Kepler's Earth-like Planets Should Not Be Confirmed without Independent Detection: The Case of Kepler-452b

Fergal Mullally[1], Susan E. Thompson[1,2,3] , Jeffrey L. Coughlin[1,2] , Christopher J. Burke[1,2,4] , and Jason F. Rowe[5]
[1] SETI Institute, 189 Bernardo Avenue, Suite 200, Mountain View, CA 94043, USA; fergal.mullally@gmail.com
[2] NASA Ames Research Center, Moffett Field, CA 94035, USA
[3] Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD 21218, USA
[4] MIT Kavli Institute for Astrophysics and Space Research, 77 Massachusetts Avenue, 37-241, Cambridge, MA 02139, USA
[5] Dept. of Physics and Astronomy, Bishop's University, 2600 College Street, Sherbrooke, QC, J1M 1Z7, Canada

Abstract

We show that the claimed confirmed planet Kepler-452b (a.k.a., K07016.01, KIC 8311864) cannot be confirmed using a purely statistical validation approach. *Kepler* detects many more periodic signals from instrumental effects than it does from transits, and it is likely impossible to confidently distinguish the two types of events at low signal-to-noise. As a result, the scenario that the observed signal is due to an instrumental artifact cannot be ruled out with 99% confidence, and the system must still be considered a candidate planet. We discuss the implications for other confirmed planets in or near the habitable zone.

1. Introduction

Kepler's results will cast a long shadow on the field of exoplanets. The abundance of exo-Earths derived from *Kepler* data will dictate the design of future direct detection missions. An oft neglected component of occurrence rate calculations is an estimate of the reliability of the underlying catalog. As is frequently mentioned in *Kepler* catalog papers (Batalha et al. 2013; Burke et al. 2014; Mullally et al. 2015; Rowe et al. 2015; Coughlin et al. 2016; Thompson et al. 2018), not every listed candidate is actually a planet. Assuming they are all planets leads to an over-estimated planet occurrence rate (see, e.g., Section 8 of Burke et al. 2015). This problem is especially pronounced for small ($\lesssim 2 R_\oplus$), long-period (>200 days) candidates, where the signals due to instrumental effects dominate over astrophysical signals.

External confirmation of *Kepler* planets as an independent measure of reliability is therefore an important part of *Kepler*'s science. While progressively more sophisticated approaches to false positive identification have been employed by successive *Kepler* catalogs, it is not possible to identify every false positive with *Kepler* data alone. Independent, external measures of catalog reliability are a necessary step toward obtaining the most accurate occurrence rates, as well as defining a set of targets for which follow-up observations can be planned. Also, the intangible benefit of being able to point to specific systems that host planets, not just candidates, should not be understated. Kepler-452b is especially interesting in this regard. The target star is similar to the Sun, the orbital period is close to one Earth year, and its measured radius ($1.6 \pm 0.2 R_\oplus$) admits the possibility of a predominantly rocky composition (Rogers 2015; Wolfgang & Lopez 2015).

While small, short-period planets can be confirmed with radial velocities (see, e.g., Marcy et al. 2014), longer-period planets are hard to detect in this manner. Torres et al. (2011), building on work by Brown (2003) and Mandushev et al. (2005), introduced the technique of statistical confirmation, which used all available observational evidence (including high-resolution imaging to identify possible background eclipsing binaries) to rule out various false positive scenarios involving eclipsing binaries to the level that the probability the observed signal was a planet exceeds the probability of an eclipsing binary by a factor of 100 or more. Morton et al. (2016) used a simplified technique that could be tractably run on the entire candidate catalog, finding and confirming 1284 new systems. Lissauer et al. (2012) deduced that the presence of two or more planet candidates around a single star meant that neither was likely to be a false positive and first suggested that a false alarm probability of <1% was sufficient to claim confirmation of a candidate. Rowe et al. (2014) used Lissauer's argument to claim statistical confirmation of 851 new planets. Rowe et al. and Morton et al. together confirmed the vast majority of confirmed *Kepler* planets.

