December 2017

'Parallel Universe' or 'Proven Future'? The Language of Dependent Means t-Test Interpretations

Anthony M. Gould
Université Laval, Quebec City, Anthony.Gould@rlt.ulaval.ca

Jean-Etienne Joullié
Gulf University for Science and Technology, Mubarak Al-Abdullah, Kuwait, jejoullie@gmail.com

Recommended Citation

'Parallel Universe' or 'Proven Future'? The Language of Dependent Means *t*-Test Interpretations

Anthony M. Gould
Laval University
Quebec City, QC

Jean-Etienne Joullié
Gulf University for Science and Technology
Mubarak Al-Abdullah, Kuwait

Of the three kinds of two-mean comparisons which judge a test statistic against a critical value taken from a Student *t*-distribution, one – the repeated measures or dependent-means application – is distinctive because it is meant to assess the value of a parameter which is not part of the natural order. This absence forces a choice between two interpretations of a significant test result and the meaning of the test hypothesis. The parallel universe view advances a conditional, backward-looking conclusion. The more practical proven future interpretation is a non-conditional proposition about what will happen if an intervention is (now) applied to each population element. Proven future conclusions are subject to the corrupting influence of time-displacement, which include the effects of learning, development, and history. These two interpretations are explored, and a proposal for new conceptual categories and nomenclature is given to distinguish them, applicable to other repeated measures procedures derived from the general linear model including ANOVA.

Keywords: *t*-test, parameter, dependent-means, language

Introduction

In the social sciences, whether knowledge is a socially constructed discourse or refers to a stable and given reality that is objectively accessible at least in principle is a question that has attracted contributions for decades. Statistical analysis, traditionally advanced as a means to measure reality, has not been spared criticism. It has been compared to storytelling and thus viewed as a form of conventional discourse (Ainsworth & Hardy, 2012). Although not taking a side in this debate, this articles illustrates the difficulty resulting from the belief the notion of a natural order refers to a given something that can be objectively known. Indeed, when

---

Dr. Anthony M. Gould is a Full Professor of Employment Relations. Email him at: Anthony.Gould@rlt.ulaval.ca.

making statistical inferences it is sometimes difficult to specify a parameter and to formulate an accurate linguistic description of what a result actually means.

There are two types of reasons why this is so: The first arises from contextual elements concerning the problem or research question itself (Lewin & Somekh, 2011; Tajfel & Fraser, 1986). In practice, this kind of concern manifests when a population is difficult to discern. The second – that which is the focus of this article – occurs because in some circumstances parameters literally do not exist because they are merely conceptual a-priori and post-hoc to an analysis. For example, assuming no control group (elements acting as their own controls) in early drug trial situations, a sample of rats may show tumor reduction following treatment with a putative anti-cancer agent. In such a case, because the treatment has been applied only to the sample, the parameter does not exist in the population at the moment the statistic is calculated. Other examples exist in diverse research paradigms, including: counselling intervention research, organizational development research, and where economic interventions are being assessed. The absence of parameter seems paradoxical, because parameter estimation is the *raison d'être* of parametric statistics.

There is a distinction between dependent means *t*-tests and other mean-related *t*-test applications. Parameters for dependent means *t*-tests literally do not exist. The problem is not that certain parameters are theoretical in the sense that the sampling distribution of means, Student *t*-distributions and the central limit theorem are abstract natural phenomena that conveniently support the logic of an analysis and can as such be simulated. Rather, the point is that dependent means *t*-test parameters are simply not out there to be discovered.

There are two possible interpretations of this non-existence and therefore two ways to interpret a significant dependent means *t*-test result. One of these is conceptually and technically sound but somewhat impractical; it is referred to here as the 'parallel universe' view, because it invites whoever reads about research results to imagine an alternative reality in which an entire population has been subjected to an intervention, rather than just a subset of elements. The other interpretation is less theoretically defensible but more practical; it is called here the 'proven future' view, because it corrals consumers of research to accept the proposition that the future is solely and exclusively determined by the past such that, if the intervention is applied to all subsequent cases, it would yield an outcome comparable to the sample-based result.

