, N. Cappelluti, H. Brunner, M. Brusa, A. Comastri, M. Elvis, A. Finoguenov, F. Fiore, A. Franceschini, R. Gilli, R. E. Griffiths, I. Lehmann, V. Mainieri, G. Matt, I. Matute, T. Miyaji, S. Molendi, S. Paltani, D. B. Sanders, N. Scoville, L. Tresse, C. M. Urry, P. Vettolani, and G. Zamorani. The xmm-newton wide-field survey in the cosmos field. i. survey description. *The Astrophysical Journal Supplement Series*, 172(1):29, Sep 2007. ISSN 0067-0049. doi: 10.1086/516576. URL https://iopscience.iop.org/article/10.1086/516576/meta.

N. P. Hathi, I. Ferreras, A. Pasquali, S. Malhotra, J. E. Rhoads, N. Pirzkal, R. A. Windhorst, and C. Xu. Stellar populations of late-type bulges at $z \simeq 1$ in the hubble ultra deep field. *The Astrophysical Journal*, 690(2):1866, Dec 2008. ISSN 0004-637X. doi: 10.1088/0004-637X/690/2/1866. URL https://dx.doi.org/10.1088/0004-637X/690/2/1866.

Andrew Hearin, Danila Korytov, Eve Kovacs, Andrew Benson, Han Aung, Christopher Bradshaw, Duncan Campbell, and LSST Dark Energy Science Collaboration. Generating synthetic cosmological data with galsampler. *Monthly Notices of the Royal Astronomical Society*, 495:5040–5051, Jun 2020. ISSN 0035-8711. doi: 10.1093/mnras/staa1495. URL http://adsabs.harvard.edu/abs/2020MNRAS.495.5040H.

Charles Henderson, Mark Connolly, Erin L. Dolan, Noah Finkelstein, Scott Franklin, Shirley Malcom, Chris Rasmussen, Kacy Redd, and Kristen St. John. Towards the stem dber alliance: Why we need a discipline-based, stem-education research community. *Journal of Geoscience Education*, 65(3):215–218, Aug 2017. ISSN 1089-9995. doi: 10.5408/1089-9995-65.3.215. URL https://doi.org/10.5408/1089-9995-65.3.215.

Richard J. Herrnstein and Charles Murray. The bell curve: Intelligence and class structure in american life. *The Free Press*, 1994.

Daniel M. Higgins, Jordan B. Peterson, Robert O. Pihl, and Alice G. M. Lee. Prefrontal cognitive ability, intelligence, big five personality, and the prediction of advanced academic and workplace performance. *Journal of Personality and Social Psychology*, 93(2):298–319, 2007. ISSN 1939-1315. doi: 10.1037/0022-3514.93.2.298.

David W. Hogg, Michael R. Blanton, Jarle Brinchmann, Daniel J. Eisenstein, David J. Schlegel, James E. Gunn, Timothy A. McKay, Hans-Walter Rix, Neta A. Bahcall, J. Brinkmann, and Avery Meiksin. The Dependence on Environment of the Color-Magnitude Relation of Galaxies. *ApJ Letters*, 601(1):L29–L32, January 2004. doi: 10.1086/381749.

Philip F Hopkins, Michael Y Grudić, Andrew Wetzel, Dušan Kereš, Claude-André Faucher-Giguère, Xiangcheng Ma, Norman Murray, and Nathan Butcher. Radiative stellar feedback in galaxy formation: Methods and physics. *Monthly Notices of the Royal Astronomical Society*, 491(3):3702–3729, Jan 2020. ISSN 0035-8711. doi: 10.1093/mnras/stz3129. URL https://doi.org/10.1093/mnras/stz3129.

Amy Hsin and Yu Xie. Explaining asian americans' academic advantage over whites. *Proceedings of the National Academy of Sciences*, 111(23):8416–8421, 2014.

Wayne Hu and Andrey V. Kravtsov. Sample variance considerations for cluster surveys. *The Astrophysical Journal*, 584(2):702–715, Feb 2003. ISSN 0004-637X. doi: 10.1086/345846. URL https://doi.org/10.1086%2F345846.

Madeline Huberth, Patricia Chen, Jared Tritz, and Timothy A. McKay. Computer-tailored student support in introductory physics. *PLOS ONE*, 10(9), Sep 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0137001. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0137001.

John D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science Engineering*, 9(3):90–95, 2007. doi: 10.1109/MCSE.2007.55.

Dragan Huterer. *A Course in Cosmology: From Theory to Practice*. Cambridge University Press, 2023. doi: 10.1017/9781009070232.

O. Ilbert, H. J. McCracken, O. Le Fèvre, P. Capak, J. Dunlop, A. Karim, M. A. Renzini, K. Caputi, S. Boissier, S. Arnouts, H. Aussel, J. Comparat, Q. Guo, P. Hudelot, J. Kartaltepe, J. P. Kneib, J. K. Krogager, E. Le Floc'h, S. Lilly, Y. Mellier, B. Milvang-Jensen, T. Moutard, M. Onodera, J. Richard, M. Salvato, D. B. Sanders, N. Scoville, J. D. Silverman, Y. Taniguchi, L. Tasca, R. Thomas, S. Toft, L. Tresse, D. Vergani, M. Wolk, and A. Zirm. Mass assembly in quiescent and star-forming galaxies since $z \simeq 4$ from ultravista. *Astronomy and Astrophysics*, 556:A55, Aug 2013. ISSN 0004-6361, 1432-0746. doi: 10.1051/0004-6361/201321100. URL https://www.aanda.org/articles/aa/abs/2013/08/aa21100-13/aa21100-13.html.

Olivier Ilbert, S Arnouts, Henry J Mccracken, M Bolzonella, Emmanuel Bertin, Olivier Le Fèvre, Yannick Mellier, G Zamorani, R Pello, Angela Iovino, et al. Accurate photometric redshifts for the cfht legacy survey calibrated using the vimos vlt deep survey. *Astronomy & Astrophysics*, 457(3):841–856, 2006.

W. Ishibashi and A. C. Fabian. Active galactic nucleus feedback and triggering of star formation in galaxies. *Monthly Notices of the Royal Astronomical Society*, 427(4):2998–3005, Dec 2012. ISSN 0035-8711. doi: 10.1111/j.1365-2966.2012.22074.x. URL https://doi.org/10.1111/j.1365-2966.2012.22074.x.

Tomoaki Ishiyama, Francisco Prada, Anatoly A Klypin, Manodeep Sinha, R Benton Metcalf, Eric Jullo, Bruno Altieri, Sofía A Cora, Darren Croton, Sylvain de la Torre, and et al. The uchuu simulations: Data release 1 and dark matter halo concentrations. *Monthly Notices of the Royal Astronomical Society*, 506(3):4210–4231, Sep 2021. ISSN 0035-8711. doi: 10.1093/mnras/stab1755. URL https://doi.org/10.1093/mnras/stab1755.

Arthur R. Jensen. Bias in mental testing. *The Free Press*, 1980.

Arthur R Jensen. The g factor: The science of mental ability. *Westport, CT: Prager*, 1998.

N. Kaiser. On the spatial correlations of abell clusters. *The Astrophysical Journal*, 284:L9, Sep 1984. ISSN 0004-637X, 1538-4357. doi: 10.1086/184341. URL http://adsabs.harvard.edu/doi/10.1086/184341.

V. Kalinova, D. Colombo, S. F. Sánchez, E. Rosolowsky, K. Kodaira, R. García-Benito, S. E. Meidt, T. A. Davis, A. B. Romeo, S.-Y. Yu, R. González Delgado, and E. a. D. Lacerda. Investigating the link between inner gravitational potential and star-formation quenching in califa galaxies. *Astronomy & Astrophysics*, 665, Sep 2022. ISSN 0004-6361, 1432-0746. doi: 10.1051/0004-6361/202243541. URL https://www.aanda.org/articles/aa/abs/2022/09/aa43541-22/aa43541-22.html.

Simona C. Kaplan, Cheri A. Levinson, Thomas L. Rodebaugh, Andrew Menatti, and Justin W. Weeks. Social anxiety and the big five personality traits: The interactive relationship of trust and openness. *Cognitive Behaviour Therapy*, 44(3):212–222, May 2015. ISSN 1650-6073. doi: 10.1080/16506073.2015.1008032. URL https://doi.org/10.1080/16506073.2015.1008032.

Brandon C. Kelly. Some aspects of measurement error in linear regression of astronomical data. *The Astrophysical Journal*, 665(2):1489–1506, Aug 2007. ISSN 0004-637X, 1538-4357. doi: 10.1086/519947. URL http://arxiv.org/abs/0705.2774. arXiv: 0705.2774.

Keunho Kim, Sangeeta Malhotra, James E. Rhoads, Bhavin Joshi, Ignacio Fererras, and Anna Pasquali. Galaxy structure, stellar populations, and star formation quenching at $0.6 \lesssim z \lesssim 1.2$. *The Astrophysical Journal*, 867(2):118, Nov 2018. ISSN 0004-637X. doi: 10.3847/1538-4357/aae488. URL https://dx.doi.org/10.3847/1538-4357/aae488.

