

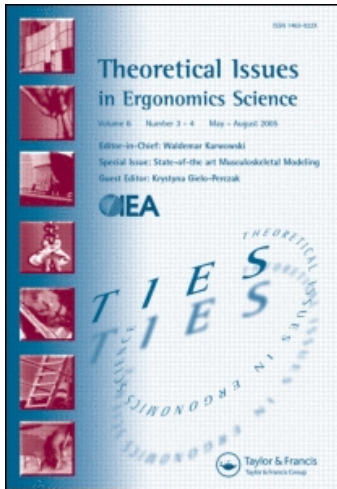
This article was downloaded by: [Rice, Stephen]

On: 29 May 2009

Access details: Access Details: [subscription number 911826838]

Publisher Taylor & Francis

Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



Theoretical Issues in Ergonomics Science

Publication details, including instructions for authors and subscription information:

<http://www.informaworld.com/smpp/title-content=t713697886>

How many people have to die over a type II error?

Stephen Rice^a; David Trafimow^a

^a New Mexico State University, Las Cruces, NM 88003-8001, USA

First Published on: 29 May 2009

To cite this Article Rice, Stephen and Trafimow, David(2009)'How many people have to die over a type II error?','Theoretical Issues in Ergonomics Science,99999:1,

To link to this Article: DOI: 10.1080/14639220902853096

URL: <http://dx.doi.org/10.1080/14639220902853096>

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: <http://www.informaworld.com/terms-and-conditions-of-access.pdf>

This article may be used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

How many people have to die over a type II error?

Stephen Rice* and David Trafimow

New Mexico State University, PO Box 30001, Las Cruces, NM 88003–8001, USA

(Received 3 August 2008; final version received 5 January 2009)

Although much concern over type I errors has permeated psychology for decades, there is less concern over type II errors. In fact, type II errors constitute a serious problem in safety research that can result in accidents and fatalities because researchers fail to reject the null hypothesis due to arbitrary probability thresholds. The purpose of this paper is to reveal how often type II errors occur and the effect they have on applied ergonomics research. Computer simulations using population parameters were generated, revealing that type II errors happen quite often, particularly with effect sizes between 0.2 and 1.2. A utility analysis also reveals that the cost of type II errors on society is much greater than it needs to be. Solutions for avoiding type II errors are discussed.

Keywords: type; errors; statistics; *p*-value; hypothesis testing

1. Introduction

Concern over errors in the null hypothesis significance testing procedure (NHSTP) has a long tradition in statistics and in other fields that employ these statistical methods (e.g. Neyman and Pearson 1933, Meehl 1978, Cohen 1994, Loftus 1996, Trafimow 2003). As several authorities have indicated (e.g. Gigerenzer 1990, Loftus 1991, Hubbard and Bayarri 2003), NHSTP is actually a conflation of the procedures proposed by Fisher and by Neyman and Pearson, with which neither would be likely to approve. To see this, consider the following quote from Gigerenzer (1990):

...almost no [social science statistics] text presented Neyman and Pearson's theory as an alternative to Fisher's, still less as a competing theory. The great mass of texts tried to fuse the controversial ideas into some *hybrid* statistical theory... Of course this meant doing the impossible. But... statisticians were eager to sell, and psychologists were eager to buy *the* method of inductive inferences. The statistical texts now taught hybrid statistics, of which neither Fisher nor, to be sure, Neyman and Pearson would have approved. The type-II errors became added to null hypothesis testing (although it could not be determined in this context), Neyman and Pearson's interpretation of the level of significance as the proportion of type-I errors in the long run became mishmashed with Fisher's and so on. Whatever the textbooks taught, it was *not* indicated that some of the ideas stemmed from Fisher, others from Neyman and Pearson. The hybrid statistics was presented anonymously, as if it were the only truth, as if there existed *only one type of statistics*. There was no mention of the existence of a deep controversy, much less of the controversial issues, nor of the existence of alternative statistical theories... (p. 208)

*Corresponding author. Email: sc_rice@yahoo.com

Of particular interest to applied researchers, who have to decide whether or not to act on the obtained data, Neyman and Pearson (1933) discussed two types of errors that they referred to as incorrectly rejecting the null hypothesis (α) and incorrectly accepting the null hypothesis (β). These are commonly referred as type I and type II errors, respectively.

Imagine that an experiment is conducted to test whether a new airbag system is more effective than a previous system. The null hypothesis states that the two systems are identical in effectiveness; the alternative hypothesis states that the new system is more effective than the previous system. Now, imagine that, in reality, the new system is no better than the old system, but the null hypothesis is rejected because a p -value less than 0.05 is found (this is an arbitrary threshold; the argument applies to any threshold given). This is an example of a type I error.

Type I errors should be avoided for theoretical and practical reasons. From a theoretical point of view, type I errors indicate that the researcher has falsely rejected a null hypothesis, which was in reality true. There is a long tradition in psychology of trying to avoid this type of error and a conservative significance level was derived early on in the history of statistics for this very purpose. Fisher (1935) argued that since one could never prove the null hypothesis, it fell to the researcher to disprove the null hypothesis instead. However, to disprove the null hypothesis, one has to set a conservative threshold for significance. He suggested a significance level of $p < 0.05$, which is still commonly used today in most psychology journals.

Many modern-day researchers are at odds with the notion of having such a threshold and some refer to it as the 'tyranny of the 0.05' (Wickens 1998, p. 18; see also Cohen (1994), Loftus (1996), Schmidt (1996)). Rosnow and Rosenthal (1989) outlined several criticisms of the current system, including:

- (a) the over-reliance on dichotomous significance-testing decisions; (b) the tendency to do many research studies in situations of low power; (c) the habit of defining the results of research in terms of significance levels alone; (d) the overemphasis on original studies and single studies at the expense of replications (p. 1276).