One-sample, dependent means, and independent means *t*-tests are intended to identify an actual state of reality, albeit one that it is not easily discoverable and therefore must be inferred from observations of samples (or more precisely sample

statistics). However, of the various two-mean comparisons available, one in particular – the repeated measures procedure – requires that a distinction be made about which of two possible interpretations should accompany a decision to reject the null hypothesis. The literature addressing the mechanics of two-mean comparisons as well as that discussing advanced repeated measures procedures which use the general linear model (e.g. ANOVA-based analyses) mostly either overlooks or does not well elucidate this point. This is unfortunate, because there are methodological and conceptual advantages that flow from giving a more nuanced understanding of the consequence of a missing parameter. Such benefits concern, from an applied perspective, interpretation; and, from a teaching/explanatory perspective, a deepening of understanding. Whatever the case, the existence of two possible interpretations of a dependent means *t*-test result has implications for experimental control which mostly have not received enough attention. They are sufficiently important to necessitate the creation of a new nomenclature and a new way of distinguishing between population frequency distributions.

Three Research Designs Necessitating a Two-Mean Comparison: The Dependent Means Case as Special

There are so-called parametric data analysis situations where population parameters literally do not exist. This is not to say that their values are impractical to calculate, nor does it imply that an understanding of their nature is not as important as it always was. Rather, some parameters do not exist in the sense that they cannot, in theory or practice, be calculated. Hence, when speaking of a parameter, for example in a dependent means *t*-test, it is especially important to be clear about the meaning of a significant test statistic and the associated decision to reject the null hypothesis. Textbooks as well as many studies that use a dependent means protocol for data analysis typically give this matter only cursory consideration (e.g. Mason, Lind, & Marchal, 1999; Gravetter & Wallnau, 1995; Wright, 1997; Baillargeon, 2012; Salkind, 2011). Such superficial or dismissive treatment leads to inadequate control of time-dependent and potentially confounding variables, and as a consequence, to imprecise or erroneous conclusions.
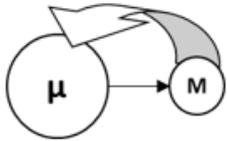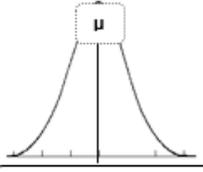
In the quest to produce statistically significant results, data analysts frequently advocate dependent means designs to reduce a sampling distribution's variation and to create a greater *t*-value, arguing that fewer degrees of freedom is a price worth paying for a smaller test statistic denominator (e.g. Gravetter & Wallnau, 1995). However, those who are less concerned with statistical significance and more

interested in contextual and ethnographic elements of a problem often favor between- over within-subjects designs (e.g. Lewin & Somekh, 2011; Adams, Khan, Raeside, & White, 2007). For these latter theorists, carry-over effects and other, more general concerns about experimental control are especially important. Such researchers are mostly satisfied that, in lieu of a control group, a matched-pair design where subjects act as their own controls is practical despite being potentially a theoretically compromised solution (Lewin & Somekh, 2011; Alasuutari, Bickman, & Brannen, 2009).
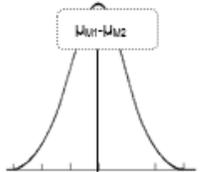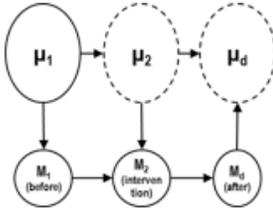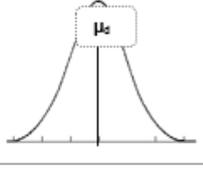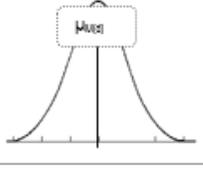
Repeated measures procedures are sometimes viewed as being plagued by the problem of time-related confounding influences (e.g. Cousineau, 2009). This concern is more fruitfully analyzed as the manifestation of an inexistent parameter. Such a perspective makes clear that two options for interpreting a significant result are possible. To understand what is meant by an inexistent parameter, three representations of typical two-mean comparison situations are presented in the first row of Figure 1. Beside each representation is a depiction of the population distribution, the sampling distribution of the mean, and formulae for calculating a test result (to be compared with an appropriate critical value drawn from a Student t-distribution when testing hypotheses). To improve clarity and concision, confidence interval formulae are not presented in Figure 1, only hypothesis testing and rejection of the null hypothesis are discussed. The conclusions and insights offered are equally relevant to confidence interval applications. Furthermore, such findings can be extended to more advanced applications of the general linear model such as ANOVA-based procedures.