However, in this paper we argue that statistical confirmation of long-period, low signal-to-noise ratio (S/N) planets is considerably more challenging than previously believed. We use Kepler-452b (Jenkins et al. 2015) as a concrete example of a planet that should not be considered confirmed to illustrate our argument, but our argument may also apply to other statistically confirmed small planets (or candidates) in or near the habitable zone of solar-type stars. Our argument relies on two lines of evidence. First, the number of instrumental false positive (hereafter called false alarm) signals seen in *Kepler* data before vetting dominates over the number of true planet signals at long period and low S/N. Second, it is difficult (if not impossible) to adequately filter out enough of those false alarms while preserving the real planets, even with visual inspection. The sample of planet candidates in this region of parameter space is sufficiently diluted by undetected false alarms that it may be impossible to conclude that any single, small, long-period planet candidate is not an instrumental feature with a confidence greater than 99% without independent observational confirmation that the transit is real. Section 7.3 of the final *Kepler* planet candidate catalog (Thompson et al. 2018) discusses the tools and techniques necessary to study catalog reliability, and we draw heavily from that analysis.
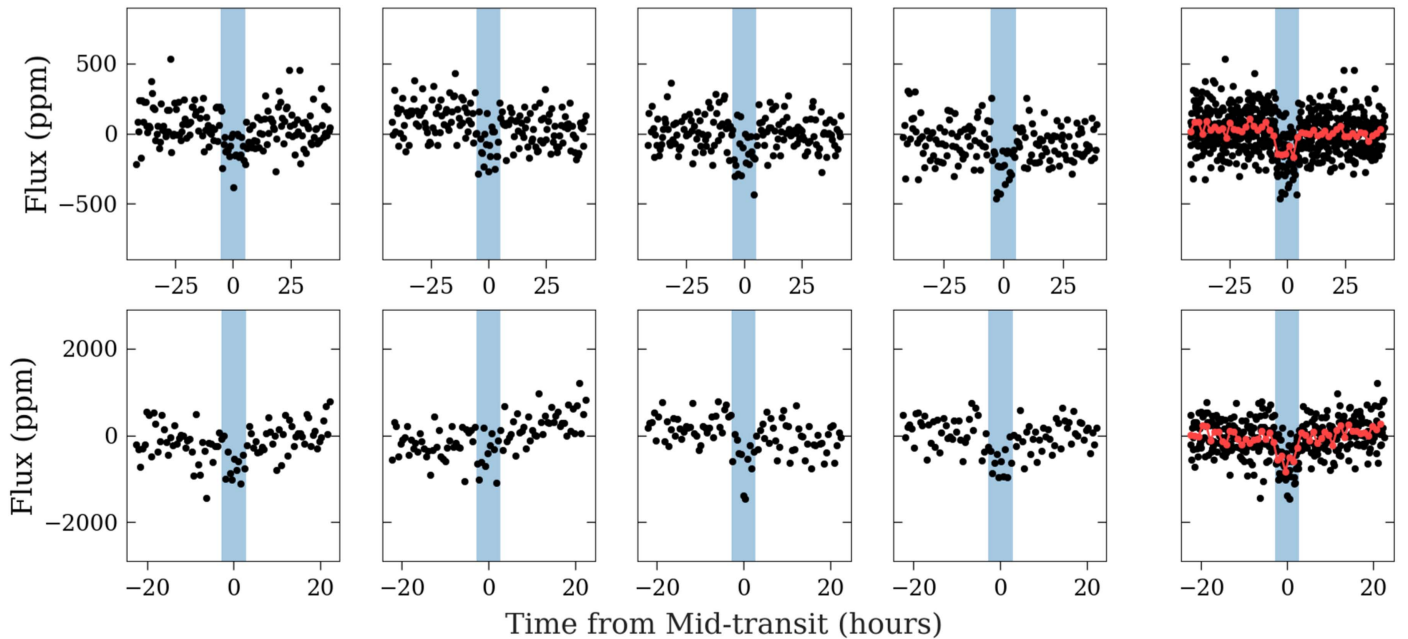
Figure 1. Each panel shows a single transit from a TCE discovered in one of our data sets. The timescale in each panel is hours from mid-transit, with the points the *Kepler* pipeline considers "in-transit" highlighted by the vertical blue bar. The panel at the far right combines all four events in a folded light curve (the red square symbols show the binned folded transit). One row shows transits from a known false alarm TCE found in the inversion run, the other from Kepler-452b. It is not clear from visual examination which one is a claimed planet and which one is a known systematic signal. (We provide the answer at the end of the acknowledgements).