From a practical point of view, type I errors can cause multiple problems for researchers and consumers. First, they give false hope that the effect is real, which can lead to sub-optimal behaviours on the part of consumers. Second, it can lead future researchers to assume the effect is real and thus base their future experimental designs on the fallacy generated from the original type I error. Third, it can induce agencies to employ this new 'improved' system with a resulting expenditure of time and money on producing a system that has no benefit.

Although many researchers have highlighted the perils of the type I error, there has been less discussion on the potential perils of type II errors. Going back to the airbag system example, imagine that the new system was actually more effective than the older system, but one failed to find a significant effect (e.g. $p > 0.05$) and thereby failed to reject the null hypothesis in favour of the alternative hypothesis. This would result in a type II error. What consequences might this produce and why might it be critical to avoid these types of errors?

From the perspective of the researcher, one might wish to avoid type II errors for several reasons. What dissertation-level student has not spent a sleepless night or two worrying about his or her data and praying that it comes out 'significant'. A non-significant result can mean another year of research before receiving that PhD. Or it can mean being denied tenure for that assistant professor. Or it can mean losing

future grant money. These are life-changing consequences for the researcher who commits a type II error.

But these consequences pale in comparison to the consequences to the consumer, particularly in human factors research. As Wickens (1998) argues, committing a type II error may be disastrous when it comes to safety research. Again imagine that one has committed a type II error when experimenting with the new airbag system. Unknowingly, a system that will save lives, perhaps thousands of lives, has been rejected. Clearly, one should consider the consequences of type II errors as a function of what type of research is being conducted and not just assume that all type I errors are more detrimental than all type II errors.

When a conservative threshold for rejection (e.g. $p < 0.05$) is used, NHSTP tends to produce more type II errors than type I errors. NHSTP is often made to be even more conservative by using two-tailed tests rather than one-tailed tests, which increases the probabilities of type II errors even further. Reviewers and editors should be cognisant that safety devices are rarely designed to result in more accidents; rather, hypotheses tend to be uni-directional in favour of the efficacy of new safety devices. Therefore, it makes sense to use one-tailed tests to investigate whether the new safety device is better than the old one.

The present authors argue that researchers should not blindly follow the standard statistical methodology without considering how the type of research affects whether one should avoid type I errors or type II errors. In addition, it is argued that the consequences of a type II error depend on more than just the type of research. It also depends on the effect size found in the experiment and the amount of cost/benefit gained from the research.

For example, if research is conducted to test a new type of kitchen timer and a small effect size is discovered, it would not be critical to avoid type II errors because the timer does not have life-changing (or life-saving) benefits and any benefits it does have are probably small. However, if one is testing a new crash avoidance warning system in future automobiles and a large effect size is seen (but the statistical test is not significant), then the cost of the type II error is dramatic. People will die because of it.

Because the authors are interested in human factors issues, particularly with regard to public safety, the remaining focus of this paper will be primarily on type II errors. It is intended to first show that type II errors occur more frequently than most people suspect. Then it is intended to show how effect size and severity interact as a function of expected value (EV). For simplicity's sake, the specific paradigm that will be used is safety research in automobiles (e.g. seatbelts, airbags, etc.). To answer these questions, computer simulations will be conducted that produce millions of 'experiments' based on preset population parameters. From these simulations, it will be shown exactly how often type II errors occur and what cost they produce as a function of effect size and severity.

2. Method

2.1. Overview

The basic strategy of the simulation was to perform an infinitely large set of experiments, whereby each experiment contained group 1 of n participants and group 2 of n participants. Although it is impossible to generate an infinite number of experiments, a computer program was used (VBA) to randomly generate 3.9 million datasets, thus assuring that the simulation data that were produced were sufficiently close to what would be generated by an infinite number of experiments.

2.2. Materials and software

The experimental simulation was conducted on a MacPro Xeon Workstation with 2xQuadCore processors and 8 gb of RAM. 3.9 million datasets (65,000 per effect size * 3 sizes of n) were generated using a VBA random number generator, which predetermined a normal distribution with fixed population means (μ) and standard deviations (σ).

2.3. Procedure

Two columns of data (experimental group and control group) were generated for each experiment. For the study purposes, the experimental group represented a new safety device for automobiles, whereas the control group represented an older type of safety device (or no safety device). The data from the experimental groups were generated using μ ranging from 1.0 to 3.0, in 0.1 increments, and $\sigma = 1.0$. The data from the control groups were generated using population $\mu = 1.0$ and $\sigma = 1.0$. This effectively manipulated the population effect sizes, which ranged from 0.0 to 2.0. For example, a simulation generating a population effect size of 1.1 would consist of a control group with population $\mu = 1.0$ and $\sigma = 1.0$ and an experimental group with population $\mu = 2.1$ and $\sigma = 1.0$.

Following generation of the experimental data, sample means and standard deviations were computed for each group. Binary scoring was used to indicate the direction of the experimental effect for each experiment, which made it possible to calculate the true proportion of experiments in which the group 1 sample mean exceeded the group 2 sample mean. Subsequently, between-participants t -tests (one-tailed and two-tailed) were conducted on each experiment, producing a p -value. Binary scoring was again used to determine if $p < 0.05$ and the means were in the right direction (i.e. correctly rejecting the null).