M Klein, J J Mohr, S Desai, H Israel, S Allam, A Benoit-Lévy, D Brooks, E Buckley-Geer, A Carnero Rosell, M Carrasco Kind, C E Cunha, L N da Costa, J P Dietrich, T F Eifler, A E Evrard, J Frieman, D Gruen, R A Gruendl, G Gutierrez, K Honscheid, D J James, K Kuehn, M Lima, M A G Maia, M March, P Melchior, F Menanteau, R Miquel, A A Plazas, K Reil, A K Romer, E Sanchez, B Santiago, V Scarpine, M Schubnell, I Sevilla-Noarbe, M Smith, M Soares-Santos, F Sobreira, E Suchyta, M E C Swanson, G Tarle, and the DES Collaboration. A multicomponent matched filter cluster confirmation tool for erosita: initial application to the rass and des-sv data sets. *Monthly Notices of the Royal Astronomical Society*, 474(3):3324–3343, Mar 2018. ISSN 0035-8711. doi: 10.1093/mnras/stx2929. URL https://doi.org/10.1093/mnras/stx2929.

M Klein, S Grandis, J J Mohr, M Paulus, T M C Abbott, J Annis, S Avila, E Bertin, D Brooks, E Buckley-Geer, A Carnero Rosell, M Carrasco Kind, J Carretero, F J Castander, C E Cunha, C B D'Andrea, L N da Costa, J De Vicente, S Desai, H T Diehl, J P Dietrich, P Doel, A E Evrard, B Flaugher, P Fosalba, J Frieman, J García-Bellido, E Gaztanaga, P A Giles, D Gruen, R A Gruendl, J Gschwend, G Gutierrez, W G Hartley, D L Hollowood, K Honscheid, B Hoyle, D J James, T Jeltema, K Kuehn, N Kuropatkin, M Lima, M A G Maia, M March, J L Marshall, F Menanteau, R Miquel, R L C Ogando, A A Plazas, A K Romer, A Roodman, E Sanchez, V Scarpine, R Schindler, S Serrano, I Sevilla-Noarbe, M Smith, R C Smith, M Soares-Santos, F Sobreira, E Suchyta, M E C Swanson, G Tarle, D Thomas, V Vikram, and the DES Collaboration. A new rass galaxy cluster catalogue with low contamination extending to $z \sim 1$ in the des overlap region. *Monthly Notices of the Royal Astronomical Society*, 488(1):739–769, Sep 2019. ISSN 0035-8711. doi: 10.1093/mnras/stz1463. URL https://doi.org/10.1093/mnras/stz1463.

Tadayuki Kodama and Nobuo Arimoto. Origin of the colour-magnitude relation of elliptical galaxies. *arXiv:astro-ph/9609160*, Sep 1996. URL http://arxiv.org/abs/astro-ph/9609160. arXiv: astro-ph/9609160.

Katherine A. Koenig, Meredith C. Frey, and Douglas K. Detterman. Act and general cognitive ability. *Intelligence*, 36(2):153–160, Mar 2008. ISSN 0160-2896. doi: 10.1016/j.intell.2007.03.005. URL https://www.sciencedirect.com/science/article/pii/S0160289607000487.

Benjamin P. Koester, Timothy A. McKay, James Annis, Risa H. Wechsler, August E. Evrard, Eduardo Rozo, Lindsey Bleem, Erin S. Sheldon, and David Johnston. Maxbcg: A red-sequence galaxy cluster finder. *The Astrophysical Journal*, 660:221–238, May 2007. ISSN 0004-637X. doi: 10.1086/512092. URL http://adsabs.harvard.edu/abs/2007ApJ...660..221K.

Benjamin P. Koester, Galina Grom, and Timothy A. McKay. Patterns of gendered performance difference in introductory stem courses, 2016. URL https://arxiv.org/abs/1608.07565.

Tutku Kolcu, Jacob P Crossett, Callum Bellhouse, and Sean McGee. Quantifying the role of ram-pressure stripping of galaxies within galaxy groups. *Monthly Notices of the Royal Astronomical Society*, 515(4):5877–5893, Oct 2022. ISSN 0035-8711. doi: 10.1093/mnras/stac2177. URL https://doi.org/10.1093/mnras/stac2177.

Meera Komarraju, Steven J. Karau, Ronald R. Schmeck, and Alen Avdic. The big five personality traits, learning styles, and academic achievement. *Personality and Individual Differences*, 51 (4):472–477, Sep 2011. ISSN 0191-8869. doi: 10.1016/j.paid.2011.04.019. URL https://www.sciencedirect.com/science/article/pii/S0191886911002194.

Danila Korytov, Andrew Hearin, Eve Kovacs, Patricia Larsen, Esteban Rangel, Joseph Hollowed, Andrew J. Benson, Katrin Heitmann, Yao-Yuan Mao, Anita Bahmanyar, Chihway Chang, Duncan Campbell, Joseph DeRose, Hal Finkel, Nicholas Frontiere, Eric Gawiser, Salman Habib, Benjamin Joachimi, François Lanusse, Nan Li, Rachel Mandelbaum, Christopher Morrison, Jeffrey A. Newman, Adrian Pope, Eli Rykoff, Melanie Simet, Chun-Hao To, Vinu Vikraman, Risa H. Wechsler, Martin White, and (The LSST Dark Energy Science Collaboration). Cosmodc2: A synthetic sky catalog for dark energy science with lsst. *The Astrophysical Journal Supplement Series*, 245(2):26, Dec 2019. ISSN 0067-0049. doi: 10.3847/1538-4365/ab510c.

Yaman Köseoglu. To what extent can the Big Five and learning styles predict academic achievement. *Journal of Education and Practice*, 7(30):43–51, 2016.

Mariska Kriek, Arjen van der Wel, Pieter G. van Dokkum, Marijn Franx, and Garth D. Illingworth. The detection of a red sequence of massive field galaxies at z 2.3 and its evolution to z 0*. *The Astrophysical Journal*, 682(2):896, Aug 2008. ISSN 0004-637X. doi: 10.1086/589677. URL https://iopscience.iop.org/article/10.1086/589677/meta.

Janusz Krywult and Agnieszka Pollo. Structure and evolution of galaxies at $z \sim 1$. In *Proceedings of the Polish Astronomical Society*, volume 7, page 245–251, Zielona Góra, Poland, Aug 2018. ISBN 978-83-950430-0-0. URL https://ui.adsabs.harvard.edu/abs/2018pas7.conf..245K. ADS Bibcode: 2018pas7.conf..245K.

M. Lacy, J. A. Surace, D. Farrah, K. Nyland, J. Afonso, W. N. Brandt, D. L. Clements, C. D. P. Lagos, C. Maraston, J. Pforr, A. Sajina, M. Sako, M. Vaccari, G. Wilson, D. R. Ballantyne, W. A. Barkhouse, R. Brunner, R. Cane, T. E. Clarke, M. Cooper, A. Cooray, G. Covone, C. D'Andrea, A. E. Evrard, H. C. Ferguson, J. Frieman, V. Gonzalez-Perez, R. Gupta, E. Hatziminaoglou, J. Huang, P. Jagannathan, M. J. Jarvis, K. M. Jones, A. Kimball, C. Lidman, L. Lubin, L. Marchetti, P. Martini, R. G. McMahon, S. Mei, H. Messias, E. J. Murphy, J. A. Newman, R. Nichol, R. P. Norris, S. Oliver, I. Perez-Fournon, W. M. Peters, M. Pierre, E. Polisensky, G. T. Richards, S. E. Ridgway, H. J. A. Röttgering, N. Seymour, R. Shirley, R. Somerville, M. A. Strauss, N. Suntzeff, P. A. Thorman, E. van Kampen, A. Verma, R. Wechsler, and W. M. Wood-Vasey. A Spitzer survey of Deep Drilling Fields to be targeted by the Vera C. Rubin Observatory Legacy Survey of Space and Time. *MNRAS*, 501(1):892–910, February 2021. doi: 10.1093/mnras/staa3714.

C. Laigle, H. J. McCracken, O. Ilbert, B. C. Hsieh, I. Davidzon, P. Capak, G. Hasinger, J. D. Silverman, C. Pichon, J. Coupon, H. Aussel, D. Le Borgne, K. Caputi, P. Cassata, Y.-Y. Chang, F. Civano, J. Dunlop, J. Fynbo, J. S. Kartaltepe, A. Koekemoer, O. Le Fèvre, E. Le Floc'h, A. Leauthaud, S. Lilly, L. Lin, S. Marchesi, B. Milvang-Jensen, M. Salvato, D. B. Sanders, N. Scoville, V. Smolcic, M. Stockmann, Y. Taniguchi, L. Tasca, S. Toft, Mattia Vaccari, and J. Zabl. The cosmos2015 catalog: Exploring the 1 ¡ z ¡ 6 universe with half a million galaxies. *The Astrophysical Journal Supplement Series*, 224(2):24, Jun 2016. ISSN 0067-0049. doi: 10.3847/0067-0049/224/2/24. URL https://dx.doi.org/10.3847/0067-0049/224/2/24.