2. The Limits of Statistical Confirmation

The basic approach for statistical validation is to compute the odds-ratio between the posterior likelihood that an observed signal is due to a planet, to the probability of some other source. The Blender approach (Torres et al. 2011), used by Jenkins et al. (2015) to confirm Kepler-452b, computes

$$\frac{P(\text{planet})}{P(\text{planet}) + P(\text{EB})} \tag{1}$$

where $P(\text{EB})$ is the posterior probability that the signal is due to an eclipsing binary, and the denominator sums to unity. Morton et al. (2016) expanded on this approach by fitting models of two types of non-transit signals to their data, effectively computing

$$\frac{P(\text{planet})}{P(\text{planet}) + P(\text{EB}) + P(\text{FA})} \tag{2}$$

where FA indicates a false alarm due to a non-astrophysical signal. Notably, they assume the priors on those false alarm models to be low, reflecting their belief that only a small fraction of their sample were non-astronomical signals.

For high S/N detections, or for candidates with periods between 50 and 200 days, $P(\text{FA})$ is indeed small and can essentially be ignored. However, Mullally et al. (2015) first noted the same is not true for candidates with periods longer than ~300 days and clustered close to the S/N threshold for detection (defined as Multiple Event Statistic, MES, >7.1). They urged caution in interpreting those candidates as evidence that there was an abundance of planets between 300 and 500 days.

The difficulty in distinguishing low signal-to-noise planets from false alarms is illustrated in Figure 1. One of the panels shows individual transits from Kepler-452b, and the other from a simulated data set mimicking the noise properties of *Kepler*

data (Coughlin 2017). (We summarize how simulated data were created in Section 3). Both are detected as TCEs (Threshold Crossing Events, or periodic dips in a light curve) by the *Kepler* pipeline (Twicken et al. 2016).

It is not obvious, even to the eye, which is the real planet, and which is the artifact. When planets and false alarms look so similar, there is no definitive test to distinguish the two. Any vetting process must trade-off between *completeness*, the fraction of bona-fide planets correctly identified (or the true positive rate), and *effectiveness*, the fraction of false alarms correctly identified (or the true negative rate).

Most of these false alarms are at long period and low S/N. At long periods, a few systematic "kinks" in a light curve can often line up, sometimes with low-amplitude stellar variability, to produce a signal that looks plausibly like a transit. These chance alignments are common for three to four transit TCEs, but their frequency declines with five or more transits. Candidates with five or more transits are less likely to be false alarms because the probability of getting five events to line up periodically is smaller (see Figure 9 of Thompson et al. 2018, for more details). The combination of less than ideal effectiveness at low S/N (because transits and false alarms are difficult to tell apart), and an over-abundance of false alarms at long period, leads to a surprisingly low catalog *reliability*, or the fraction of claimed planet candidates actually due to transits and not due to instrumental effects.

3. The Reliability of Earth-like Planet Candidates

The catalog of Thompson et al. (2018) uses a number of tests, collectively called the Robovetter (Thompson et al. 2015; Coughlin et al. 2016; Mullally et al. 2016), to identify false positive and false alarm TCEs. Each test must be tuned to balance the competing demands of completeness (that as many true planets as possible are passed by each test) and effectiveness (that as many false positives of the targeted type
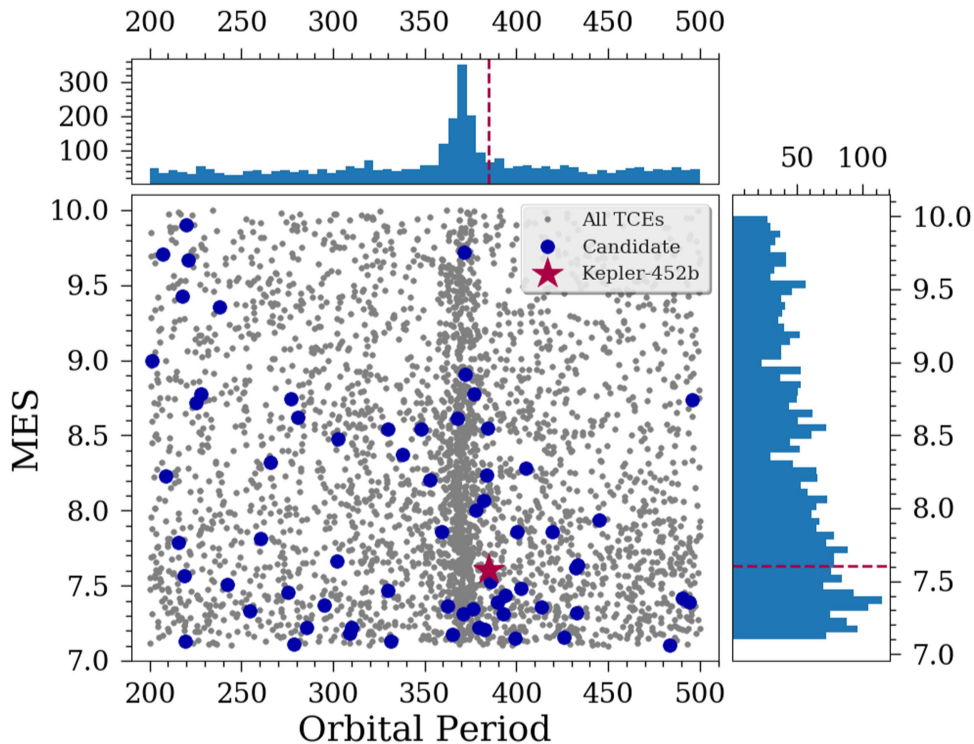
Figure 2. Distribution of planet candidates (large blue circles) and non-candidate TCEs (small gray circles) from the DR25 catalog. The marginal histograms in period (and MES) are shown in the top (and right) panels. The rejected population is dominated by TCEs caused by instrumental systematics in the light curves. Because there are so many of these instrumental false alarms, the small number that slip through vetting is a significant fraction of the planet candidate catalog in this region of parameter space. The reliability of any individual planet candidate (or the probability it is due to an astrophysical event and not an instrumental feature) is too low to allow any single candidate to be confirmed without additional observational evidence. The location of Kepler-452b is indicated by the red star.