From these data, it was possible to determine the true probability of making a correct decision ($p[C]$) for each effect size. Each effect size was then converted into a correlation coefficient (see Equation (1)), where R is the population correlation coefficient and E is the effect size (see Rosenthal and Rosnow (1991)).

$$R = \frac{E}{\sqrt{4 + E^2}} \quad (1)$$

Using the result from Equation (1) (R), the binomial effect size display was used to calculate the percentage of times the new safety device prevents an accident when the null hypothesis is always rejected, using Equation (2), where OP is the optimal probability under this condition (see Rosenthal and Rosnow (1991)); OP is the probability, assuming that the null hypothesis has been rejected, of obtaining the desired result in practice (e.g. preventing the accident with the new safety device). To use the binomial effect size display in this context, it is assumed that the probability of preventing an accident is 50% when the old safety device is used. Consequently, as long as the effect size is positive, which it always is in the present computer simulations, OP always exceeds 50%.

$$OP = \frac{R}{2} + 0.5 \quad (2)$$

Following this, the percentage of accidents that the new safety device prevented was calculated based on the traditional method of rejecting the null when $p < 0.05$, using Equation (3), where EP is the expected probability using NHSTP and C is the probability

of making a correct decision. For example, when $n = 10$, and the effect size is 0.7, OP is 0.665 and EP is 0.572.

$$EP = (C * OP) + ((1 - C) * 0.5) \tag{3}$$

Finally, EP and OP was multiplied by 2 (low), 6 (moderate) or 10 (severe) to simulate the optimal utility and expected utility of potentially preventing various levels of damage to the individual in the case of an accident. For example, a value of 2 might indicate the driver was potentially prevented from being bruised, a value of 6 might indicate potential prevention of broken bones and a value of 10 would indicate potential prevention of fatalities. Obviously, these values are arbitrary but they should nevertheless serve to illustrate the importance of considering the severity of accidents as well as their probabilities. It was hypothesised that the difference between OP and EP would be magnified as the severity of the accident increased.

3. Results

The following sections describe the findings from two types of simulations. The first section simulates the probability of correctly rejecting the null hypothesis in the correct direction as a function of the effect size. The second section concerns the expected cost of using the standard NHSTP given the probability of making the right decision and the utility of making the right decision.

3.1. The probability of correctly rejecting the null hypothesis

Figure 1 illustrates the probability of correctly rejecting the null hypothesis given that the population effect sizes range from 0.0 to 2.0. Obviously, in cases where the population

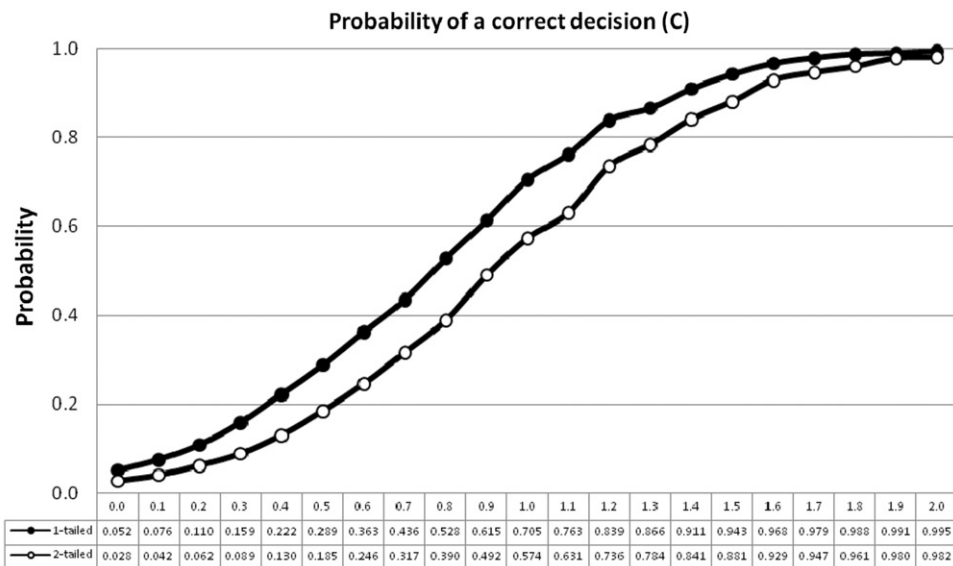


Figure 1. Probability of making a correct decision as a function of effect size, for t -tailed and two-tailed t -tests.

effect size exceeds 0.0, the null hypothesis is always wrong and should be rejected. However, as Figure 1 demonstrates, the probability of making the correct decision is not very impressive, particularly at low population effect sizes. This probability becomes even lower when using two-tailed *t*-tests. Not surprisingly, this probability increases as the effect sizes increase and as the sample sizes increase.

3.2. The potential cost of using NHSTP

As was described earlier, the population effect size was translated into the percentage of people who would be helped by the safety device assuming that the null hypothesis was correctly rejected in every study, which is the *OP* of success. By arbitrarily assuming various values for helping people to avoid injuries of differing severities, labelled as ‘low’ (2), ‘moderate’ (6) and ‘severe’ (10), it was also possible to calculate the optimal utility of correctly rejecting the null hypothesis and always using the safety device.

Alternatively, given that NHSTP is used, and so the null hypothesis is not always rejected, it is clear that optimal utility will not be obtained. But by combining the probability of correctly rejecting the null hypothesis with the arbitrarily assigned utilities, it is a simple matter to find expected utilities that can then be compared with optimal utilities. The difference between an expected utility and the corresponding optimal utility is the benefit that society foregoes by the dependence of engineering psychologists on NHSTP.

Figure 2 illustrates the foregoing considerations when there are 10 participants in each condition. Not surprisingly, as the utility of the safety device increases from ‘low’ to ‘moderate’ to ‘severe’, so does the optimal utility as well as the expected utility. Also not surprisingly, optimal utility tends to exceed expected utility. But Figure 2 also shows some important interactions. For instance, optimal and expected utilities tend to converge at extremely low or extremely high effect sizes, whereas the difference between them tends to max out at moderate effect sizes.