C Laigle, C Pichon, S Arnouts, H J McCracken, Y Dubois, J Devriendt, A Slyz, D Le Borgne, A Benoit-Lévy, Ho Seong Hwang, O Ilbert, K Kraljic, N Malavasi, Changbom Park, and D Vibert. Cosmos2015 photometric redshifts probe the impact of filaments on galaxy properties. *Monthly Notices of the Royal Astronomical Society*, 474(4):5437–5458, Mar 2018. ISSN 0035-8711. doi: 10.1093/mnras/stx3055. URL https://doi.org/10.1093/mnras/stx3055.

Dustin Lang, David W. Hogg, and David J. Schlegel. WISE Photometry for 400 Million SDSS Sources. *The Astronomical Journal*, 151(2):36, February 2016. doi: 10.3847/0004-6256/151/2/36.

R. Laureijs, J. Amiaux, S. Arduini, J. L. Auguères, J. Brinchmann, R. Cole, M. Cropper, C. Dabin, L. Duvet, A. Ealet, B. Garilli, P. Gondoin, L. Guzzo, J. Hoar, H. Hoekstra, R. Holmes, T. Kitching,

T. Maciaszek, Y. Mellier, F. Pasian, W. Percival, J. Rhodes, G. Saavedra Criado, M. Sauvage, R. Scaramella, L. Valenziano, S. Warren, R. Bender, F. Castander, A. Cimatti, O. Le Fèvre, H. Kurki-Suonio, M. Levi, P. Lilje, G. Meylan, R. Nichol, K. Pedersen, V. Popa, R. Rebolo Lopez, H. W. Rix, H. Rottgering, W. Zeilinger, F. Grupp, P. Hudelot, R. Massey, M. Meneghetti, L. Miller, S. Paltani, S. Paulin-Henriksson, S. Pires, C. Saxton, T. Schrabback, G. Seidel, J. Walsh, N. Aghanim, L. Amendola, J. Bartlett, C. Baccigalupi, J. P. Beaulieu, K. Benabed, J. G. Cuby, D. Elbaz, P. Fosalba, G. Gavazzi, A. Helmi, I. Hook, M. Irwin, J. P. Kneib, M. Kunz, F. Mannucci, L. Moscardini, C. Tao, R. Teyssier, J. Weller, G. Zamorani, M. R. Zapatero Osorio, O. Boulade, J. J. Foumond, A. Di Giorgio, P. Guttridge, A. James, M. Kemp, J. Martignac, A. Spencer, D. Walton, T. Blümchen, C. Bonoli, F. Bortoletto, C. Cerna, L. Corcione, C. Fabron, K. Jahnke, S. Ligori, F. Madrid, L. Martin, G. Morgante, T. Pamplona, E. Prieto, M. Riva, R. Toledo, M. Trifoglio, F. Zerbi, F. Abdalla, M. Douspis, C. Grenet, S. Borgani, R. Bouwens, F. Courbin, J. M. Delouis, P. Dubath, A. Fontana, M. Frailis, A. Grazian, J. Koppenhöfer, O. Mansutti, M. Melchior, M. Mignoli, J. Mohr, C. Neissner, K. Noddle, M. Poncet, M. Scodeggio, S. Serrano, N. Shane, J. L. Starck, C. Surace, A. Taylor, G. Verdoes-Kleijn, C. Vuerli, O. R. Williams, A. Zacchei, B. Altieri, I. Escudero Sanz, R. Kohley, T. Oosterbroek, P. Astier, D. Bacon, S. Bardelli, C. Baugh, F. Bellagamba, C. Benoist, D. Bianchi, A. Biviano, E. Branchini, C. Carbone, V. Cardone, D. Clements, S. Colombi, C. Conselice, G. Cresci, N. Deacon, J. Dunlop, C. Fedeli, F. Fontanot, P. Franzetti, C. Giocoli, J. Garcia-Bellido, J. Gow, A. Heavens, P. Hewett, C. Heymans, A. Holland, Z. Huang, O. Ilbert, B. Joachimi, E. Jennins, E. Kerins, A. Kiessling, D. Kirk, R. Kotak, O. Krause, O. Lahav, F. van Leeuwen, J. Lesgourgues, M. Lombardi, M. Magliocchetti, K. Maguire, E. Majerotto, R. Maoli, F. Marulli, S. Maurogordato, H. McCracken, R. McLure, A. Melchiorri, A. Merson, M. Moresco, M. Nonino, P. Norberg, J. Peacock, R. Pello, M. Penny, V. Pettorino, C. Di Porto, L. Pozzetti, C. Quercellini, M. Radovich, A. Rassat, N. Roche, S. Ronayette, E. Rossetti, B. Sartoris, P. Schneider, E. Semboloni, S. Serjeant, F. Simpson, C. Skordis, G. Smadja, S. Smartt, P. Spano, S. Spiro, M. Sullivan, A. Tilquin, R. Trotta, L. Verde, Y. Wang, G. Williger, G. Zhao, J. Zoubian, and E. Zucca. Euclid Definition Study Report. *arXiv e-prints*, art. arXiv:1110.3193, October 2011.

Joel Leja, Joshua S. Speagle, Yuan-Sen Ting, Benjamin D. Johnson, Charlie Conroy, Katherine E. Whitaker, Erica J. Nelson, Pieter van Dokkum, and Marijn Franx. A new census of the 0.2 ¡ z ¡ 3.0 universe. ii. the star-forming sequence. *The Astrophysical Journal*, 936(2):165, Sep 2022. ISSN 0004-637X. doi: 10.3847/1538-4357/ac887d. URL https://dx.doi.org/10.3847/1538-4357/ac887d.

Ya-Ping Li, Feng Yuan, Houjun Mo, Doosoo Yoon, Zhaoming Gan, Luis C. Ho, Bo Wang, Jeremiah P. Ostriker, and Luca Ciotti. Stellar and agn feedback in isolated early-type galaxies: The role in regulating star formation and ism properties. *The Astrophysical Journal*, 866(1):70, Oct 2018. ISSN 0004-637X. doi: 10.3847/1538-4357/aade8b. URL https://dx.doi.org/10.3847/1538-4357/aade8b.

Yen-Ting Lin, Joseph J. Mohr, and S. Adam Stanford. K-band properties of galaxy clusters and groups: Luminosity function, radial distribution, and halo occupation number. *The Astrophysical Journal*, 610:745–761, Aug 2004. ISSN 0004-637X. doi: 10.1086/421714. URL http://adsabs.harvard.edu/abs/2004ApJ...610..745L.

Yen-Ting Lin, Bau-Ching Hsieh, Sheng-Chieh Lin, Masamune Oguri, Kai-Feng Chen, Masayuki Tanaka, I.-Non Chiu, Song Huang, Tadayuki Kodama, Alexie Leauthaud, and et al. First results on the cluster galaxy population from the subaru hyper suprime-cam survey. iii. brightest cluster galaxies, stellar mass distribution, and active galaxies. *The Astrophysical Journal*, 851(2):139, Dec 2017. ISSN 1538-4357. doi: 10.3847/1538-4357/aa9bf5. URL http://arxiv.org/abs/1709.04484. arXiv: 1709.04484.

Jason M Lindo, Isaac D Swensen, and Glen R Waddell. Are big-time sports a threat to student achievement? *American Economic Journal: Applied Economics*, 4(4):254–274, 2012.

Steven Lonn and Benjamin Koester. Rearchitecting data for researchers: A collaborative model for enabling institutional learning analytics in higher education. *Journal of Learning Analytics*, 6(22):107–119, Jul 2019. ISSN 1929-7750. doi: 10.18608/jla.2019.62.8. URL https://learning-analytics.info/index.php/JLA/article/view/6201.

Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, Nov 2017. URL http://arxiv.org/abs/1705.07874. arXiv:1705.07874 [cs, stat].

Piero Madau and Mark Dickinson. Cosmic star formation history. *Annual Review of Astronomy and Astrophysics*, 52(1):415–486, Aug 2014. ISSN 0066-4146, 1545-4282. doi: 10.1146/annurev-astro-081811-125615. URL http://arxiv.org/abs/1403.0007. arXiv: 1403.0007.

Piero Madau, Henry C. Ferguson, Mark E. Dickinson, Mauro Giavalisco, Charles C. Steidel, and Andrew Fruchter. High-redshift galaxies in the Hubble Deep Field: colour selection and star formation history to z~4. *MNRAS*, 283(4):1388–1404, December 1996. doi: 10.1093/mnras/283.4.1388.

John Magorrian, Scott Tremaine, Douglas Richstone, Ralf Bender, Gary Bower, Alan Dressler, S. M. Faber, Karl Gebhardt, Richard Green, Carl Grillmair, and et al. The demography of massive dark objects in galaxy centers. *The Astronomical Journal*, 115(6):2285, Jun 1998. ISSN 1538-3881. doi: 10.1086/300353. URL https://dx.doi.org/10.1086/300353.

Gary A. Mamon and Joseph Silk. The current status of galaxy formation. *Research in Astronomy and Astrophysics*, 12(8):917, Aug 2012. ISSN 1674-4527. doi: 10.1088/1674-4527/12/8/004. URL https://dx.doi.org/10.1088/1674-4527/12/8/004.

Justin Matejka and George Fitzmaurice. Same stats, different graphs: generating datasets with varied appearance and identical statistics through simulated annealing. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pages 1290–1294, 2017.