as possible are rejected by the test). The large number of tests applied means that each test must be tuned for high completeness, or the number of bona-fide planets incorrectly rejected quickly becomes unacceptably large.

To test the completeness, effectiveness, and reliability of the planet candidate catalog, the *Kepler* mission produced three synthetic data sets. Christiansen ([2017](#)) measured completeness by injecting artificial transits into *Kepler* light curves and seeing how many were recovered by the *Kepler* pipeline. Producing a sample of transit free light curves with realistic noise and systematic properties was more difficult. The mission tried two approaches, each with their own strengths and weaknesses, as discussed in more detail in Thompson et al. ([2018](#)). The first approach (called "inversion") was to invert the light curves (so that a flux decrement becomes an excess). The second approach was to shuffle the light curves in time, or "scramble" them, to remove the phase coherence of the real transits. All simulated data sets are available for download at the NASA Exoplanet Archive.[6]

In the discussion that follows, we will consider the performance of the Robovetter for TCEs with periods in the range 200–500 days and MES (Multiple Event Statistic) or the S/N of detection of the TCE in TPS (Christiansen et al. [2012](#)) less than 10. This parameter space comfortably brackets Kepler-452b and contains sufficient TCEs to perform a statistical analysis. For brevity, we will refer to this sample as the Earth-like sample because it encompasses the rocky, habitable-zone planets *Kepler* was designed to detect (Koch et al. [2010](#)). We assume for simplicity that completeness and

effectiveness are constant across this box, although in reality the Robovetter performance declines with increasing period and decreasing MES.

Thompson et al. ([2018](#)) ran the Robovetter on a union of TCEs produced by inversion and the first scrambling run (Coughlin [2017](#)).[7] They measure effectiveness as the fraction of these simulated TCEs correctly identified as false alarms in the given parameter space. The Robovetter performs extremely well on these simulated TCEs, with an effectiveness of 98.3%. For every 1000 false alarms, the Robovetter correctly identifies 983 of them and mis-classifies only 17 as planet candidates.

The original observations (i.e., the real data) produced a huge number of (mostly false alarm) Earth-like TCEs, as shown in Figure [2](#). The Robovetter examined 3341 TCEs with periods between 200 and 500 days and MES < 10 in the original data for DR25. Of these, 3274 are identified as not-transit-like (i.e., unlikely to be either due to a transit or eclipse; not-transit-like is the label given by the Robovetter to TCEs without a realistic transit shape). The measured effectiveness suggests that an additional 56 false alarm TCEs were mistakenly labeled as planet candidates.

There are only 67 planet candidates in this region of parameter space, but if 56 are likely uncaught false alarms, then only 11 are bona-fide transit signals. The reliability of the sample (the fraction of planet candidates that are actually due to an astrophysical event such as a transit or eclipse) is then $11/67 = 16.0\%$. This low reliability means that there is only a 16.0% chance that any one planet candidate is actually due to a planet transit, and not a systematic event, far lower than the

---

[6] Simulated data available at https://exoplanetarchive.ipac.caltech.edu/docs/KeplerSimulated.html.