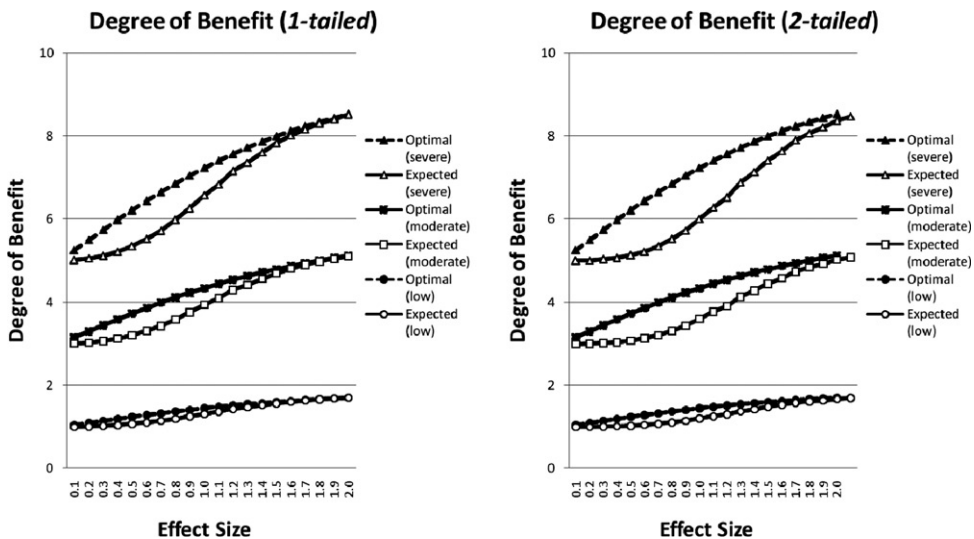


Figure 2. Degree of benefit of making a correct decision for optimal utility vs expected utility, for *t*-tailed and two-tailed *t*-tests, as a function of severity.

Downloaded By: [Rice, Stephen] At: 18:17 29 May 2009

effect sizes very few people are being helped by the safety device and so little is lost by not using it. In addition, at large effect sizes the likelihood of failing to reject the null hypothesis is extremely small. It is only when the effect size is moderate that: (a) the safety device provides enough help to matter; (b) there is a reasonable likelihood of failing to reject the null hypothesis.

This interaction is qualified by the utility of the help. When the utility of the help is low, then the difference between expected utility and optimal utility is minimised. But as the two topmost curves in Figure 2 demonstrate, under high utility the difference between expected utility and optimal utility is impressive, particularly at moderate effect sizes. Finally, Tables 1–4 demonstrate that these effects are magnified even more when two-tailed tests are used rather than one-tailed tests. As an example of the importance of distinguishing between optimal and expected probabilities for severe accidents, consider that when the population effect size is 0.7, $EV = 5.75$ and $OP = 6.65$, for a difference of 0.90. In contrast, when the accidents are mild, $EV = 1.14$, $OP = 1.33$, for a difference of only 0.19 (see Table 2).

3.3. The effects of sample size

Tables 1–4 show how increases in sample size can modulate the effects described above. Table 1 demonstrates these effects when the sample size is very small ($n = 6$). While many ergonomics researchers would avoid such a low n , one must be careful not to assume that increasing n for any given experiment is a simple matter of gathering more participants. For example, in aviation research, it is often quite difficult to find an infinite number of licensed pilots that are willing to participate in experimental research. The problem becomes exacerbated when experimental designs prevent the use of within-participants analysis, forcing researchers to run between-participants designs due to issues of fatigue, contamination between conditions, etc. (e.g. Dixon and Wickens 2006). Thus, it is often the case that researchers are limited to a low n when conducting aviation research (e.g. Thomas and Wickens 2004).

Tables 2–4 present data from sample sizes of 10, 20 and 30, respectively. Clearly, as the sample size is increased, the gap between OP and EV becomes increasingly smaller, leading to fewer opportunities for type II errors. Thus, when possible, it would be beneficial to increase n as a means of reducing type II errors. However, this does not fully resolve the issue, as even when $n = 30$, there is a noticeable difference between OP and EV , particularly when accidents are severe and effect sizes are small.

4. Discussion

Previous researchers (e.g. Murphy and Myors 2004) have shown mathematically that NHSTP results in high probabilities of type II errors when p is set at, or below, the traditional level of 0.05. The computer simulations resulted in similar findings and reinforce previous conclusions about the high probability of failing to reject the null hypothesis when it should be rejected. It is important to note just how great the probability of a type II error is, particularly when the effect size is low to moderate. In fact, it takes a population effect size of 1.0 just to reach a level where the researcher is correct 60–70% of the time, depending on whether he/she used a one-tailed or two-tailed t -test. Effect sizes of 1.0 are often considered to be large (Cohen 1992); many studies reveal effect sizes that are much smaller, but are still taken quite seriously (Cohen 1992). An example would be the

Table 1. Data from the simulation ($n = 6$).