Rebecca L. Matz, Benjamin P. Koester, Stefano Fiorini, Galina Grom, Linda Shepard, Charles G. Stangor, Brad Weiner, and Timothy A. McKay. Patterns of gendered performance differences in large introductory courses at five research universities. *AERA Open*, 3(4):2332858417743754, Oct 2017. ISSN 2332-8584. doi: 10.1177/2332858417743754. URL https://doi.org/10.1177/2332858417743754.

Ian McConachie, Gillian Wilson, Ben Forrest, Z. Cemile Marsan, Adam Muzzin, M. C. Cooper, Marianna Annunziatella, Danilo Marchesini, Jeffrey C. C. Chan, Percy Gomez, Mohamed H. Abdullah, Paolo Saracco, and Julie Nantais. Spectroscopic Confirmation of a Protocluster at z = 3.37 with a High Fraction of Quiescent Galaxies. *ApJ*, 926(1):37, February 2022. doi: 10.3847/1538-4357/ac2b9f.

H. J. McCracken, B. Milvang-Jensen, J. Dunlop, M. Franx, J. P. U. Fynbo, O. Le Fèvre, J. Holt, K. I. Caputi, Y. Goranova, F. Buitrago, J. P. Emerson, W. Freudling, P. Hudelot, C. López-Sanjuan, F. Magnard, Y. Mellier, P. Møller, K. K. Nilsson, W. Sutherland, L. Tasca, and J. Zabl. Ultravista: a new ultra-deep near-infrared survey in cosmos. *Astronomy and Astrophysics*, 544: A156, Aug 2012. ISSN 0004-6361, 1432-0746. doi: 10.1051/0004-6361/201219507. URL https://www.aanda.org/articles/aa/abs/2012/08/aa19507-12/aa19507-12.html.

Robert R. McCrae, Antonio Terracciano, and Personality Profiles of Cultures Project. Universal features of personality traits from the observer's perspective: data from 50 cultures. *Journal of Personality and Social Psychology*, 88(3):547–561, Mar 2005. ISSN 0022-3514. doi: 10.1037/0022-3514.88.3.547.

Sara McLanahan, Laura Tach, and Daniel Schneider. The causal effects of father absence. *Annual review of sociology*, 39:399–427, Jul 2013. ISSN 0360-0572. doi: 10.1146/annurev-soc-071312-145704. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3904543/.

P. Melchior and A. D. Goulding. Filling the gaps: Gaussian mixture models from noisy, truncated or incomplete samples. *Astronomy and Computing*, 25:183–194, Oct 2018. ISSN 2213-1337. doi: 10.1016/j.ascom.2018.09.013. URL https://www.sciencedirect.com/science/article/pii/S2213133718300489.

Robert Metcalfe, Simon Burgess, and Steven Proud. Student effort and educational attainment: Using the england football team to identify the education production function. *Centre for Market and Public Organisation*, 2011.

Ben Moore, Neal Katz, George Lake, Alan Dressler, and Augustus Oemler. Galaxy harassment and the evolution of clusters of galaxies. *Nature*, 379(6566):613–616, February 1996. doi: 10.1038/379613a0.

John Moustakas, Alison L. Coil, James Aird, Michael R. Blanton, Richard J. Cool, Daniel J. Eisenstein, Alexander J. Mendez, Kenneth C. Wong, Guangtun Zhu, and Stéphane Arnouts. Primus: Constraints on star formation quenching and galaxy merging, and the evolution of the stellar mass function from z=0–1. *The Astrophysical Journal*, 767(1):50, Mar 2013. ISSN 0004-637X. doi: 10.1088/0004-637X/767/1/50. URL https://doi.org/10.1088/0004-637x/767/1/50.

S. G. Murray, C. Power, and A. S. G. Robotham. Hmfcalc: An online tool for calculating dark matter halo mass functions. *Astronomy and Computing*, 3–4:23–34, Nov 2013. ISSN 2213-1337. doi: 10.1016/j.ascom.2013.11.001.

Adam Muzzin, Gillian Wilson, H. K. C. Yee, Henk Hoekstra, David Gilbank, Jason Surace, Mark Lacy, Kris Blindert, Subhabrata Majumdar, Ricardo Demarco, and et al. Spectroscopic confirmation of two massive red-sequence-selected galaxy clusters atz∼ 1.2 in the sparcs-north cluster survey. *The Astrophysical Journal*, 698(2):1934–1942, Jun 2009. ISSN 0004-637X. doi: 10.1088/0004-637X/698/2/1934. URL https://doi.org/10.1088/0004-637x/698/2/1934.

J Myles, A Alarcon, A Amon, C Sánchez, S Everett, J DeRose, J McCullough, D Gruen, G M Bernstein, M A Troxel, S Dodelson, A Campos, N MacCrann, B Yin, M Raveri, A Amara, M R Becker, A Choi, J Cordero, K Eckert, M Gatti, G Giannini, J Gschwend, R A Gruendl, I Harrison, W G Hartley, E M Huff, N Kuropatkin, H Lin, D Masters, R Miquel, J Prat, A Roodman, E S Rykoff, I Sevilla-Noarbe, E Sheldon, R H Wechsler, B Yanny, T M C Abbott, M Aguena, S Allam, J Annis, D Bacon, E Bertin, S Bhargava, S L Bridle, D Brooks, D L Burke, A Carnero Rosell, M Carrasco Kind, J Carretero, F J Castander, C Conselice, M Costanzi, M Crocce, L N da Costa, M E S Pereira, S Desai, H T Diehl, T F Eifler, J Elvin-Poole, A E Evrard, I Ferrero, A Ferté, B Flaugher, P Fosalba, J Frieman, J García-Bellido, E Gaztanaga, T Giannantonio, S R Hinton, D L Hollowood, K Honscheid, B Hoyle, D Huterer, D J James, E Krause, K Kuehn, O Lahav, M Lima, M A G Maia, J L Marshall, P Martini, P Melchior, F Menanteau, J J Mohr, R Morgan, J Muir, R L C Ogando, A Palmese, F Paz-Chinchón, A A Plazas, M Rodriguez-Monroy, S Samuroff, E Sanchez, V Scarpine, L F Secco, S Serrano, M Smith, M Soares-Santos, E Suchyta, M E C Swanson, G Tarle, D Thomas, C To, T N Varga, J Weller, and W Wester. Dark energy survey year 3 results: redshift calibration of the weak lensing source galaxies. *Monthly Notices of the Royal Astronomical Society*, 505(3):4249–4277, Aug 2021. ISSN 0035-8711. doi: 10.1093/mnras/stab1515. URL https://doi.org/10.1093/mnras/stab1515.

Thorsten Naab and Jeremiah P. Ostriker. Theoretical challenges in galaxy formation. *Annual Review of Astronomy and Astrophysics*, 55(1):59–109, 2017. doi: 10.1146/annurev-astro-081913-040019. URL https://doi.org/10.1146/annurev-astro-081913-040019.

Ulric Neisser, Gwyneth Boodoo, Thomas J. Bouchard Jr., A. Wade Boykin, Nathan Brody, Stephen J. Ceci, Diane F. Halpern, John C. Loehlin, Robert Perloff, Robert J. Sternberg, and Susana Urbina. Intelligence: Knowns and unknowns. *American Psychologist*, 51:77–101, 1996. ISSN 1935-990X. doi: 10.1037/0003-066X.51.2.77.

Dylan Nelson, Annalisa Pillepich, Volker Springel, Rainer Weinberger, Lars Hernquist, Ruediger Pakmor, Shy Genel, Paul Torrey, Mark Vogelsberger, Guinevere Kauffmann, and et al. First results from the illustristng simulations: the galaxy color bimodality. *Monthly Notices of the Royal Astronomical Society*, 475(1):624–647, Mar 2018. ISSN 0035-8711, 1365-2966. doi: 10.1093/mnras/stx3040. URL http://arxiv.org/abs/1707.03395. arXiv: 1707.03395.

Atsushi J. Nishizawa, Masamue Oguri, Taira Oogi, Surhud More, Takahiro Nishimichi, Masahiro Nagashima, Yen-Ting Lin, Rachel Mandelbaum, Masahiro Takada, Neta Bahcall, and et al. First results on the cluster galaxy population from the subaru hyper suprime-cam survey. ii. faint end color-magnitude diagrams and radial profiles of red and blue galaxies at $0.1 < z < 1.1$. *Publications of the Astronomical Society of Japan*, 70(SP1), Jan 2018. ISSN 0004-6264, 2053-051X. doi: 10.1093/pasj/psx106. URL http://arxiv.org/abs/1709.01136. arXiv: 1709.01136.

Julie Noble and Richard Sawyer. Predicting different levels of academic success in college using high school GPA and ACT composite score. *ACT Research Report Series*, Aug 2002. URL https://eric.ed.gov/?id=ED469746.

Masamune Oguri. A cluster finding algorithm based on the multiband identification of red sequence galaxies. *Monthly Notices of the Royal Astronomical Society*, 444(1):147–161, Oct 2014. ISSN 0035-8711. doi: 10.1093/mnras/stu1446. URL https://doi.org/10.1093/mnras/stu1446.