[7] Three scrambling runs were created, but we only use the first one for consistency with Thompson et al. ([2018](#)).

99% threshold for statistical confirmation suggested by Lissauer et al. (2012), or the 99.76% level confidence claimed by Jenkins et al. (2015) for Kepler-452b. The astrophysical false positive rate is clearly only part of the puzzle. For Earth-like candidates, the instrumental false alarm rate must be included in order to statistically confirm planet candidates.

Worse, our ability to estimate reliability is limited by the small numbers of simulations in the parameter space of interest. If we assume the number of uncaught false alarms is drawn from a Poisson distribution, our uncertainty in the number of false alarms is 56 ± 7.5. The uncertainty in the reliability is then $16 \pm 11\%$. Not only do we fail to measure a false positive probability of $<1\%$, our current data prevents us from measuring the false positive probability with a precision of $<1\%$.

Our analysis is slightly over-simplified for clarity. A more detailed calculation would correct the number of measured false alarms for the small number of bona-fide transits mistakenly rejected as false alarms. It should also consider the reasons the Robovetter failed simulated transits, as some rejected TCEs were rejected as variable stars, eclipsing binaries etc. However, none of these improvements can hope to raise the measured reliability to $>99\%$. The error in the computed reliability due to these simplifications is likely small compared to the systematic error in measured effectiveness due to the different abundances of false alarms in the original and simulated data. In particular, assuming constant reliability across the sample slightly over-estimates the reliability of the Kepler-452b, as the effectiveness decreases with decreasing numbers of transits.

The fidelity of the simulated false alarm populations is discussed in some detail in Section 4.2 of Thompson et al. (2018). Combining the TCEs from the inversion run and one scrambling run produces a distribution of simulated TCEs as a function of period that matches the observed distribution quite well, but not exactly. Because the sources of the observed false alarm TCEs are not precisely understood, we cannot rule out the possibility that the simulated data sets are over-producing a certain kind of systematic signal that the robovetter is good at detecting, while under-producing a different kind of signal the robovetter struggles with, thereby overestimating our measured reliability. The converse may also be true. The similarity in the period distributions implies that the simulated populations are well matched to the observed ones, but there is still a small systematic error in our measured reliability that we do not yet know how to measure.

Our argument does not directly apply to the majority of statistically confirmed *Kepler* planets. Transits with high S/N are much less likely to be systematics and much easier to identify as such if they are. The two largest contributions to the set of statistical confirmations, Rowe et al. (2014) and Morton et al. (2016), both restrict their analysis to candidates detected with an S/N > 10. Kepler-452b has the lowest S/N of all the long-period planets with claimed confirmations, and thus has the most tenuous claim to confirmation. However, 99% reliability is a high threshold to reach at low S/N, and the confidence with which many of the other small, habitable-zone planets were claimed to be confirmed is likely over-stated. We leave the analysis of these systems as future work.

3.1. Raising Reliability

In this section, we discuss some refinements to our analysis that increase the estimated reliability of some of the candidates in the Earth-like sample. However, none of these refinements increase the reliability to the desired 99% level.

3.1.1. Improved Parameter Space Selection

Following Thompson et al. (2018), if we restrict our analysis to TCEs around main-sequence stars (4000 K $< T_{\mathrm{eff}} <$ 7000 K and $\log g > 4.0$), we are looking at signals observed around photometrically quieter stars. The Robovetter performs much better for these stars, and the measured reliability increases to $\approx$50%. Further restricting the sample to exclude the period range with the highest false alarm rate (360–380 days, as shown in Figure 2) increases the reliability to only 54%.

This is close to the ceiling for reliability that can be achieved by analyzing smaller parameter spaces. Choosing narrower slices of parameter space can produce small gains in effectiveness but comparatively larger changes in the numbers of candidates. The measured reliability increases slowly (if at all) and becomes noisier, as the effects of small number statistics begin to dominate.

3.1.2. High-score TCEs

Thompson et al. (2018) include a score metric, which attempts to quantify the confidence the Robovetter places in its classification of a TCE as a planet candidate or a false alarm. The score is not the probability that a TCE is a planet. High scores (>0.8) indicate strong confidence that a TCE is a candidate, while low scores (<0.2) indicate high confidence that a TCE is either a false positive or false alarm. An intermediate score indicates that the TCE was close to the threshold for failure in one or more of the Robovetter metrics. Kepler-452b has a score of 0.77.