Effect size	Mean 1	SD 1	Mean 2	SD 2	1-tail t -test	2-tail t -test	Optimal	EV	Optimal (low)	Optimal (moderate)	Optimal (severe)	EV (low)	EV (moderate)	EV (severe)
0.0	1.00	0.95	0.99	0.95	0.047	0.024	0.500	0.500	1.000	3.000	5.000	1.000	3.000	5.000
0.1	1.10	0.95	0.99	0.95	0.079	0.038	0.525	0.502	1.050	3.150	5.250	1.004	3.012	5.020
0.2	1.21	0.95	0.99	0.95	0.096	0.052	0.550	0.505	1.100	3.229	5.498	1.010	3.029	5.048
0.3	1.29	0.95	0.99	0.95	0.127	0.073	0.574	0.509	1.148	3.445	5.742	1.019	3.057	5.094
0.4	1.40	0.96	0.99	0.95	0.152	0.087	0.598	0.515	1.196	3.588	5.981	1.030	3.089	5.149
0.5	1.50	0.95	0.99	0.95	0.194	0.116	0.621	0.523	1.243	3.728	6.213	1.047	3.141	5.235
0.6	1.60	0.95	0.99	0.95	0.257	0.167	0.644	0.537	1.287	3.862	6.437	1.074	3.221	5.369
0.7	1.70	0.96	0.99	0.95	0.298	0.187	0.665	0.549	1.330	3.991	6.652	1.098	3.295	5.492
0.8	1.80	0.96	0.99	0.95	0.358	0.244	0.686	0.566	1.371	4.114	6.857	1.133	3.399	5.665
0.9	1.90	0.96	0.99	0.95	0.443	0.304	0.705	0.591	1.410	4.231	7.052	1.182	3.545	5.908
1.0	2.00	0.95	0.99	0.95	0.490	0.362	0.724	0.609	1.447	4.342	7.236	1.219	3.657	6.095
1.1	2.09	0.95	0.99	0.95	0.545	0.407	0.741	0.631	1.482	4.446	7.410	1.262	3.787	6.312
1.2	2.21	0.94	0.99	0.95	0.629	0.477	0.757	0.662	1.514	4.543	7.572	1.324	3.971	6.618
1.3	2.30	0.96	0.99	0.95	0.683	0.538	0.772	0.686	1.545	4.635	7.725	1.372	4.116	6.860
1.4	2.38	0.96	0.99	0.95	0.721	0.584	0.787	0.707	1.573	4.720	7.867	1.413	4.240	7.067
1.5	2.51	0.96	0.99	0.95	0.791	0.667	0.800	0.737	1.600	4.800	8.000	1.474	4.423	7.372
1.6	2.60	0.95	0.99	0.95	0.827	0.703	0.812	0.758	1.625	4.874	8.123	1.516	4.549	7.582
1.7	2.70	0.95	0.99	0.95	0.863	0.756	0.824	0.779	1.648	4.943	8.238	1.559	4.677	7.795
1.8	2.80	0.95	0.99	0.95	0.899	0.802	0.834	0.801	1.669	5.007	8.345	1.601	4.803	8.005
1.9	2.91	0.94	0.99	0.95	0.926	0.852	0.844	0.819	1.689	5.066	8.444	1.637	4.912	8.187
2.0	3.01	0.96	0.99	0.95	0.945	0.881	0.854	0.834	1.707	5.121	8.536	1.668	5.005	8.341

EV = expected value.

Table 2. Data from the simulation ($n = 10$).

Effect size	Mean 1	SD 1	Mean 2	SD 2	1-tail t -test	2-tail t -test	Optimal	EV	Optimal (low)	Optimal (moderate)	Optimal (severe)	EV (low)	EV (moderate)	EV (severe)
0.0	1.00	0.97	1.00	0.98	0.052	0.028	0.500	0.500	1.000	3.000	5.000	1.000	3.000	5.000
0.1	1.11	0.98	1.00	0.98	0.076	0.042	0.525	0.502	1.050	3.150	5.250	1.004	3.011	5.019
0.2	1.20	0.97	1.00	0.98	0.110	0.062	0.550	0.505	1.100	3.299	5.498	1.011	3.033	5.054
0.3	1.30	0.98	1.00	0.98	0.159	0.089	0.574	0.512	1.148	3.445	5.742	1.024	3.071	5.118
0.4	1.40	0.98	1.00	0.98	0.222	0.130	0.598	0.522	1.196	3.588	5.981	1.044	3.131	5.218
0.5	1.50	0.97	1.00	0.98	0.289	0.185	0.621	0.535	1.243	3.728	6.213	1.070	3.210	5.350
0.6	1.61	0.97	1.00	0.98	0.363	0.246	0.644	0.552	1.287	3.862	6.437	1.104	3.312	5.521
0.7	1.69	0.98	1.00	0.98	0.436	0.317	0.665	0.572	1.330	3.991	6.652	1.144	3.432	5.719
0.8	1.80	0.98	1.00	0.98	0.528	0.390	0.686	0.598	1.371	4.114	6.857	1.196	3.588	5.980
0.9	1.91	0.98	1.00	0.98	0.615	0.492	0.705	0.626	1.410	4.231	7.052	1.252	3.757	6.261
1.0	2.01	0.98	1.00	0.98	0.705	0.574	0.724	0.658	1.447	4.342	7.236	1.315	3.946	6.576
1.1	2.10	0.97	1.00	0.98	0.763	0.631	0.741	0.684	1.482	4.446	7.410	1.367	4.102	6.837
1.2	2.20	0.97	1.00	0.98	0.839	0.736	0.757	0.716	1.514	4.543	7.572	1.432	4.295	7.158
1.3	2.30	0.98	1.00	0.98	0.866	0.784	0.772	0.736	1.545	4.635	7.725	1.472	4.416	7.360
1.4	2.40	0.97	1.00	0.98	0.911	0.841	0.787	0.761	1.573	4.720	7.867	1.522	4.566	7.611
1.5	2.50	0.97	1.00	0.98	0.943	0.881	0.800	0.783	1.600	4.800	8.000	1.566	4.697	7.829
1.6	2.60	0.98	1.00	0.98	0.968	0.929	0.812	0.802	1.625	4.874	8.123	1.604	4.813	8.022
1.7	2.70	0.97	1.00	0.98	0.979	0.947	0.824	0.817	1.648	4.943	8.238	1.634	4.902	8.170
1.8	2.79	0.97	1.00	0.98	0.988	0.961	0.834	0.830	1.669	5.007	8.345	1.661	4.983	8.305
1.9	2.91	0.97	1.00	0.98	0.991	0.980	0.844	0.841	1.689	5.066	8.444	1.683	5.048	8.413
2.0	3.00	0.97	1.00	0.98	0.995	0.982	0.854	0.852	1.707	5.121	8.536	1.704	5.111	8.518

EV = expected value.

