How doctors conceptualise *P* values

A mixed methods study

Chun Wah Michael Tam,
Abeer Hasan Khan, Andrew Knight, Joel Rhee, Karen Price, Katrina McLean

Background and objectives
Researchers and clinicians have been criticised for frequently misinterpreting and misusing *P* values. This study sought to understand how general practitioners (GPs) in Australia and New Zealand conceptualise *P* values presented in the manner typically encountered in a medical publication.

Methods
This mixed-methods study used quantitative and qualitative questions embedded in an online questionnaire and delivered through an Australian and New Zealand GP-specific Facebook group in 2017. It included questions that elaborated on the participant's conceptualisation of '*P* = 0.05' within a scenario and tested their *P* value interpretation ability and confidence.

Results
There were 247 participants who completed the questionnaire. Participant conceptualisations of *P* values were described using six thematic categories. The most common (and erroneous) conceptualisation was that *P* values numerically indicated a 'real-world probability'. No demographic factor, including research experience, seemed associated with better interpretation ability. A confidence–ability gap was detected.

Discussion
P value misunderstanding is pervasive and might be influenced by a few central misconceptions. Statistics education for clinicians should explicitly address the most common misconceptions.

IN 2016, the American Statistical Association (ASA), the international peak body for professional statisticians, took an unprecedented step by publishing a policy statement detailing the correct context, process and purpose in using *P* values, claiming that they were 'too often misunderstood and misused in the broader research community'.[1] In this open rebuke, the ASA defined the *P* value as:

> *... the probability under a specified statistical model that a statistical summary of the data (eg the sample mean difference between two compared groups) would be equal to or more extreme than its observed value.*[1]

There have been numerous criticisms over the past century concerning the misinterpretation and misuse of *P* values, together with warnings against making unjustified inferences from data.[1–4] Specifically, *P* values are often misinterpreted as providing far stronger evidence than is actually the case,[2] which not only has a harmful impact on the understanding of medical research, but potentially on the delivery of patient care. Research on the statistical knowledge of doctors is consistent with these concerns. In one study, only 42% of medical residents were able to correctly define *P* values in a simple true/false question;[5] another study identified a substantial gap between clinician confidence in interpreting *P* values and their actual ability.[6] Although the literature seems to confirm that clinicians frequently interpret *P* values incorrectly, there is much less evidence on how and why they (mis)conceptualise *P* values. This is important, considering the reported confidence–ability gap and the ubiquity of *P* values in health literature.

Our study sought to describe and categorise what and how clinicians (in this case, Australian and New Zealand general practitioners [GPs]) conceptualise *P* values presented in the manner that it is typically encountered in a medical publication. These results may help inform the provision of targeted statistics education to clinicians.

Methods

This was a mixed methods study using quantitative and qualitative questions embedded in a short online questionnaire, conducted in mid-2017. The questionnaire was delivered through an Australian and New Zealand Facebook group, 'GPs Down Under' (GPDU). At the time of the study, GPDU had approximately 4000 members, all authenticated as either GPs or general practice registrars (vocational trainees).[7,8]

Questionnaire

The questionnaire was delivered through SurveyMonkey, a web-based platform, and linked in a post in the GPDU Facebook group. It remained open for data collection until three consecutive days of no responses, which was at approximately six weeks.

Questions related to participants' personal and professional demographics, research and teaching experience, followed by sequential targeted questions designed to:

1. qualitatively elaborate on the participant's conceptualisation of '*P* = 0.05' within a scenario
2. test *P* value interpretation ability (whether they identified a common erroneous interpretation) with a dichotomous choice question
3. measure their confidence in their answer with a Likert scale (Box 1).