If we repeat our analysis on the small, long-period TCEs around FGK dwarves, but require a score $\geqslant 0.77$ to consider a TCE a candidate, we find an effectiveness in simulated data of 99.97%, and a reliability of the catalog (between periods of 200–500 days and MES < 10) of 92% based on nine planet candidates. While this reliability is much stronger than before, it is still some distance from the 99% reliability required to claim statistical confirmation. The effectiveness estimate is also based on a small number of candidates and simulated TCEs passing the Robovetter. If the number of incorrectly identified simulated TCEs changes by just one, the measured reliability changes by 8%. Even if the measured reliability for high-score TCEs was $>99\%$, the uncertainty in the measured reliability casts sufficient doubt to counter any claimed confirmation.

3.1.3. Multiple Planet Systems

Lissauer et al. (2012) first noted that the presence of multiple planet candidates around a single target star was strong evidence that all candidates in that system were planets and not eclipsing binaries. The argument is essentially that while planet detections and eclipsing binary detections are both rare, multiple detectable planets in a single system are much more likely than an accidental geometric alignment between a planet host system and a background eclipsing binary. As recognized by Rowe et al. (2014), a similar argument does not automatically apply to false positives. Lissauer assumed that

planetary systems and eclipsing binaries are distributed uniformly across all targets. Systematic signals come from a variety of sources, some of which are concentrated in specific regions of the focal plane, or types of stars. Some targets are considerably more likely to see multiple false alarm TCEs than a uniform distribution would predict. It may be possible to apply the Lissauer analysis once a richer model of the false alarm distribution is known.

3.1.4. Relying on Earlier Catalogs

The reliability of a candidate may be higher if it appears in multiple *Kepler* catalogs. The TPS algorithm underwent extensive development between the Q1-Q16 catalog of Mullally et al. (2015), the DR24 catalog of Coughlin et al. (2016), and the DR25 catalog (the three catalogs to include 4 years of *Kepler* data). The different catalogs could, in principle, be used as quasi-independent detections at the low S/N limit. Unfortunately, due to the lack of diagnostic runs on the earlier catalogs (reliability is unmeasured for DR24, and neither reliability nor completeness for Q1-Q16), it is impossible to quantify the improved reliability of a candidate in this manner. For the specific case of Kepler-452b, it was detected in both the DR24 and DR25 catalogs, but was not detected in Q1-Q16, even though all four transits were observed.

4. Discussion and Conclusion

We show statistical validation is insufficient to confirm Kepler-452b at the 99% level, and that Kepler-452b should no longer be considered a confirmed planet. Although use of this threshold to claim confirmation is somewhat arbitrary, it has seen widespread adoption in the community. To our knowledge, there are no claims in the literature of the claimed confirmation of a transiting planet with confidence less than this threshold.

Our simplest calculation argues that there is only a 16% chance that the detected signal on the target star is due to a transit, while our most optimistic calculation sets the probability at 92%. Even this most optimistic assessment is still an order of magnitude less confident than our threshold, and nearly two orders less than that claimed in the discovery paper (99.76%, Jenkins et al. 2015). While the exact choice of threshold to claim confirmation may be a matter of taste and convention, setting the threshold low enough to admit Kepler-452b would result in significant contamination of the confirmed planet catalog with false alarms and negate the goals of creating a confirmed planet list.

It may be possible to devise a new statistical approach that validates Kepler-452b. However, given the difficulties of creating simulated data that reproduces the time-varying, non-Gaussian noise properties of *Kepler* data, and the considerable development effort already invested in the Robovetter, we are doubtful such a method will be devised in the near future. Instead, we advocate that the transit events should be directly confirmed by independent observations, such as the Hubble follow-up of Kepler-62 by Burke (2017).

The un-confirmation of Kepler-452b is interesting in its own right, given the similarity between its measured parameters and those of the Earth, but our result has implications for other long-period confirmed *Kepler* planets. Torres et al. (2017) claimed to statistically confirm six habitable-zone candidates originally detected with long-period and MES < 12. These planets were all detected at higher S/N than Kepler-452b and are therefore less likely to be caused by an instrumental artifact. However, the lack of precision for our reliability estimates in this work does raise a concern for these other systems. The burden of proof for claiming that a planet exists lies in demonstrating that the probability of all other scenarios is definitely less than the chosen threshold. If the likelihood of a transit signal is drawn from a distribution (i.e., a posterior) that overlaps significantly with the threshold for acceptance, then that burden of proof is not met. For Kepler-452b, the uncertainty in our measurement of reliability is an order of magnitude higher than what is required for confirmation.