In row 1.1 of Figure 1, the one-sample case, a statistic (mean) is calculated and indirectly compared to a parameter which actually exists but which is 'hidden', difficult or impractical to discover. (The statistic is compared indirectly because it is compared with the mean for the sampling distribution of the mean that is equal to but not the same as the population mean.) The fact that the parameter is hiding in such situations can be appreciated with a thought experiment. Imagine that, at the same time that a statistic is being calculated using the formula in the last cell of row 1.1, another person is calculating the actual population parameter (μ). In such a case, the aim would be to see how close an obtained statistic falls from the calculated specified population value. Now also imagine that an analyst substituted the parameter (mean) for the population (μ) for the parameter (mean) of its sampling distribution of the mean (μ_M). This manipulation would allow an obtained statistic (M) to be compared with a parameter value of interest (μ) rather than with its proxy value, the mean for the sampling distribution of the mean (μ_M). If this

Graphic Depiction of key elements of analysis	Depiction of the Population Frequency Distribution[1]	Depiction of the Sampling Distribution of the Mean[2]	Formula for calculating a test statistic
1.1 One-Sample t-test μ → M	μ Distribution exists as part of the natural order	μM	$t = (M - \mu_M)/\tilde{o}_M$ (with n - 1 degrees of freedom)
1.2 Independent Means t-test μ₁ μ₂ μ₁- μ₂ M₁ → M₂ → M₁-M₂	μ₁-μ₂ Distribution exists as part of the natural order	μ_M1-M2	$t = \{(M_1 - M_2) - (\mu_{M1} - _{M2})\}/\tilde{o}_{M1-M2}$ (with n₁ + n₂ − 2 degrees of freedom)
1.3 Dependent Means t-test μ₁ μ₂ μd M₁ (before) → M₂ (intervention) → Md (after)	μd Distribution **does not** exist as part of the natural order	μ_M(d)	$t = (M_d - \mu_{M(d)})/\tilde{o}_{M(d)}$ (with n - 1 degrees of freedom)
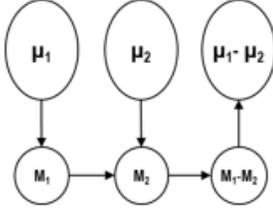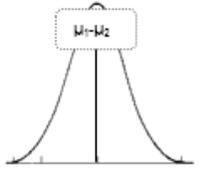
Figure 1. Key elements of three kinds of two-mean comparison *t*-test applications

Note: [1]Depicted as normal although, due to the Central Limit Theorem, a normally distributed population frequency is not essential for applying a *t*-test procedure. [2]Depicted as normal because it is assumed that samples used to create the sampling distribution of the mean are of a size $n > 12$ and $n < 30$ (Central Limit Theorem)

were to occur, $t = (M - \mu_M) / \tilde{o}_M$ would become $t = (M - \mu) / \tilde{o}_M$. Like the orthodox technique, this manipulation would yield the correct result. However, it would be bizarre because it would be standardizing a score (μ) from a non-related distribution,

the population distribution, with two elements (M and $\tilde{o}_M$) from the sampling distribution of the mean. Aside from being conceptually unsound, this change in formula adds an arduous and unrealistic additional step to the procedure.

However, parameter substitution could be done and could be synchronized to coincide with calculation of the test statistic (t_{obtained}). In such a case, the analyst would make the rejection/non rejection decision at the same moment that they became aware of the real population mean. They would have used a procedure which bypasses the step which relies on the central limit theorem to prove that $\mu = \mu_M$ (Gravetter & Wallnau, 1995; Wright, 1997; Baillargeon, 2012). One-sample t-test situations of the kind just described (and depicted in row 1.1 of Figure 1) are somewhat rare but nonetheless occasionally important. For example, to measure the temperature of bath water, it is possible to put a thermometer in the bath itself or to take a cupful of water as a representative sample of the liquid and plunge a thermometer therein to record a reading. In this latter case, the researcher would subsequently make an inference about the bath temperature from the cup temperature.