Ann Owens, Sean F. Reardon, and Christopher Jencks. Income segregation between schools and school districts. *American Educational Research Journal*, 53(4):1159–1197, Aug 2016. ISSN 0002-8312. doi: 10.3102/0002831216652722. URL https://doi.org/10.3102/0002831216652722.

Matthew S. Panizzon, Eero Vuoksimaa, Kelly M. Spoon, Kristen C. Jacobson, Michael J. Lyons, Carol E. Franz, Hong Xian, Terrie Vasilopoulos, and William S. Kremen. Genetic and environmental influences on general cognitive ability: Is *g* a valid latent construct? *Intelligence*, 43:65–76, 2014. ISSN 0160-2896. doi: https://doi.org/10.1016/j.intell.2014.01.008. URL https://www.sciencedirect.com/science/article/pii/S0160289614000099.

Ernest T. Pascarella, Christopher T. Pierson, Gregory C. Wolniak, and Patrick T. Terenzini. First-generation college students. *The Journal of Higher Education*, 75(3):249–284, May 2004. ISSN 0022-1546. doi: 10.1080/00221546.2004.11772256. URL https://doi.org/10.1080/00221546.2004.11772256.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Ying-jie Peng, Simon J. Lilly, Katarina Kovač, Micol Bolzonella, Lucia Pozzetti, Alvio Renzini, Gianni Zamorani, Olivier Ilbert, Christian Knobel, Angela Iovino, Christian Maier, Olga Cucciati, Lidia Tasca, C. Marcella Carollo, John Silverman, Pawel Kampczyk, Loic de Ravel, David Sanders, Nicholas Scoville, Thierry Contini, Vincenzo Mainieri, Marco Scodeggio, Jean-Paul Kneib, Olivier Le Fèvre, Sandro Bardelli, Angela Bongiorno, Karina Caputi, Graziano Coppa, Sylvain de la Torre, Paolo Franzetti, Bianca Garilli, Fabrice Lamareille, Jean-Francois Le Borgne, Vincent Le Brun, Marco Mignoli, Enrique Perez Montero, Roser Pello, Elena Ricciardelli, Masayuki Tanaka, Laurence Tresse, Daniela Vergani, Niraj Welikala, Elena Zucca, Pascal Oesch, Ummi Abbas, Luke Barnes, Rongmon Bordoloi, Dario Bottini, Alberto Cappi, Paolo Cassata, Andrea Cimatti, Marco Fumana, Gunther Hasinger, Anton Koekemoer, Alexei Leauthaud, Dario Maccagni, Christian Marinoni, Henry McCracken, Pierdomenico Memeo, Baptiste Meneux, Preethi Nair, Cristiano Porciani, Valentina Presotto, and Roberto Scaramella. Mass and environment as drivers of galaxy evolution in sdss and zcosmos and the origin of the schechter function. *The Astrophysical Journal*, 721(1):193–221, Aug 2010. ISSN 0004-637X. doi: 10.1088/0004-637X/721/1/193. URL https://doi.org/10.1088/0004-637x/721/1/193.

Annalisa Pillepich, Volker Springel, Dylan Nelson, Shy Genel, Jill Naiman, Rüdiger Pakmor, Lars Hernquist, Paul Torrey, Mark Vogelsberger, Rainer Weinberger, and Federico Marinacci. Simulating galaxy formation with the illustristng model. *Monthly Notices of the Royal Astronomical Society*, 473(3):4077–4106, Jan 2018. ISSN 0035-8711. doi: 10.1093/mnras/stx2656. URL https://doi.org/10.1093/mnras/stx2656.

Nolan Pope, Richard Patterson, Uros Petronijevic, and Philip Oreopoulos. The disappointing impact of encouraging students to study more, Nov 2018. URL https://cepr.org/voxeu/columns/disappointing-impact-encouraging-students-study-more.

Arthur E. Poropat. A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin*, 135(2):322–338, 2009. ISSN 1939-1455. doi: 10.1037/a0014996.

William H. Press and Paul Schechter. Formation of galaxies and clusters of galaxies by self-similar gravitational condensation. *The Astrophysical Journal*, 187:425–438, Feb 1974. ISSN 0004-637X. doi: 10.1086/152650. ADS Bibcode: 1974ApJ...187..425P.

A Repp and H Ebeling. Science from a glimpse: Hubble snapshot observations of massive galaxy clusters. *Monthly Notices of the Royal Astronomical Society*, 479(1):844–864, Sep 2018. ISSN 0035-8711. doi: 10.1093/mnras/sty1489.

Michelle Richardson, Charles Abraham, and Rod Bond. Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin*, 138:353–387, 2012. ISSN 1939-1455. doi: 10.1037/a0026838. URL https://doi.org/10.1037/a0026838.

Susan D Ridgell and John W Lounsbury. Predicting academic success: General intelligence," big five" personality traits, and work drive. *College Student Journal*, 38(4):607–619, 2004.

Sonia Roccas, Lilach Sagiv, Shalom H. Schwartz, and Ariel Knafo. The big five personality factors and personal values. *Personality and Social Psychology Bulletin*, 28(6):789–801, Jun 2002. ISSN 0146-1672. doi: 10.1177/0146167202289008. URL https://doi.org/10.1177/0146167202289008.

M K Rodriguez Wimberly, M C Cooper, D C Baxter, M Boylan-Kolchin, J S Bullock, S P Fillingham, A P Ji, L V Sales, and J D Simon. Sizing from the smallest scales: the mass of the milky way. *Monthly Notices of the Royal Astronomical Society*, 513(4):4968–4982, Jul 2022. ISSN 0035-8711. doi: 10.1093/mnras/stac1265. URL https://doi.org/10.1093/mnras/stac1265.

Carlos Felipe Rodríguez-Hernández, Eduardo Cascallar, and Eva Kyndt. Socio-economic status and academic performance in higher education: A systematic review. *Educational Research Review*, 29:100305, Feb 2020. ISSN 1747-938X. doi: 10.1016/j.edurev.2019.100305. URL https://www.sciencedirect.com/science/article/pii/S1747938X18302744.

Cristóbal Romero and Sebastián Ventura. Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6):601–618, Nov 2010. ISSN 1558-2442. doi: 10.1109/TSMCC.2010.2053532.

E. Rozo, E. S. Rykoff, A. Abate, C. Bonnett, M. Crocce, C. Davis, B. Hoyle, B. Leistedt, H. V. Peiris, R. H. Wechsler, T. Abbott, F. B. Abdalla, M. Banerji, A. H. Bauer, A. Benoit-Lévy, G. M. Bernstein, E. Bertin, D. Brooks, E. Buckley-Geer, D. L. Burke, D. Capozzi, A. Carnero Rosell, D. Carollo, M. Carrasco Kind, J. Carretero, F. J. Castander, M. J. Childress, C. E. Cunha, C. B. D'Andrea, T. Davis, D. L. DePoy, S. Desai, H. T. Diehl, J. P. Dietrich, P. Doel, T. F. Eifler, A. E. Evrard, A. Fausti Neto, B. Flaugher, P. Fosalba, J. Frieman, E. Gaztanaga, D. W. Gerdes, K. Glazebrook, D. Gruen, R. A. Gruendl, K. Honscheid, D. J. James, M. Jarvis, A. G. Kim, K. Kuehn, N. Kuropatkin, O. Lahav, C. Lidman, M. Lima, M. A. G. Maia, M. March, P. Martini, P. Melchior, C. J. Miller, R. Miquel, J. J. Mohr, R. C. Nichol, B. Nord, C. R. O'Neill, R. Ogando, A. A. Plazas, A. K. Romer, A. Roodman, M. Sako, E. Sanchez, B. Santiago, M. Schubnell, I. Sevilla-Noarbe, R. C. Smith, M. Soares-Santos, F. Sobreira, E. Suchyta, M. E. C. Swanson, J. Thaler, D. Thomas, S. Uddin, V. Vikram, A. R. Walker, W. Wester, Y. Zhang, and L. N. da Costa. redmagic: Selecting luminous red galaxies from the des science verification data. *Monthly Notices of the Royal Astronomical Society*, 461(2):1431–1450, Sep 2016. ISSN 0035-8711, 1365-2966. doi: 10.1093/mnras/stw1281. URL https://doi.org/10.1093/mnras/stw1281. arXiv: 1507.05460.

Eduardo Rozo, Eli S. Rykoff, August Evrard, Matthew Becker, Timothy McKay, Risa H. Wechsler, Benjamin P. Koester, Jiangang Hao, Sarah Hansen, Erin Sheldon, and et al. Constraining the scatter in the mass-richness relation of maxbcg clusters with weak lensing and x-ray data. *The Astrophysical Journal*, 699(1):768–781, Jun 2009a. ISSN 0004-637X. doi: 10.1088/0004-637X/699/1/768. URL https://doi.org/10.1088/0004-637x/699/1/768.

Eduardo Rozo, Eli S. Rykoff, Benjamin P. Koester, Timothy McKay, Jiangang Hao, August Evrard, Risa H. Wechsler, Sarah Hansen, Erin Sheldon, David Johnston, and et al. Improvement of the richness estimates of maxbcg clusters. *The Astrophysical Journal*, 703(1):601–613, Aug 2009b. ISSN 0004-637X. doi: 10.1088/0004-637X/703/1/601. URL https://doi.org/10.1088/0004-637x/703/1/601.