The inability to confirm individual planet candidates by statistical techniques should not be considered a failure of the *Kepler* mission (or of the statistical techniques themselves, which are applicable when their underlying assumptions remain valid). *Kepler*'s goal was to establish the frequency of Earth-like planets in the Galaxy, and the conclusion that at least 2% (and possibly as many as 25%) of stars host an Earth-size planet in the shorter-period range of 50–300 days (Burke et al. 2015) is not strongly affected by this result. The reliability of the *Kepler* sample in this regime is much higher, and sensitivity of occurrence rate calculations is less sensitive to catalog reliability. It will be necessary to account for catalog reliability to properly extend occurrence rate calculations out toward the longer periods that encompass the habitable zone around G-type stars.

Facility: Kepler.

ORCID iDs

Susan E. Thompson ⑩ https://orcid.org/0000-0001-7106-4683
Jeffrey L. Coughlin ⑩ https://orcid.org/0000-0003-1634-9672
Christopher J. Burke ⑩ https://orcid.org/0000-0002-7754-9486
Jason F. Rowe ⑩ https://orcid.org/0000-0002-5904-1865

References

Batalha, N. M., Rowe, J. F., Bryson, S. T., et al. 2013, ApJS, 204, 24
Brown, T. M. 2003, ApJL, 593, L125
Burke, C. 2017, Completing Kepler's Mission to Determine the Frequency of Earth-like Planets, HST Proposal 15129
Burke, C. J., Bryson, S. T., Mullally, F., et al. 2014, ApJS, 210, 19
Burke, C. J., Christiansen, J. L., Mullally, F., et al. 2015, ApJ, 809, 8
Christiansen, J. L. 2017, Kepler Science Document, KSCI-19110-001, https://exoplanetarchive.ipac.caltech.edu/docs/KSCI-19110-001.pdf

Christiansen, J. L., Jenkins, J. M., Caldwell, D. A., et al. 2012, PASP, 124, 1279

Coughlin, J. L. 2017, Kepler Science Document, KSCI-19114-002, https://exoplanetarchive.ipac.caltech.edu/docs/KSCI-19114-002.pdf

Coughlin, J. L., Mullally, F., Thompson, S. E., et al. 2016, ApJS, 224, 12

Jenkins, J. M., Twicken, J. D., Batalha, N. M., et al. 2015, AJ, 150, 56

Koch, D. G., Borucki, W. J., Basri, G., et al. 2010, ApJL, 713, L79

Lissauer, J. J., Marcy, G. W., Rowe, J. F., et al. 2012, ApJ, 750, 112

Mandushev, G., Torres, G., Latham, D. W., et al. 2005, ApJ, 621, 1061

Marcy, G. W., Isaacson, H., Howard, A. W., et al. 2014, ApJS, 210, 20

Morton, T. D., Bryson, S. T., Coughlin, J. L., et al. 2016, ApJ, 822, 86

Mullally, F., Coughlin, J. L., Thompson, S. E., et al. 2015, ApJS, 217, 31

Mullally, F., Coughlin, J. L., Thompson, S. E., et al. 2016, PASP, 128, 074502

Rogers, L. A. 2015, ApJ, 801, 41

Rowe, J. F., Bryson, S. T., Marcy, G. W., et al. 2014, ApJ, 784, 45

Rowe, J. F., Coughlin, J. L., Antoci, V., et al. 2015, ApJS, 217, 16

Thompson, S. E., Coughlin, J. L., Hoffman, K., et al. 2018, ApJS, 235, 38

Thompson, S. E., Mullally, F., Coughlin, J., et al. 2015, ApJ, 812, 46

Torres, G., Fressin, F., Batalha, N. M., et al. 2011, ApJ, 727, 24

Torres, G., Kane, S. R., Rowe, J. F., et al. 2017, AJ, 154, 264

Twicken, J. D., Jenkins, J. M., Seader, S. E., et al. 2016, AJ, 152, 158

Wolfgang, A., & Lopez, E. 2015, ApJ, 806, 183