Row 1.2 of Figure 1 is a depiction of an independent means application of a t-test. This technique has some conceptual similarity with the one-sample t-test case discussed above; for example, the parameter of interest here ($\mu_1 - \mu_2$) exists to be discovered before, during, and after the calculation of a test statistic. Once again, in a sense, the parameter may be thought of as hiding, but nonetheless part of the natural order. Another thought experiment makes this clear. Imagine that there was a binary independent variable, say gender, taking two possible values, male or female in this instance. Imagine that an interval-scaled dependent variable was height; the hypothesis for testing (H_1) was that males are taller than females in the population; and the null hypothesis (H_0) was that males are not taller than females in the population. The characteristics of this problem require analysis using an independent means t-test. Specifically, a test t-value (t-statistic) could be calculated to be compared with a critical value drawn from a Student t-distribution (with $n_1 + n_2 - 2$ degrees of freedom). In such a case, the statistic ($M_M - M_F$) is tangible. Further, at the instant of its calculation, there also exists a real, equally tangible population mean for height for males (μ_M) and a real, equally tangible population mean for height for females (μ_F). Although these two population values and their difference ($\mu_M - \mu_F$) may be difficult or impractical to calculate, they are a real part of the natural order that is perhaps hidden but nonetheless there at the moment that a t_{obtained} value is being compared to a t_{critical} value.

When it comes to the dependent means t-test – a protocol which mostly aims to assess the consequence of an intervention – the parameter does not exist at the

moment a statistic is calculated. It is not part of reality because each element of the population is not yet present in the post-intervention state. Hence, in such a case, a $t_{obtained}$ cannot be calculated through substituting μ_d for $\mu_{d(mean)}$. The broken-line depictions in the first cell of row 1.3 of Figure 1, where the distributions of (μ_2) and (μ_d) are stylized portrayals of the population in the post-intervention state, indicate this state of affairs. As such, together they represent one member of a family of future scenarios. Textbook authors are typically unclear about this point, with disparate recommendations offered for drawing a conclusion (e.g. Adams et al., 2011; Cousineau, 2009; Mason et al., 1999; Gravetter & Wallnau, 1995; Wright, 1997; Baillargeon, 2012; Salkind, 2011). In fact, scholars are inclined to offer one of either two antithetical ways of addressing this missing parameter and its associated missing sampling distribution. The first of these is technically and conceptually correct but impossible to operationalize (e.g. Gravetter & Wallnau, 1995; Wright, 1997; Levin & Fox, 2000); it will henceforth be referred to as the 'parallel universe' view.

For example, Gravetter and Wallnau (1995), in commenting on a significant result for a matched-pair intervention procedure aimed at controlling asthma symptoms using relaxation, concluded, "Relaxation training resulted in a decrease in the number of doses of medication needed to control asthma symptoms. This reduction was statistically significant, $t(4) = -3.72$, $p = 0.05$, two-tailed" (p. 256). Similarly, Wright (1997) interpreted the result of a significant dependent-means *t*-test by saying "an effect was detected" (p. 53). Furthermore, Levin and Fox (2000), in drawing a conclusion about a significant dependent-means *t*-statistic concerning the efficacy of a remedial math intervention, stated, "The remedial math program has produced a statistically significant improvement in math ability" (p. 227). In each of these cases, past tense conditional verb conjugation was used. Hence, in rejecting the null-hypothesis, a backward-looking conditional inference was invoked with the linguistic structure: If each element of the population had been subjected to the same intervention that was applied to the sample, there would have been a difference in the pre- and post-intervention population means.