Analyses

Qualitative

Participants' free-text responses to the 'qualitative' question (Box 1) were analysed to identify thematic categories. We undertook this analysis from a constructivist perspective (that the researcher constructs knowledge through their interaction with the data, with an emphasis on **understanding** the phenomena of the participant responses),[9] with the framing that each response would include the participant's conceptualisation of *P* values. NVivo 11 software was used to code the data. The first step in the analysis process was assessing the responses line by line and identifying 'in vivo' codes – verbatim statements from the participants. These codes were abstracted to higher level concepts, and then finally to categories/themes. Interim categories/themes were analysed and discussed by AHKh and MT until consensus was reached, then shared with the remainder of the research team for discussion and approval.

Quantitative

Descriptive data analysis of participant demographics and responses to the 'dichotomous choice' and 'confidence' questions (Box 1) was conducted using IBM SPSS Statistics 24 and Microsoft Excel. Exploratory analyses were conducted using independent sample t-tests, Mann-Whitney U tests and Pearson Chi-squared tests to examine the effects of continuous, ordinal and categorical variables respectively on participant responses to the 'dichotomous choice' question and to the reported confidence in their interpretations.

Mixed methods

Following the construction of the thematic categories, each participant response to the 'qualitative question' was assigned one category. If there was more than one concept present in the response, the predominant one was chosen. AHKh and MT undertook this process together in one sitting until consensus was reached for all participant responses. The enumerated data of how participants conceptualised '*P* = 0.05' in terms of the thematic categories were explored for associations with demographic factors, responses to the 'dichotomous choice' and 'confidence' questions.

Ethics approval

This study was approved by the UNSW Sydney Human Research Ethics Advisory Panel (reference number: HC17503).

Box 1. The three targeted questions in the questionnaire

The 'qualitative' question

The scenario:

A study comparing a new antihypertensive to an older agent, with blood pressure as the primary outcome, is published in a medical journal. In the article's conclusion, the authors claim that, the new drug was superior to the old drug at lowering blood pressure (*P* = 0.05).

Question:

In no more than 2–3 sentences, describe what '*P* = 0.05' means in the above statement.
[Free text response]

The 'dichotomous choice' question

[The text of 'the scenario' is repeated in full, with the additional statement]

A reader makes the following interpretation:

'This means that there is a 5% probability that this result is due to chance alone, or, there is a 95% probability that the conclusion is true.'

Question:

Please select the option that BEST MATCHES your understanding of *P* values:

- The reader's interpretation is mostly FALSE [note: correct response]
- The reader's interpretation is mostly TRUE

The 'confidence' question

Question:

Please indicate how CONFIDENT you are of your answer:
Not at all – Slightly – Somewhat – Very – Entirely
[Labelled 5-point Likert scale]

Results

Participants

The questionnaire was open from 5 July 2017 to 18 August 2017. There were 247 respondents who completed the questionnaire out of 272 who started (91% completion rate). In brief, a preponderance of the respondents were female, Australian residents, Fellows of The Royal Australian College of General Practitioners, and reported some research and teaching experience (Table 1).

Qualitative results

Nine participants submitted a blank response or indicated that they could not provide an answer to the 'qualitative question'. Using the data from the remaining 238 participants, six thematic categories emerged from their free-text responses.

Real-world probability

Many respondents conceptualised the *P* value as numerically indicating the natural probability of some phenomenon – for instance, a 95% or 5% chance of the truth or falsity of a hypothesis in the real world.

> *We are 95% sure that the new drug is superior to the old drug. Or there is 5% chance the drugs perform equally well.* [Participant 98]

Some participants seemed to comfortably include this (mis)conceptualisation with other technical concepts of statistics:

> *Assuming a Gaussian distribution and appropriate sample size, this means*

> *that the difference (better than control) with new drug is due to a real effect with a 95% probability (ie less than 1/20 chance of this effect being due to chance or error not a real effect, assuming normal distribution, control, adequate sample size).* [Participant 103]

Threshold reasoning

Many respondents saw '*P* = 0.05' as indicating a cut-off in their interpretation of whether there was evidence of an empirical effect. Interestingly, interpretations from both sides of the threshold were described – superiority:

> *That the new drug was statistically significant to show superiority over the old drug. The new drug is better than the old drug.* [Participant 138]