E. Rykoff. lambda is biased low when trying to reproduce the des y1 redmapper run. https://github.com/erykoff/redmapper/issues/78#issuecomment-860775426, Jun 2021.

E. S. Rykoff, E. Rozo, M. T. Busha, C. E. Cunha, A. Finoguenov, A. Evrard, J. Hao, B. P. Koester, A. Leauthaud, B. Nord, and et al. redmapper. i. algorithm and sdss dr8 catalog. *The Astrophysical Journal*, 785(2):104, Apr 2014. ISSN 0004-637X. doi: 10.1088/0004-637X/785/2/104. URL https://doi.org/10.1088%2F0004-637x%2F785%2F2%2F104.

E. S. Rykoff, E. Rozo, D. Hollowood, A. Bermeo-Hernandez, T. Jeltema, J. Mayers, A. K. Romer, P. Rooney, A. Saro, C. Vergara Cervantes, R. H. Wechsler, H. Wilcox, T. M. C. Abbott, F. B. Abdalla, S. Allam, J. Annis, A. Benoit-Lévy, G. M. Bernstein, E. Bertin, D. Brooks, D. L. Burke, D. Capozzi, A. Carnero Rosell, M. Carrasco Kind, F. J. Castander, M. Childress, C. A. Collins, C. E. Cunha, C. B. D'Andrea, L. N. da Costa, T. M. Davis, S. Desai, H. T. Diehl, J. P. Dietrich, P. Doel, A. E. Evrard, D. A. Finley, B. Flaugher, P. Fosalba, J. Frieman, K. Glazebrook, D. A. Goldstein, D. Gruen, R. A. Gruendl, G. Gutierrez, M. Hilton, K. Honscheid, B. Hoyle, D. J. James, S. T. Kay, K. Kuehn, N. Kuropatkin, O. Lahav, G. F. Lewis, C. Lidman, M. Lima, M. A. G. Maia, R. G. Mann, J. L. Marshall, P. Martini, P. Melchior, C. J. Miller, R. Miquel, J. J. Mohr,

R. C. Nichol, B. Nord, R. Ogando, A. A. Plazas, K. Reil, M. Sahlén, E. Sanchez, B. Santiago, V. Scarpine, M. Schubnell, I. Sevilla-Noarbe, R. C. Smith, M. Soares-Santos, F. Sobreira, J. P. Stott, E. Suchyta, M. E. C. Swanson, G. Tarle, D. Thomas, D. Tucker, S. Uddin, P. T. P. Viana, V. Vikram, A. R. Walker, Y. Zhang, and DES Collaboration. The RedMaPPer Galaxy Cluster Catalog From DES Science Verification Data. *The Astrophysical Journal Supplement Series*, 224(1):1, May 2016. doi: 10.3847/0067-0049/224/1/1.

Paul R. Sackett, Matthew J. Borneman, and Brian S. Connelly. High stakes testing in higher education and employment: Appraising the evidence for validity and fairness. *American Psychologist*, 63:215–227, 2008. ISSN 1935-990X. doi: 10.1037/0003-066X.63.4.215.

B. Sartoris, A. Biviano, C. Fedeli, J. G. Bartlett, S. Borgani, M. Costanzi, C. Giocoli, L. Moscardini, J. Weller, B. Ascaso, S. Bardelli, S. Maurogordato, and P. T. P. Viana. Next generation cosmology: constraints from the Euclid galaxy cluster survey. *MNRAS*, 459(2):1764–1780, June 2016. doi: 10.1093/mnras/stw630.

Guan Saw, Chi-Ning Chang, and Hsun-Yu Chan. Cross-sectional and longitudinal disparities in stem career aspirations at the intersection of gender, race/ethnicity, and socioeconomic status. *Educational Researcher*, 47(8):525–531, Nov 2018. ISSN 0013-189X. doi: 10.3102/0013189X18787818. URL https://doi.org/10.3102/0013189X18787818.

Kevin Schawinski. Black hole – galaxy co-evolution. *arXiv: 1206.2661 [astro-ph]*, Jun 2012. doi: 10.48550/arXiv.1206.2661. URL http://arxiv.org/abs/1206.2661. arXiv:1206.2661 [astro-ph].

Kevin Schawinski, Daniel Thomas, Marc Sarzi, Claudia Maraston, Sugata Kaviraj, Seok-Joo Joo, Sukyoung K. Yi, and Joseph Silk. Observational evidence for agn feedback in early-type galaxies. *Monthly Notices of the Royal Astronomical Society*, 382(4):1415–1431, Dec 2007. ISSN 0035-8711. doi: 10.1111/j.1365-2966.2007.12487.x. URL https://doi.org/10.1111/j.1365-2966.2007.12487.x.

Kevin Schawinski, C. Megan Urry, Brooke D. Simmons, Lucy Fortson, Sugata Kaviraj, William C. Keel, Chris J. Lintott, Karen L. Masters, Robert C. Nichol, Marc Sarzi, and et al. The green valley is a red herring: Galaxy zoo reveals two evolutionary pathways towards quenching of star formation in early- and late-type galaxies. *Monthly Notices of the Royal Astronomical Society*, 440(1):889–907, May 2014. ISSN 0035-8711. doi: 10.1093/mnras/stu327. URL https://doi.org/10.1093/mnras/stu327.

E. Schinnerer, C. L. Carilli, N. Z. Scoville, M. Bondi, P. Ciliegi, P. Vettolani, O. Le Fèvre, A. M. Koekemoer, F. Bertoldi, and C. D. Impey. The vla-cosmos survey. i. radio identifications from the pilot project. *The Astronomical Journal*, 128(5):1974, Nov 2004. ISSN 1538-3881. doi: 10.1086/424860. URL https://dx.doi.org/10.1086/424860.

S J Schmidt, A I Malz, J Y H Soo, I A Almosallam, M Brescia, S Cavuoti, J Cohen-Tanugi, A J Connolly, J DeRose, P E Freeman, M L Graham, K G Iyer, M J Jarvis, J B Kalmbach, E Kovacs, A B Lee, G Longo, C B Morrison, J A Newman, E Nourbakhsh, E Nuss, T Pospisil, H Tranin, R H Wechsler, R Zhou, R Izbicki, and (The LSST Dark Energy Science Collaboration).

Evaluation of probabilistic photometric redshift estimation approaches for the rubin observatory legacy survey of space and time (lsst). *Monthly Notices of the Royal Astronomical Society*, 499 (2):1587–1606, Oct 2020. ISSN 0035-8711. doi: 10.1093/mnras/staa2799.

David P. Schmitt, Anu Realo, Martin Voracek, and Jüri Allik. Why can't a man be more like a woman? sex differences in big five personality traits across 55 cultures. *Journal of Personality and Social Psychology*, 94(1):168–182, Jan 2008. ISSN 0022-3514. doi: 10.1037/0022-3514.94.1.168.

N. Scoville, H. Aussel, M. Brusa, P. Capak, C. M. Carollo, M. Elvis, M. Giavalisco, L. Guzzo, G. Hasinger, C. Impey, J.-P. Kneib, O. LeFevre, S. J. Lilly, B. Mobasher, A. Renzini, R. M. Rich, D. B. Sanders, E. Schinnerer, D. Schminovich, P. Shopbell, Y. Taniguchi, and N. D. Tyson. The cosmic evolution survey (cosmos): Overview. *The Astrophysical Journal Supplement Series*, 172(1):1, Sep 2007. ISSN 0067-0049. doi: 10.1086/516585. URL https://iopscience.iop.org/article/10.1086/516585/meta.