The second way of dealing with the missing parameter conundrum is to assert or assume tacitly that an intervention on the population of the kind that was applied to the sample will affect the population as it affected the sample (e.g. Mason et al., 1999; Levin & Rubin, 1998; Elliott & Woodward, 2007). Such conclusions are often parsed in the present or future tense. For example, Mason et al. (1999), wrote "is a difference" (p. 369) to describe the state of a population following a significant dependent-means *t*-test result. This conclusion has a prospective focus. Similarly, Levin and Rubin (1998), in commenting on significantly improved typing speeds

for secretaries using new word-processing software in a pre-test/post-test design, concluded "The difference in typing speed can be attributed to the different word processors" (p. 473). A similar conclusion was offered by Elliott and Woodward (2007) who, in interpreting the effectiveness of a weight-loss regime following a significant dependent means t-test result stated, "We reject H_0 and conclude that the diet is effective; $t(14) = 2.567$, one-tailed, $p = 0.001$" (p. 73).

These interpretations of a dependent-means protocol make a claim about the future. Owing to control-related confounding influences arising from time-displacement, such 'proven future' inferences are technically and conceptually more spurious than those concerning a parallel universe. However, proven future interpretations are attractive because they are inclined to be practical. They are, in a sense, the *raison d'être* for a repeated measures protocol (Fortin, Côte, & Filion, 2006).

These examples were extracted from data analysis textbooks. Textbook authors typically do not give a clear rationale for the way their conclusions are formulated, which makes them appear arbitrary. It is therefore not surprising to find that the applied literature perpetuates this lack of clarity. However, this literature also reveals some patterns. For example, medical research tends to favor a proven future interpretation. Thus, Kutcher, Wei, and Morgan (2015) concluded their study by noting that "these results [i.e., those obtained from their intervention] suggest a simple but effective approach to improving MHL [mental health literacy] in young people" (p. 580). Comparable inferences are found in studies investigating educational-type interventions. According to Baykara, Demir, and Yaman (2015), "ethics education given to students enables them to distinguish ethical violations" (p. 661); Azarbarzin, Malekian, and Taleghani (2015) inferred "supportive-educative programs can enhance some aspects of quality of life" (p. 577). Additional examples are Lau, Li, Mak, and Chung (2004), Scott and Graham (2015), Garst and Ozier (2015), or again, Pritchard, Hansen, Scarboro, and Melnic (2015). This dominance of the proven future interpretation is understandable because medical or educational interventions are intended to be therapeutic and/or remedying. In such circumstances, incentives often exist strongly to push the case for a putative treatment.

Compared with proven future interpretations, parallel universe inferences are less commonly found in the applied literature, although they do exist. For example, Fee, Gray, and Lu (2013), in commenting on improvement in cross-cultural awareness following a stint in a foreign country, reported "expatriates' level of cognitive complexity increased significantly during the 12-month study period" (p. 299).

Two Interpretations of a Significant Dependent-Means *t*-Test Result: Parallel Universe versus Proven Future

The parallel universe interpretation of a significant dependent-means *t*-test result uses past-tense conditional verb conjugation to describe what would have happened if each member of a population had been subjected to the intervention which was applied to the sample. This emphasis is on an imagined alternative reality in which each element of a population is subjected to a treatment protocol at the same moment that elements of the sample were so subjected. Remembering that such an interpretation is offered following the finding of a significant *t*-test result, it can only be hypothetical because it is not possible to go back in time. In this sense, analysts who rely on it offer a conclusion that has limited practical utility.

Alternatively, a significant dependent means *t*-test result may be used to reject the null hypothesis with the conclusion that, if an intervention is applied to all members of a population, the post-hoc mean would be different to the a-priori mean. This proven future statement has implications for practice. However, it is less likely than the parallel universe view to reflect reality, because it is vulnerable to a source of control-related error, namely time-displacement effects. Time-displacement is an umbrella term covering at least three circumscribed classes of phenomena: development/maturation, non-treatment-related learning, and historical events (Fortin et al., 2006). Each of these has implications for interpretation of a significant dependent means *t*-test finding that deserve further discussion.