Table 3. Data from the simulation ($n = 20$).

Effect size	Mean 1	SD 1	Mean 2	SD 2	1-tail t -test	2-tail t -test	Optimal	EV	Optimal (low)	Optimal (moderate)	Optimal (severe)	EV (low)	EV (moderate)	EV (severe)
0.0	1.00	0.99	1.00	0.99	0.050	0.023	0.500	0.500	1.000	3.000	5.000	1.000	3.000	5.000
0.1	1.11	0.99	1.00	0.99	0.107	0.059	0.525	0.503	1.050	3.150	5.250	1.005	3.016	5.027
0.2	1.20	0.98	1.00	0.91	0.157	0.094	0.550	0.508	1.100	3.299	5.498	1.016	3.047	5.078
0.3	1.30	0.99	1.00	0.99	0.232	0.154	0.574	0.517	1.148	3.445	5.742	1.034	3.103	5.172
0.4	1.40	0.99	1.00	0.99	0.361	0.236	0.598	0.535	1.196	3.588	5.981	1.071	3.212	5.354
0.5	1.50	0.99	1.00	0.99	0.460	0.357	0.621	0.556	1.243	3.728	6.213	1.112	3.335	5.558
0.6	1.61	0.98	1.00	0.99	0.602	0.449	0.644	0.586	1.287	3.862	6.437	1.173	3.519	5.865
0.7	1.69	0.99	1.00	0.99	0.691	0.571	0.665	0.614	1.330	3.991	6.652	1.228	3.685	6.141
0.8	1.80	0.99	1.00	0.99	0.809	0.685	0.686	0.650	1.371	4.114	6.857	1.300	3.901	6.502
0.9	1.91	0.99	1.00	0.99	0.877	0.797	0.705	0.680	1.410	4.231	7.052	1.360	4.080	6.799
1.0	2.01	0.99	1.00	0.99	0.929	0.861	0.724	0.708	1.447	4.342	7.236	1.415	4.246	7.077
1.1	2.10	0.98	1.00	0.99	0.957	0.914	0.741	0.731	1.482	4.446	7.410	1.461	4.384	7.306
1.2	2.20	0.99	1.00	0.99	0.981	0.963	0.757	0.752	1.514	4.543	7.572	1.505	4.514	7.524
1.3	2.30	0.99	1.00	0.99	0.992	0.986	0.772	0.770	1.545	4.635	7.725	1.541	4.622	7.703
1.4	2.40	0.99	1.00	0.99	0.998	0.995	0.787	0.786	1.573	4.720	7.867	1.572	4.717	7.862
1.5	2.50	0.99	1.00	0.99	1.000	0.997	0.800	0.800	1.600	4.800	8.000	1.600	4.800	8.000
1.6	2.60	0.99	1.00	0.99	1.000	0.997	0.812	0.812	1.625	4.874	8.123	1.625	4.874	8.123
1.7	2.70	0.99	1.00	0.99	1.000	1.000	0.824	0.824	1.648	4.943	8.238	1.648	4.943	8.238
1.8	2.79	0.98	1.00	0.99	1.000	0.999	0.834	0.834	1.669	5.007	8.345	1.669	5.007	8.345
1.9	2.91	0.98	1.00	0.99	1.000	1.000	0.844	0.844	1.689	5.066	8.444	1.689	5.066	8.444
2.0	3.00	0.99	1.00	0.99	1.000	0.999	0.854	0.854	1.707	5.121	8.536	1.707	5.121	8.536

EV = expected value.

Table 4. Data from the simulation ($n = 30$).

Effect size	Mean 1	SD 1	Mean 2	SD 2	1-tail t -test	2-tail t -test	Optimal	EV	Optimal (low)	Optimal (moderate)	Optimal (severe)	EV (low)	EV (moderate)	EV (severe)
0.0	0.00	0.00	0.00	0.00	0.000	0.000	0.500	0.500	1.000	3.000	5.000	1.000	3.000	5.000
0.1	1.11	0.99	1.00	1.00	0.110	0.060	0.525	0.503	1.050	3.150	5.250	1.005	3.016	5.027
0.2	1.20	0.98	1.00	1.00	0.185	0.123	0.550	0.509	1.100	3.299	5.498	1.018	3.055	5.092
0.3	1.30	1.00	1.00	1.00	0.300	0.208	0.574	0.522	1.148	3.445	5.742	1.045	3.134	5.223
0.4	1.40	1.00	1.00	1.00	0.470	0.327	0.598	0.546	1.196	3.588	5.981	1.092	3.277	5.461
0.5	1.49	0.99	1.00	1.00	0.598	0.482	0.621	0.573	1.243	3.728	6.213	1.145	3.435	5.726
0.6	1.61	0.99	1.00	1.00	0.763	0.640	0.644	0.610	1.287	3.862	6.437	1.219	3.658	6.097
0.7	1.69	1.00	1.00	1.00	0.838	0.745	0.665	0.638	1.330	3.991	6.652	1.277	3.831	6.385
0.8	1.79	1.00	1.00	1.00	0.907	0.835	0.686	0.668	1.371	4.114	6.857	1.337	4.010	6.684
0.9	1.91	0.99	1.00	1.00	0.975	0.945	0.705	0.700	1.410	4.231	7.052	1.400	4.200	7.001
1.0	2.01	0.99	1.00	1.00	0.978	0.960	0.724	0.719	1.447	4.342	7.236	1.438	4.313	7.188
1.1	2.09	0.99	1.00	1.00	0.997	0.980	0.741	0.740	1.482	4.446	7.410	1.480	4.441	7.402
1.2	2.20	0.99	1.00	1.00	0.993	0.993	0.757	0.756	1.514	4.543	7.572	1.511	4.533	7.555
1.3	2.30	0.99	1.00	1.00	1.000	1.000	0.772	0.772	1.545	4.635	7.725	1.545	4.635	7.725
1.4	2.40	1.00	1.00	1.00	1.000	1.000	0.787	0.787	1.573	4.720	7.867	1.573	4.720	7.867
1.5	2.50	0.99	1.00	1.00	1.000	1.000	0.800	0.800	1.600	4.800	8.000	1.600	4.800	8.000
1.6	2.60	1.00	1.00	1.00	1.000	1.000	0.812	0.812	1.625	4.874	8.123	1.625	4.874	8.123
1.7	2.70	1.00	1.00	1.00	1.000	1.000	0.824	0.824	1.648	4.943	8.238	1.648	4.943	8.238
1.8	2.79	0.99	1.00	1.00	1.000	1.000	0.834	0.834	1.669	5.007	8.345	1.669	5.007	8.345
1.9	2.91	0.99	1.00	1.00	1.000	1.000	0.844	0.844	1.689	5.066	8.444	1.689	5.066	8.444
2.0	3.00	0.99	1.00	1.00	1.000	1.000	0.854	0.854	1.707	5.121	8.536	1.707	5.121	8.536