... and not superior:

> *As it is not less than 0.05, then there is not a statistically significant difference. Therefore the new drug can not be considered superior to the old drug based on this study.* [Participant 46]

Statistical significance

Many participants described '*P* = 0.05' using the words 'statistical significance'. However, some participants seemed to have conceptualised this in a limited, self-referential manner – that this was the explanation in and of itself:

> *P = 0.05 means that the result is statistically significant.* [Participant 179]

There was some indication that these words remain with the decay in statistical knowledge:

> *... can't remember a single thing, other than it means it is statistically significant in a research approved way.* [Participant 236]

Within the context of null-hypothesis statistical testing

A minority of participants conceptualised *P* values explicitly within the framework of null-hypothesis statistical testing. Of all the categories, this was the one that

Table 1. Demographics of survey respondents

Total number of participants	n	247
Gender	Male – n (%)	83 (33.6)
	Female – n (%)	163 (66.0)
	Pefer not to say – n (%)	1 (0.4)
Age (years)	Mean (range), SD	39.6 (25–67), 8.9
Country of residence	Australia – n (%)	239 (96.8)
	New Zealand – n (%)	8 (3.2)
Year of graduation (medical school)	Median (range)	2005 (1973–2015)
Country of medical degree attainment	Australia – n (%)	203 (82.2)
	New Zealand – n (%)	8 (3.2)
	United Kingdom – n (%)	15 (6.1)
	Other countries – n (%)	21 (8.5)
Years working in general practice	Mean (range), SD	9.7 (0 to 40), 8.8
Specialist general practitioner or general practice registrar (vocational trainee)	Specialist GP – n (%)	195 (78.9)
	General practice registrar – n (%)	52 (21.1)
Specialist general practitioner qualification type*	FRACGP – n (% of all)	176 (71.3)
	FRNZCGP – n (% of all)	9 (3.6)
	FACRRM – n (% of all)	11 (4.5)
	FARGP – n (% of all)	7 (2.8)
	Other equivalent – n (% of all)	10 (4.0)
Highest postgraduate degree awarded	Primary medical degree – n (%)	111 (44.9)
	Graduate certificate – n (%)	10 (4.0)
	Graduate diploma – n (%)	68 (27.5)
	Masters (coursework) – n (%)	38 (15.4)
	Masters (research) – n (%)	8 (3.2)
	Doctorate – n (%)	12 (4.9)
Teach or supervise learners†	Yes – n (%)	177 (71.7)
	No – n (%)	70 (28.3)
Any research experience‡	Yes – n (%)	203 (82.2)
	No – n (%)	44 (17.8)
Been a researcher§	Yes – n (%)	92 (37.2)
	No – n (%)	155 (62.8)

**Participants can hold more than one qualification †Including medical students, hospital doctors-in-training, general practice registrars, or specialist general practitioners*
‡Includes having been a research participant, or recruited patients for a study
§Including having been a primary investigator, coinvestigator, received a research grant, and/or undertaken a higher degree by research
FARGP, Fellowship in Advanced Rural General Practice; FACRRM, Fellow of the Australian College of Rural and Remote Medicine; FRACGP, Fellow of the Royal Australian College of General Practitioners; FRNZCGP, Fellow of the Royal New Zealand College of General Practitioners

most aligned with the actual definition of *P* values:

> *'P = 0.05' in this study means that, if there was really no difference between the old and the new antihypertensives, then there would only be a 5% chance that we would have seen a blood pressure difference between the two drugs as large as was seen.* [Participant 30]

No meaningful interpretation

This group of participants indicated that *P* could not be interpreted in the question:

> *It doesn't have a meaning in this context ...* [Participant 164]

> *... statement is misleading. P value here suggests result is statistically significant but need to clarify more before claiming this.* [Participant 196]

Study quality

These participants conceptualised *P* in the question as indicating the quality of the study itself, rather than necessarily referring to the results. For instance:

> *Result is likely to be real rather coincidental. A well powered study.* [Participant 73]

Descriptive and exploratory quantitative results

Dichotomous choice question responses

'Mostly FALSE' was the correct response to the question; 'this means that there is a 5% probability that this result is due to chance alone, or, there is a 95% probability that the conclusion is true'.[2,10] Only 29% (95% confidence interval: 24, 35) or 72/247 respondents answered this question of *P* value interpretation ability correctly.