L. F. Secco, S. Samuroff, E. Krause, B. Jain, J. Blazek, M. Raveri, A. Campos, A. Amon, A. Chen, C. Doux, A. Choi, D. Gruen, G. M. Bernstein, C. Chang, J. DeRose, J. Myles, A. Ferté, P. Lemos, D. Huterer, J. Prat, M. A. Troxel, N. MacCrann, A. R. Liddle, T. Kacprzak, X. Fang, C. Sánchez, S. Pandey, S. Dodelson, P. Chintalapati, K. Hoffmann, A. Alarcon, O. Alves, F. Andrade-Oliveira, E. J. Baxter, K. Bechtol, M. R. Becker, A. Brandao-Souza, H. Camacho, A. Carnero Rosell, M. Carrasco Kind, R. Cawthon, J. P. Cordero, M. Crocce, C. Davis, E. Di Valentino, A. Drlica-Wagner, K. Eckert, T. F. Eifler, M. Elidaiana, F. Elsner, J. Elvin-Poole, S. Everett, P. Fosalba, O. Friedrich, M. Gatti, G. Giannini, R. A. Gruendl, I. Harrison, W. G. Hartley, K. Herner, H. Huang, E. M. Huff, M. Jarvis, N. Jeffrey, N. Kuropatkin, P.-F. Leget, J. Muir, J. Mccullough, A. Navarro Alsina, Y. Omori, Y. Park, A. Porredon, R. Rollins, A. Roodman, R. Rosenfeld, A. J. Ross, E. S. Rykoff, J. Sanchez, I. Sevilla-Noarbe, E. S. Sheldon, T. Shin, A. Troja, I. Tutusaus, T. N. Varga, N. Weaverdyck, R. H. Wechsler, B. Yanny, B. Yin, Y. Zhang, J. Zuntz, T. M. C. Abbott, M. Aguena, S. Allam, J. Annis, D. Bacon, E. Bertin, S. Bhargava, S. L. Bridle, D. Brooks, E. Buckley-Geer, D. L. Burke, J. Carretero, M. Costanzi, L. N. da Costa, J. De Vicente, H. T. Diehl, J. P. Dietrich, P. Doel, I. Ferrero, B. Flaugher, J. Frieman, J. García-Bellido, E. Gaztanaga, D. W. Gerdes, T. Giannantonio, J. Gschwend, G. Gutierrez, S. R. Hinton, D. L. Hollowood, K. Honscheid, B. Hoyle, D. J. James, T. Jeltema, K. Kuehn, O. Lahav, M. Lima, H. Lin, M. A. G. Maia, J. L. Marshall, P. Martini, P. Melchior, F. Menanteau, R. Miquel, J. J. Mohr, R. Morgan, R. L. C. Ogando, A. Palmese, F. Paz-Chinchón, D. Petravick, A. Pieres, A. A. Plazas Malagón, M. Rodriguez-Monroy, A. K. Romer, E. Sanchez, V. Scarpine, M. Schubnell, D. Scolnic, S. Serrano, M. Smith, M. Soares-Santos, E. Suchyta, M. E. C. Swanson, G. Tarle, D. Thomas, and C. To. Dark energy survey year 3 results: Cosmology from cosmic shear and robustness to modeling uncertainty. *Physical Review D*, 105(2):023515, Jan 2022. doi: 10.1103/PhysRevD.105.023515. URL https://link.aps.org/doi/10.1103/PhysRevD.105.023515.

S. F. Shandarin and Ya. B. Zeldovich. The large-scale structure of the universe: Turbulence, intermittency, structures in a self-gravitating medium. *Reviews of Modern Physics*, 61(2): 185–220, Apr 1989. doi: 10.1103/RevModPhys.61.185. URL https://link.aps.org/doi/10.1103/RevModPhys.61.185.

Melanie Simet, Tom McClintock, Rachel Mandelbaum, Eduardo Rozo, Eli Rykoff, Erin Sheldon, and Risa H. Wechsler. Erratum: Weak lensing measurement of the mass–richness relation of sdss redmapper clusters. *Monthly Notices of the Royal Astronomical Society*, 480(4):5385–5385, Nov 2018. ISSN 0035-8711. doi: 10.1093/mnras/sty2318. URL https://academic.oup.com/mnras/article/480/4/5385/5079647.

Amber B. Simmons and Andrew F. Heckler. Grades, grade component weighting, and demographic disparities in introductory physics. *Physical Review Physics Education Research*, 16(2):020125, Oct 2020. doi: 10.1103/PhysRevPhysEducRes.16.020125. URL https://link.aps.org/doi/10.1103/PhysRevPhysEducRes.16.020125.

Varsha Singh, Sonika Thakral, Kunal Singh, and Rahul Garg. Examining cognitive sex differences in elite math intensive education: Preliminary evidence from a gender inequitable country. *Trends in Neuroscience and Education*, 26:100172, Mar 2022. ISSN 2211-9493. doi: 10.1016/j.tine.2022.100172. URL https://www.sciencedirect.com/science/article/pii/S2211949322000011.

Diana Lopes Soares, Gina C. Lemos, Ricardo Primi, and Leandro S. Almeida. The relationship between intelligence and academic achievement throughout middle school: The role of students' prior academic performance. *Learning and Individual Differences*, 41:73–78, Jul 2015. ISSN 1041-6080. doi: 10.1016/j.lindif.2015.02.005.

Rachel S. Somerville and Romeel Davé. Physical models of galaxy formation in a cosmological framework. *Annual Review of Astronomy and Astrophysics*, 53(1):51–113, 2015. doi: 10.1146/annurev-astro-082812-140951. URL https://doi.org/10.1146/annurev-astro-082812-140951.

J. S. Speagle, C. L. Steinhardt, P. L. Capak, and J. D. Silverman. A highly consistent framework for the evolution of the star-forming "main sequence" from $z \sim 0$–6. *The Astrophysical Journal Supplement Series*, 214(2):15, Sep 2014. ISSN 0067-0049. doi: 10.1088/0067-0049/214/2/15.

D. Spergel, N. Gehrels, C. Baltay, D. Bennett, J. Breckinridge, M. Donahue, A. Dressler, B. S. Gaudi, T. Greene, O. Guyon, C. Hirata, J. Kalirai, N. J. Kasdin, B. Macintosh, W. Moos, S. Perlmutter, M. Postman, B. Rauscher, J. Rhodes, Y. Wang, D. Weinberg, D. Benford, M. Hudson, W. S. Jeong, Y. Mellier, W. Traub, T. Yamada, P. Capak, J. Colbert, D. Masters, M. Penny, D. Savransky, D. Stern, N. Zimmerman, R. Barry, L. Bartusek, K. Carpenter, E. Cheng, D. Content, F. Dekens, R. Demers, K. Grady, C. Jackson, G. Kuan, J. Kruk, M. Melton, B. Nemati, B. Parvin, I. Poberezhskiy, C. Peddie, J. Ruffa, J. K. Wallace, A. Whipple, E. Wollack, and F. Zhao. Wide-Field InfrarRed Survey Telescope-Astrophysics Focused Telescope Assets WFIRST-AFTA 2015 Report. *arXiv e-prints*, art. arXiv:1503.03757, March 2015.

Markus Wolfgang Hermann Spitzer. Just do it! study time increases mathematical achievement scores for grade 4-10 students in a large longitudinal cross-country study. *European Journal of Psychology of Education*, 37(1):39–53, Mar 2022. ISSN 1878-5174. doi: 10.1007/s10212-021-00546-0. URL https://doi.org/10.1007/s10212-021-00546-0.

Volker Springel. E pur si muove: Galilean-invariant cosmological hydrodynamical simulations on a moving mesh. *Monthly Notices of the Royal Astronomical Society*, 401(2):791–851, Jan 2010. ISSN 0035-8711. doi: 10.1111/j.1365-2966.2009.15715.x.

Volker Springel, Simon D. M. White, Adrian Jenkins, Carlos S. Frenk, Naoki Yoshida, Liang Gao, Julio Navarro, Robert Thacker, Darren Croton, John Helly, and et al. Simulating the joint evolution of quasars, galaxies and their large-scale distribution. *Nature*, 435(7042):629–636, Jun 2005. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature03597. URL http://arxiv.org/abs/astro-ph/0504097. arXiv: astro-ph/0504097.

Volker Springel, Carlos S. Frenk, and Simon D. M. White. The large-scale structure of the universe. *Nature*, 440(70887088):1137–1144, Apr 2006. ISSN 1476-4687. doi: 10.1038/nature04805. URL https://www.nature.com/articles/nature04805.

Dominik Steinhauser, Sabine Schindler, and Volker Springel. Simulations of ram-pressure stripping in galaxy-cluster interactions. *Astronomy & Astrophysics*, 591:A51, Jul 2016. ISSN 0004-6361, 1432-0746. doi: 10.1051/0004-6361/201527705. URL https://www.aanda.org/articles/aa/abs/2016/07/aa27705-15/aa27705-15.html.

P Steyrleithner, G Hensler, and A Boselli. The effect of ram-pressure stripping on dwarf galaxies. *Monthly Notices of the Royal Astronomical Society*, 494(1):1114–1127, May 2020. ISSN 0035-8711. doi: 10.1093/mnras/staa775. URL https://doi.org/10.1093/mnras/staa775.

Ralph Stinebrickner and Todd R. Stinebrickner. The causal effect of studying on academic performance. *The B.E. Journal of Economic Analysis & Policy*, 8(1), Jun 2008. ISSN 1935-1682. doi: 10.2202/1935-1682.1868. URL https://www.degruyter.com/document/doi/10.2202/1935-1682.1868/html.

Iskra Strateva, Željko Ivezić, Gillian R. Knapp, Vijay K. Narayanan, Michael A. Strauss, James E. Gunn, Robert H. Lupton, David Schlegel, Neta A. Bahcall, Jon Brinkmann, and et al. Color separation of galaxy types in the sloan digital sky survey imaging data. *The Astronomical Journal*, 122:1861–1874, Oct 2001. ISSN 0004-6256. doi: 10.1086/323301. URL http://adsabs.harvard.edu/abs/2001AJ....122.1861S.

Kori J. Stroub and Meredith P. Richards. From resegregation to reintegration: Trends in the racial/ethnic segregation of metropolitan public schools, 1993–2009. *American Educational Research Journal*, 50(3):497–531, Jun 2013. ISSN 0002-8312. doi: 10.3102/0002831213478462. URL https://doi.org/10.3102/0002831213478462.

M. Symeonidis, N. Maddox, M. J. Jarvis, M. J. Micha lowski, P. Andreani, D. L. Clements, G. De Zotti, S. Duivenvoorden, J. Gonzalez-Nuevo, E. Ibar, R. J. Ivison, L. Leeuw, M. J. Page, R. Shirley, M. W. L. Smith, and M. Vaccari. The star formation rates of QSOs. *MNRAS*, 514(3): 4450–4464, August 2022. doi: 10.1093/mnras/stac1359.