EV = expected value.

traffic alert and collision avoidance system that has had some success in the real world (Williamson and Spencer 1989). Thus, one must be aware of the danger of the type II error in these types of studies.

In addition, because utility was included in the calculations, it was also possible to simulate the harm to society that is caused by failing to reject the null. To the extent that the accidents prevented by the safety device are serious, the harm to society caused by failing to reject the null hypothesis when it should be rejected becomes increasingly important. The computer simulations also demonstrate that there is an interaction between the size of the effect and the degree of harm prevented by the device when it works; harm to society from type II error is maximised at moderate effect sizes and when the safety device prevents severe injury. This is an important point to drive home, because the authors suspect that many, if not most, ergonomics studies reveal effect sizes in this range (0.2–1.2). And yet, as can be seen in Figure 2, the gap between optimal utility and expected utility is quite large in this range, with more than a 10% difference in some cases. Thus, it is important to note that the authors are not talking about errors in extreme cases where effect sizes are so tiny that no one would report them anyway, or when effect sizes are so large that the danger of a type II error almost disappears. The biggest danger of a type II error occurs when effect sizes are in the moderate range.

So what can be done to reduce type II errors? There is a variety of potential solutions. One possibility is to increase the alpha level so that p no longer has to be less than the traditional level of 0.05 to enable the null hypothesis to be rejected. Using one-tailed tests instead of two-tailed tests has this effect, although the present simulations show that it is clearly not enough to overcome the entire problem (see Figure 2). Another possibility is to have more participants; as the number of participants increases, so does power (as is shown by Tables 1–3). However, as mentioned previously, this is not always possible, particularly in areas of research that require participants with specialised skills. For example, in aviation, researchers frequently need participants who are licensed pilots (e.g. Wickens *et al.* 2003, Thomas and Wickens 2004), a requirement that is not easy to fill. Even research conducted in aviation laboratories that have access to a flight school nearby often is limited by the number of pilots that can be found. Aviation experiments can be long and fatigue-inducing, thereby preventing the use of within-participants designs (e.g. Dixon *et al.* 2005). Furthermore, it often is the case that once pilots are exposed to one condition, they cannot participate in other conditions due to contamination effects (e.g. Dixon and Wickens 2006).

Furthermore, even if researchers can increase n to 10–20 participants, that is not always sufficient to overcome the loss of power, particularly if the phenomenon of interest is a rare event. For example, one might be interested in how pilots respond to emergency situations; it is a necessary (and fortunate!) fact that these situations occur rarely. Thus, when conducting experimental research on these types of issues, pilots only can be exposed to very few events (sometimes only one) (e.g. Wickens *et al.* 2006); otherwise, the realism of the event is destroyed. The only way to overcome this loss of power is to increase n even more.

Yet another possibility is for the researcher to set up the experiment in as careful a way as possible so as to eliminate as much error variance as possible, although the increase in experimental precision might come at a cost with regard to experimental realism or the ability to generalise the findings beyond the experimental context. Another alternative might be to use the epistemic ratio procedure (Trafimow 2005, 2006), whereby the safety researcher would first decide what the minimum effect size would have to be to justify the adoption of the proposed safety device and use this as the alternative hypothesis. A more

extreme solution would be to employ a strategy where one always rejects the null and simply relies on the sample means as the best indicators of the population means. Finally, researchers might consider a recommendation by Loftus (1996) to add a third category to the reporting of results: accept, reject, or withhold judgement pending the acquisition of new data.

To be clear, the purpose of this paper is not to dictate to the field of ergonomics which method should be used to avoid type II errors. The purpose herein is simply to show the dangers of type II errors, provide possible solutions for avoiding them and let researchers and designers decide what they wish to do with this information. Regardless of whether researchers choose to increase alpha, increase the number of participants, use more precise experimental designs, employ the epistemic ratio procedure, use sample means as indicators of population means or include a 'withhold judgement' category to the reporting of results, the result should be fewer type II errors and higher prevention of injuries or deaths.

In summary, ergonomics editors and reviewers should be flexible when evaluating research studies with important safety implications and consider that it might be more important to avoid a type II error than a type I error, particularly when the rejection of the research could lead to loss of life.

4.1. Other fields of research

For the purposes of this paper, an automobile safety paradigm was used; however, the present findings can be generalised to many other areas of research. For example, safety research spans a wide variety of paradigms, including aviation, transportation, manufacturing, workplace design, etc. In each of these areas, researchers are attempting to provide safety devices and/or techniques that prevent accidents and fatalities. The present authors believe that many researchers in these areas understand the dangers of type II errors and are taking precautions against them. It is also realised that not all researchers are equally aware of how insidious a type II error can be and it is hoped that this paper helps to convince the sceptics that it is a very serious matter indeed; often a matter of life and death.