No demographic factor seemed associated with better performance (giving the correct response), including having been a researcher (yes versus no: 28% versus 30%, X^2, $P = 0.89$) or having a higher degree by research (yes versus no: 30% versus 29%, X^2, $P = 1.0$).

Confidence

Conversely, participant confidence in their response to the 'dichotomous choice' question seemed strongly associated with several factors.

- Men were more confident than women (Figure 1).
- Those with research experience were more confident (Mann-Whitney U, $P = 0.009$).

Notably, there was a confidence–ability gap – self-reported confidence was not associated with better performance in the 'dichotomous choice' question (Mann-Whitney U, $P = 0.132$).

Mixed methods results

The two most common conceptualisations of '$P = 0.05$' were 'real-world probability' and 'threshold reasoning', comprising 83% of responses (Figure 2).

Those who gave the incorrect response to the 'dichotomous question' on the interpretation of '$P = 0.05$' were much more likely to have given a prior free-text response that was categorised to 'real-world probability' (67% versus 24%; Figure 3).

Discussion

Our study identified six ways clinicians conceptualise '$P = 0.05$' when encountered in a short statement designed to mimic the concluding statement in a medical journal abstract. The predominant conceptualisation of over half the respondents is that the *P* value numerically represents the natural probability of something in the real world – typically the 'null hypothesis' or that the 'result is due to chance'. In designing the questionnaire, we had predicted that this might be a common misconception, so the 'dichotomous choice' question asked participants explicitly whether this real-world probability interpretation of the *P* value was true (Box 1). Most respondents (71%) answered in the affirmative and, therefore, incorrectly.

Although it may be tempting to simply attribute this result to clinician statistical innumeracy, it might not be the best explanation for this finding. Rather than ignorance, the fact that the majority were mistaken suggests the presence of an active and pervasive misunderstanding. Notably, there was no evidence of participants who would have likely received postgraduate statistics training performing better. Few participants conceptualised '$P = 0.05$' within the context of null-hypothesis statistical testing as the major concept in their free-text responses.

We propose that probabilities tend to be intuitively understood in concrete and absolute terms – that these apply to 'real-world' phenomena. For instance, a recent systematic review on evidence-based risk communication found that expressing probabilities as event rates or natural frequencies, and using absolute risks to describe risks and benefits, improved patient understanding.[11] *P* values are a summary of a statistical model – there is no direct numerical interpretation of their value in the concrete real world. As such, they are intrinsically counter-intuitive. Decay in statistical knowledge may lead clinicians to use the intuitive, but erroneous, 'real-world probability' conceptualisation of *P* values. As this appears to be the predominant conceptualisation, GPs may never encounter dissonance and may become increasingly comfortable with their interpretations. This could be an explanation for the confidence–ability gap we observed, consistent with prior research.[6] The unfortunate effect is that although the confident are no more likely to be wrong, they are much more likely to be **confidently** wrong.

Other authors have previously published excellent lists of 12[2] and 25[10] common misconceptions in interpreting *P* values in the academic literature. These can be overwhelming, especially for someone who confidently holds an intuitive but mistaken understanding of *P* values. What our results suggest is that priority should be placed on addressing two major misconceptions – first that *P* values refer to a 'real-world probability', and second that *P* values should be interpreted using 'threshold reasoning'.