Alexander S. Szalay, Jim Gray, Ani R. Thakar, Peter Z. Kunszt, Tanu Malik, Jordan Raddick, Christopher Stoughton, and Jan vandenBerg. The sdss skyserver. In *SIGMOD '02: Proceedings of the 2002 ACM SIGMOD international conference on Management of data*, page 570–581.

Association for Computing Machinery, Jun 2002. ISBN 978-1-58113-497-1. URL https://doi.org/10.1145/564691.564758.

Kunio Takezawa. *Introduction to nonparametric regression*. John Wiley & Sons, 2005.

University of Michigan. Chatgpt teach-out, Mar 2023. URL https://online.umich.edu/teach-outs/chatgpt-teach-out/.

Nishmin Unwalla. Comparative analysis of study habits between males and females. *International Journal of Innovative Science and Research Technology*, 5(7):182–187, 2020.

Stefan van der Walt, S. Chris Colbert, and Gael Varoquaux. The numpy array: A structure for efficient numerical computation. *Computing in Science Engineering*, 13(2):22–30, 2011. doi: 10.1109/MCSE.2011.37.

F. Vazza and A. Feletti. The quantitative comparison between the neuronal network and the cosmic web. *Frontiers in Physics*, 8, 2020. ISSN 2296-424X. URL https://www.frontiersin.org/articles/10.3389/fphy.2020.525731.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

G. Mark Voit. Tracing cosmic evolution with clusters of galaxies. *Reviews of Modern Physics*, 77 (1):207–258, April 2005. doi: 10.1103/RevModPhys.77.207.

Benedetta Vulcani, Bianca M. Poggianti, Rory Smith, Alessia Moretti, Yara L. Jaffé, Marco Gullieuszik, Jacopo Fritz, and Callum Bellhouse. The relevance of ram pressure stripping for the evolution of blue cluster galaxies as seen at optical wavelengths. *The Astrophysical Journal*, 927(1):91, Mar 2022. ISSN 0004-637X. doi: 10.3847/1538-4357/ac4809. URL https://dx.doi.org/10.3847/1538-4357/ac4809.

MaryBeth Walpole. Socioeconomic status and college: How SES affects college experiences and outcomes. *The Review of Higher Education*, 27(1):45–73, 2003. ISSN 1090-7009. doi: 10.1353/rhe.2003.0044. URL https://muse.jhu.edu/article/46608.

Noah Weaverdyck, Dhayaa Anbajagane, and August E. Evrard. Differential Assessment, Differential Benefit: Four-year Problem Roulette Analysis of STEM Practice Study. In *Proceedings of the Seventh ACM Conference on Learning @ Scale*, L@S '20, page 293–296. Association for Computing Machinery, Aug 2020. ISBN 978-1-4503-7951-9. doi: 10.1145/3386527.3406731. URL https://doi.org/10.1145/3386527.3406731.

Risa H. Wechsler and Jeremy L. Tinker. The connection between galaxies and their dark matter halos. *Annual Review of Astronomy and Astrophysics*, 56(1):435–487, Sep 2018. ISSN 0066-4146, 1545-4282. doi: 10.1146/annurev-astro-081817-051756. URL http://arxiv.org/abs/1804.03097. arXiv: 1804.03097.

Risa H. Wechsler, Joseph DeRose, Michael T. Busha, Matthew R. Becker, Eli Rykoff, and August Evrard. Addgals: Simulated sky catalogs for wide field galaxy surveys. *arXiv e-prints*, page arXiv:2105.12105, May 2021. URL https://ui.adsabs.harvard.edu/abs/2021arXiv210512105W.

Rainer Weinberger, Volker Springel, Lars Hernquist, Annalisa Pillepich, Federico Marinacci, Rüdiger Pakmor, Dylan Nelson, Shy Genel, Mark Vogelsberger, Jill Naiman, and Paul Torrey. Simulating galaxy formation with black hole driven thermal and kinetic feedback. *Monthly Notices of the Royal Astronomical Society*, 465(3):3291–3308, Mar 2017. ISSN 0035-8711, 1365-2966. doi: 10.1093/mnras/stw2944. URL http://arxiv.org/abs/1607.03486. arXiv: 1607.03486.

Yanna Weisberg, Colin DeYoung, and Jacob Hirsh. Gender differences in personality across the ten aspects of the Big Five. *Frontiers in Psychology*, 2, 2011. ISSN 1664-1078. doi: 10.3389/fpsyg.2011.00178. URL https://doi.org/10.3389/fpsyg.2011.00178.

Andrew R. Wetzel, Jeremy L. Tinker, and Charlie Conroy. Galaxy evolution in groups and clusters: star formation rates, red sequence fractions and the persistent bimodality. *Monthly Notices of the Royal Astronomical Society*, 424(1):232–243, Jul 2012. ISSN 0035-8711. doi: 10.1111/j.1365-2966.2012.21188.x. URL https://doi.org/10.1111/j.1365-2966.2012.21188.x.

Katherine E. Whitaker, Mariska Kriek, Pieter G. van Dokkum, Rachel Bezanson, Gabriel Brammer, Marijn Franx, and Ivo Labbé. A large population of massive compact post-starburst galaxies at $z > 1$: Implications for the size evolution and quenching mechanism of quiescent galaxies. *The Astrophysical Journal*, 745(2):179, Jan 2012. ISSN 0004-637X. doi: 10.1088/0004-637X/745/2/179. URL https://doi.org/10.1088/0004-637x/745/2/179.

J Harold Williams. Intelligence and delinquency. *J. Am. Inst. Crim. L. & Criminology*, 6:696, 1915.

Rik J. Williams, Ryan F. Quadri, Marijn Franx, Pieter van Dokkum, and Ivo Labbé. Detection of quiescent galaxies in a bicolor sequence from z = 0–2. *The Astrophysical Journal*, 691(2):1879–1895, Feb 2009. ISSN 0004-637X. doi: 10.1088/0004-637X/691/2/1879. URL https://doi.org/10.1088/0004-637x/691/2/1879.

Gillian Wilson, Adam Muzzin, H. K. C. Yee, Mark Lacy, Jason Surace, David Gilbank, Kris Blindert, Henk Hoekstra, Subhabrata Majumdar, Ricardo Demarco, and et al. Spectroscopic confirmation of a massive red-sequence-selected galaxy cluster atz= 1.34 in the sparcs-south cluster survey. *The Astrophysical Journal*, 698(2):1943–1950, Jun 2009. ISSN 0004-637X. doi: 10.1088/0004-637X/698/2/1943. URL https://doi.org/10.1088/0004-637x/698/2/1943.

Billy Wong, Yuan-Li Tiffany Chiu, Órla Meadhbh Murray, Jo Horsburgh, and Meggie Copsey-Blake. ‘Biology is easy, physics is hard’: Student perceptions of the ideal and the typical student across STEM higher education. *International Studies in Sociology of Education*, 0(0):1–22, 2022. doi: 10.1080/09620214.2022.2122532. URL https://doi.org/10.1080/09620214.2022.2122532.

Guy Worthey. Comprehensive stellar population models and the disentanglement of age and metallicity effects. *The Astrophysical Journal Supplement Series*, 95:107, Nov 1994. ISSN 0067-0049. doi: 10.1086/192096. URL https://ui.adsabs.harvard.edu/abs/1994ApJS...95..107W.

Chunliang Yang, Liang Luo, Miguel A. Vadillo, Rongjun Yu, and David R. Shanks. Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. *Psychological Bulletin*, 147:399–435, 2021. ISSN 1939-1455. doi: 10.1037/bul0000309.

Xiaohu Yang, H. J. Mo, and Frank C. van den Bosch. Galaxy groups in the sdss dr4. ii. halo occupation statistics. *The Astrophysical Journal*, 676:248–261, Mar 2008. ISSN 0004-637X. doi: 10.1086/528954. URL http://adsabs.harvard.edu/abs/2008ApJ...676..248Y.

Taniguchi Yoshiaki, Scoville N. Z, Sanders D. B, Mobasher B, Aussel H, Capak P, Ajiki M, Murayama T, Miyazak S, Komiyama Y, Shioya Y, and Nagao T. The hst cosmos project: Contribution from the subaru telescope. *Journal of The Korean Astronomical Society*, 38 (2):187–190, 2005. ISSN 1225-4614. doi: 10.5303/JKAS.2005.38.2.187. URL http://koreascience.or.kr/article/JAKO200502637760892.page.

Aeron Zentner, Raissa Covit, and Daniela Guevara. Exploring the relationship between grit and study habits among two-year college students. *SSRN*, Apr 2018. doi: 10.2139/ssrn.3171299. URL https://papers.ssrn.com/abstract=3171299.

Barry J. Zimmerman and Anastasia Kitsantas. Homework practices and academic achievement: The mediating role of self-efficacy and perceived responsibility beliefs. *Contemporary Educational Psychology*, 30(4):397–417, Oct 2005. ISSN 0361-476X. doi: 10.1016/j.cedpsych.2005.05.003. URL https://www.sciencedirect.com/science/article/pii/S0361476X05000329.