There are also many areas of applied research where type II errors may not be a significant concern. For example, when developing new medications, it may be more important to avoid type I errors (incorrectly rejecting the null), because the results of a type I error may lead to dangerous and expensive medications that cause harm to, or do nothing for, the patient. Take, for example, a cancer patient who is spending thousands of dollars on a medication that does nothing for him/her, when instead he/she should be taking a different medication that might help. An even worse scenario would be this same patient taking a medication that is actually causing harm to him/her.

In basic research, the dangers of type II errors may also not be very severe. For example, when conducting basic visual perception experiments, a type II error is probably not going to lead to someone's death. Basic researchers are interested in topics that often do not generalise readily to application, so any errors they make do not have catastrophic consequences. This is not a criticism of basic research, as the authors highly value their basic colleagues (and in fact, both authors of this paper conduct and have published basic research); it is simply a natural fact that basic research has different implications from applied research.

5. Conclusion

Type II errors have not received as much ‘press’ in the field of psychology as have type I errors (Murphy and Myors 2004) and most statistics courses focus much more on avoiding type I errors than type II errors. However, there have been several important advocates of avoiding type II errors in applied research (e.g. Wickens 1998). The present authors have heeded their warning and have presented simulated data that reveal just how dangerous a type II error can be and what might be its cost to society.

Although researchers may differ on which remedy they prefer for reducing type II errors, and it is likely to depend on the particular situation at hand, as well as on the mathematical or philosophical virtuosity of the researcher involved, there can be little doubt that researchers should consider carefully the virtues and detractions associated with each of them. The present computer simulations demonstrate that such careful consideration has the potential to greatly increase the benefits to society that accrue from the field of ergonomics.

Acknowledgements

The authors wish to thank an anonymous reviewer for the valuable suggestions on a previous version of this paper.

References

- Cohen, J., 1992. A power primer. *Psychological Bulletin*, 112, 155–159.
- Cohen, J., 1994. The earth is round ($p < 0.05$). *American Psychologist*, 49, 997–1003.
- Dixon, S.R. and Wickens, C.D., 2006. Automation reliability in unmanned aerial vehicle control: a reliance-compliance model of automation dependence in high workload. *Human Factors*, 48 (3), 474–486.
- Dixon, S.R., Wickens, C.D., and Chang, D., 2005. Mission control of multiple unmanned aerial vehicles: a workload analysis. *Human Factors*, 47 (3), 479–487.
- Fisher, R.A., 1935. *The design of experiments*. Edinburgh: Oliver & Boyd.
- Gigerenzer, G., 1990. *The empire of chance: how probability changed science and everyday life*. Oxford: Cambridge University Press.
- Hubbard, R. and Bayarri, M.J., 2003. Confusion over measures of evidence (p 's) versus errors (α 's) in classical statistical testing. *The American Statistician*, 57, 171–182.
- Loftus, G.R., 1991. On the tyranny of hypothesis testing in the social sciences. *Contemporary Psychology*, 36 (2), 102–105.
- Loftus, G.R., 1996. Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science*, 5, 161–171.
- Meehl, P.E., 1978. Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, 46, 806–834.
- Murphy, K.R. and Myors, B., 2004. *Statistical power analysis: a simple and general model for traditional and modern hypothesis testing*. 2nd ed. London: Lawrence Erlbaum Associates.
- Neyman, J. and Pearson, E.S., 1933. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231, 289–337.
- Rosenthal, R. and Rosnow, R.L., 1991. *Essentials of behavioral research: methods and data analysis*. New York: McGraw-Hill, Inc.
- Rosnow, R.L. and Rosenthal, R., 1989. Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, 44 (10), 1276–1284.

- Schmidt, F.L., 1996. Statistical significance testing and cumulative knowledge in psychology: implications for the training of researchers. *Psychological Methods*, 1, 115–129.
- Thomas, L.C. and Wickens, C.D., 2004. Eye-tracking and individual differences in off-normal event detection when flying with a synthetic vision system display. In: *Proceedings of the 48th annual meeting of the Human Factors and Ergonomics Society*. Santa Monica, CA: HFES.
- Trafimow, D., 2003. Hypothesis testing and theory evaluation at the boundaries: surprising insights from Bayes's theorem. *Psychological Review*, 110, 526–535.
- Trafimow, D., 2005. The ubiquitous Laplacian assumption: reply to Lee and Wagenmakers. *Psychological Review*, 112, 669–674.
- Trafimow, D., 2006. Using epistemic ratios to evaluate hypotheses: an imprecision penalty for imprecise hypotheses. *Genetic, Social, and General Psychology Monographs*, 132 (4), 431–462.
- Wickens, C., 1998. Commonsense statistics. *Ergonomics in Design*, 6 (4), 18–22.
- Wickens, C.D., Dixon, S.R., and Johnson, N., 2006. Imperfect diagnostic automation: an experimental examination of priorities and threshold setting. In: *Proceedings of the 50th Annual Meeting of the Human Factors and Ergonomics Society*. Santa Monica CA: HFES.
- Wickens, C.D., et al., 2003. The influences of display highlighting and size and event eccentricity for aviation surveillance. In: *Proceedings of the 47th annual meeting of the Human Factors and Ergonomics Society*. Santa Monica, CA: HFES.
- Williamson, T. and Spencer, N.A., 1989. Development and operation of the Traffic Alert and Collision Avoidance System (TCAS). *Proceedings of the IEEE*, 77 (11), 1735–1744.

About the authors

Stephen Rice is an Assistant Professor of Psychology at New Mexico State University. His main areas of interest are automation, disaster warning systems, and aviation psychology.

David Trafimow is a Professor of Psychology at New Mexico State University. His main areas of interest are in social cognition, philosophy of science, and methodology.