Given their counterintuitive nature, ensuring that all clinicians can correctly define *P* values may be a lost cause. We have consciously chosen to not attempt to provide our own 'simple' definition of *P* values in this paper beyond the ASA's informal definition.[1] Instead, we wonder whether

introducing rules of thumb,[12] with a focus on explicitly countering these two common misconceptions (ie emphasising that *P* values **are not** real-world probabilities and **should not** be interpreted using thresholds), may help reduce the misinterpretation and misuse of *P* values. Furthermore, reducing the emphasis on 'statistical significance' and concentrating on the effect size estimate and its imprecision may improve the interpretation of results.[2]

Strengths and limitations

In this study, we were able to collect data on how clinicians conceptualise *P* values in context in their own words. We had a sufficiently large response such that we had ample free-text data for the construction of the thematic categories. By 'mixing in' the quantitative elements of the study with the qualitative analysis, we were able to build a more coherent description of the phenomena. Our questionnaire was carefully designed to ask participants for a free-text response to the 'qualitative question' first to avoid the risk of a biased response due to priming.

A significant limitation is that it is probable that the respondents to this study were not representative of Australian GPs. Of the approximate 4000 members of the GPDU Facebook group, only a fraction participated. The participant demographics (Table 1) suggest that GPs who received their medical education internationally are underrepresented.

As the qualitative analyses were performed on short free-text responses, it is possible that we are missing depth. We opted not to design this study with interviews because of limited resources. Lastly, the inferential quantitative analyses performed in this study should not be considered definitive. They need to be interpreted as exploratory and hypothesis-generating, and used to help construct the mixed methods analysis only.

Implications for general practice

- GPs and general practice registrars of all levels of research experience commonly conceptualise *P* values as indicating a 'real-world probability'. This interpretation is incorrect.
- Consistent with prior research pertaining to other clinicians, there may be a substantial confidence–ability gap in interpreting *P* values.
- Statistics education for clinicians should consider explicitly addressing the most common misconceptions.

Authors

Chun Wah Michael Tam BSc (Med), MBBS, MMH (GP), FRACGP, Staff Specialist, Academic Primary and Integrated Care Unit, South Western Sydney Local Health District and Ingham Institute of Applied Medical Research; Conjoint Senior Lecturer, School of Public Health and Community Medicine, UNSW Sydney, NSW. m.tam@unsw.edu.au

Abeer Hasan Khan BMed, medical student, UNSW Medicine, Sydney, NSW

Andrew Knight MBBS, MMedSci, FRACGP, FAICD, Acting Director, Academic Primary and Integrated Care Unit, South Western Sydney Local Health District and Ingham Institute of Applied Medical Research; Conjoint Senior Lecturer, School of Public Health and Community Medicine, UNSW Sydney, NSW

Joel Rhee BSc (Med), MBBS (Hons), GCULT, PhD, FRACGP, Associate Professor in General Practice, Graduate School of Medicine, University of Wollongong, NSW

Karen Price MBBS, Dip Adolescent Medicine, FRACGP, Clinical academic, Department of General Practice, Monash University, Vic; moderator, GPs Down Under

Katrina McLean BMLSc (Dist), PGDip Sport Med, MBChB, FRNZCGP, FRACGP, Assistant Professor, Faculty of Health Sciences and Medicine, Bond University, Qld; moderator, GPs Down Under

Competing interests: None

Funding: AHKh received a medical student scholarship from GP Synergy Ltd to undertake this research project.

Provenance and peer review: Not commissioned, externally peer reviewed.

Acknowledgements

We thank Drs Nicole Higgins and Tim Leeuwenberg, who were co-investigators in this project and moderators of GPDU, and contributed to the review of the study protocol. Ms Sarah Jacob (UNSW Sydney) and Ms Charmaine Rodricks (GP Unit) assisted with some of the administrative logistics. We thank GP Synergy Ltd who provided a scholarship that supported AHKh. Lastly, we acknowledge and thank all the doctors who participated in this study.

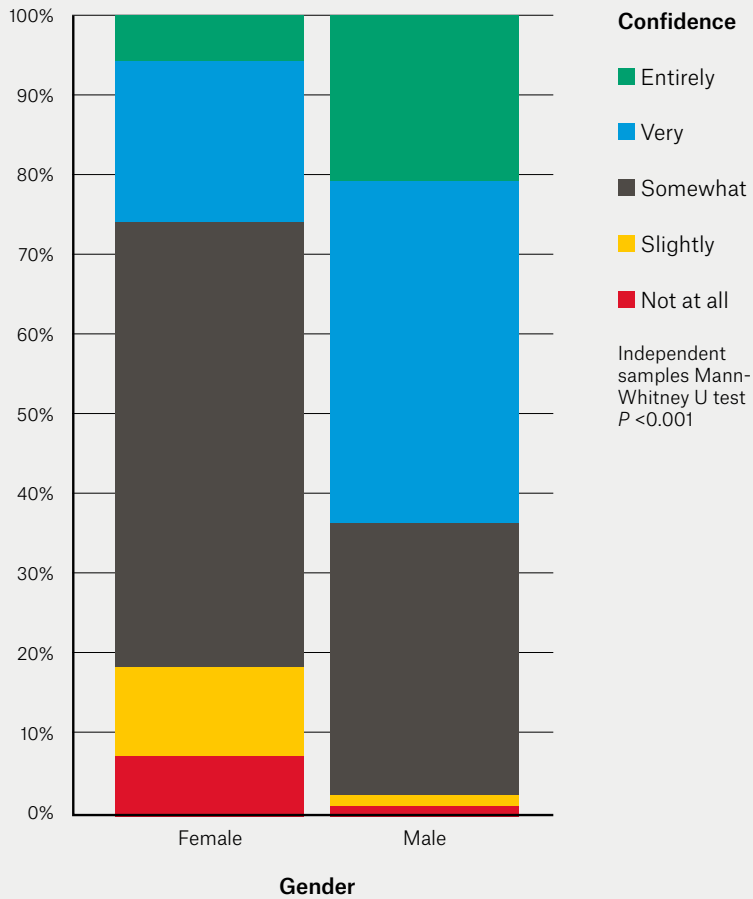

Figure 1. Confidence in 'dichotomous choice' question performance by gender

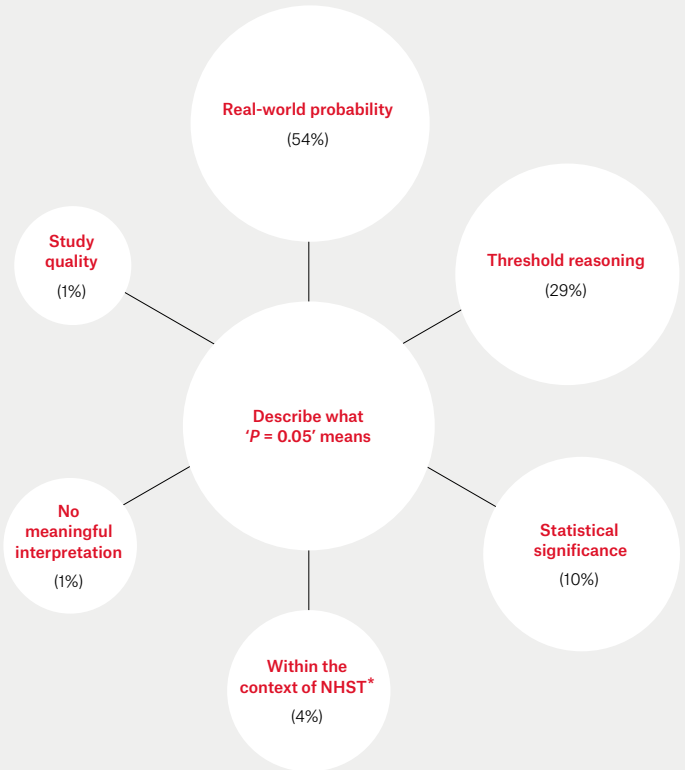

Figure 2. Thematic categories of how participants conceptualised '*P* = 0.05'
**NHST: Null-hypothesis statistical testing*
The percentages are the proportion that the thematic category was the predominant concept in the participant responses.

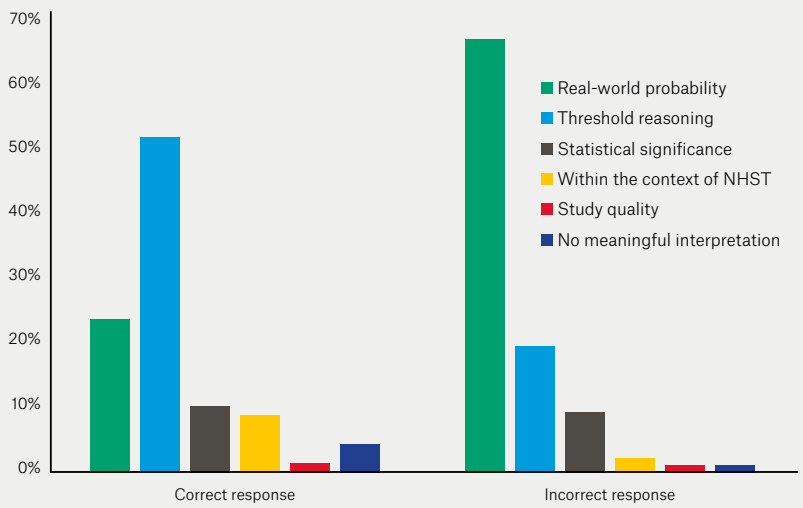

Figure 3. Association between '*P* = 0.05' conceptualisation and interpretation
NHST, null-hypothesis statistical testing

References

1. Wasserstein RL, Lazar NA. The ASA's statement on P values: Context, process, and purpose. Am Stat 2016;70(2):129–31.
2. Goodman S. A dirty dozen: Twelve P value misconceptions. Semin Hematol 2008;45(3):135–40. doi: 10.1053/j.seminhematol.2008.04.003.
3. Nuzzo R. Scientific method: Statistical errors. Nature 2014;506(7487):150–52. doi: 10.1038/506150a.
4. Siegfried T. P value ban: Small step for a journal, giant leap for science. ScienceNews 17 March 2015. Available at www.sciencenews.org/blog/context/P value-ban-small-step-journal-giant-leap-science [Accessed 7 June 2018].
5. Anderson BL, Williams S, Schulkin J. Statistical literacy of obstetrics-gynecology residents. J Grad Med Educ 2013;5(2):272–75. doi: 10.4300/JGME-D-12-00161.1.
6. Windish DM, Huot SJ, Green ML. Medicine residents' understanding of the biostatistics and results in the medical literature. JAMA 2007;298(9):1010–22.
7. Price K. GPs Down Under: Who are we and what do we do? MJA InSight. 22 January 2018.
8. Knibbs J. Are you a member of Australia's underground GP college? The Medical Republic. Surry Hills, Sydney: The Moose Republic. 10 March 2017.
9. Costantino TE. Constructivism. In: Given L, editor. The SAGE encyclopedia of qualitative research methods. Thousand Oaks, CA: SAGE Publications Inc, 2008.
10. Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests, *P* values, confidence intervals, and power: A guide to misinterpretations. Eur J Epidemiol 2016;31(4):337–50. doi: 10.1007/s10654-016-0149-3.
11. Zipkin DA, Umscheid CA, Keating NL, et al. Evidence-based risk communication: A systematic review. Ann Intern Med 2014;161(4):270–80. doi: 10.7326/M14-0295.
12. Ryan MT, Kolodner JL. Using 'rules of thumb' practices to enhance conceptual understanding and scientific reasoning in project-based inquiry classrooms. Proceedings of the 6th International Conference on Learning Sciences. Santa Monica, CA: International Society of the Learning Sciences, 2004: p. 449–56.

correspondence **ajgp@racgp.org.au**