Development/maturation is a relatively permanent change in behavior, values, or cognition that cannot be accounted for by experience or a health-related event such as illness or injury (Demetriou, 1998; Upton, 2011). For example, normally developing babies do not learn to walk (Upton, 2011); rather, regardless of whether they have been trained or otherwise instructed, babies typically take their first steps at between 11 to 15 months of age with a normal distribution of habitual infant bipedalism centered on a mean of 13 months (Upton, 2011). Suppose there were theoretical grounds for challenging the maturational perspective of the emergence of upright walking behavior in healthy human beings and that it was possible to train babies to walk earlier than they otherwise would. Further, suppose that there was a misconception about when infants generally walk upright and that it was believed that they mostly crawl until they are at least 18 months old. In such a case, it would be worthwhile to take a sample of 11 to 12-month old babies, measure their propensity towards bipedalism on a suitable scale, subject them to 'mobility-training', and then re-measure them on the same instrument in a matched-pair one-tailed hypothesis-test design.

If this analysis did not yield a significant dependent means t-test result, it would have if the three phases of the study (pre-test measurement, intervention, and post-test measurement) had been instituted over a longer period. What this suggests is that a significant test result can be caused by a third, maturation-related variable and thus have nothing to do with the intervention. However, it is also possible for a particular cohort of babies ('late walkers'), a treatment intervention decreases the mean age of bipedalism from 14 months to 13 months. To demonstrate such an effect through a repeated measures protocol, the delay between each phase of the study must be kept minimal; the longer the delay, the less informative would be the assessed post-intervention state. Hence, when there are two population frequency distributions depicted on a single axis with values representing points in time as levels of an independent variable (prior to walking training and following walking training in this example), a proven-future interpretation of a significant test statistic is likely to be misleading. Hence, claims about what will occur remain at best ambiguous and at worst spurious when maturation is an alternative explanation for change on the dependent variable. A key problem here is that the existence of such a competing explanation is not necessarily known.

Learning is a relatively permanent change in behavior, value, or cognition that occurs as a consequence of experience and not as a result of either maturation or a traumatic health-related event (Schacter, Gilbert, & Wegner, 2011). For example, school district officials may implement a stranger-danger initiative to discourage children from accepting lifts with adults who they do not know. The creators of such an initiative want to determine if their intervention changes behavior. They institute a pre-test/post-test repeated measures assessment protocol (before and after the stranger-danger intervention using a sample) and establish a suitable index of childhood propensity to accept rides from strangers as a dependent variable.

In a case like this, the null hypothesis is typically that the stranger-danger intervention does not influence children to be less inclined to accept a ride from people they do not know and the one-tailed test hypothesis is that the stranger-danger intervention makes children less inclined to accept a ride from people they do not know. Suppose that a significant t-test value for this analysis is not obtained at the orthodox Type-1 error rate of $\alpha = 0.01$, and therefore the null hypothesis cannot be rejected. However, the researchers notice that if they make the test slightly less conservative, say by adjusting the Type-1 error rate to $\alpha = 0.05$, they obtain a significant result and can therefore reject the null hypothesis. Further, assume that, on the night of the intervention, there is a lead news story about the abduction and murder of a child who accepted a ride from a stranger.

This story is followed up over the ensuing days. In such a case, the timing of the three elements of the study, once again, becomes important. From a practical standpoint, it is likely that the almost significant *t*-test result would underestimate the effect of the intervention if it were – in the future – carried out on all members of the population. Hence, in this example, a *t*-test that was non-significant on one day may be significant on the following day (i.e., after the nightly news). This phenomenon, once again, highlights the importance of keeping the amount of time between pre-measurement, intervention, and post-measurement to a minimum, although it is unclear how small that minimum should be. What is generally true is that, the more protracted this delay, the more the subjects (or sample elements) are exposed to stimuli which can elicit learning of an unplanned and uncontrolled, but nonetheless systematic, nature.

An historical occurrence is a time-displacement effect that can be viewed as a special case of learning; special in the sense that it is not cyclical or typical. (Certain events are rare or one-off in nature; they are not amenable to measurement even if they can be said to create a 'new normal.' The assassination of President Kennedy is a case in point: it was an unprecedented event in twentieth-century American history and created new and enduring anxieties about the welfare of political leaders. Conversely, child abductions and murders are unfortunately recurring events in large societies; new cases regularly appear and their occurrence can be quantified in probabilistic terms where a numerator is non-zero and is all instances where an event could have occurred.)

For example, suppose that on September 10, 2001, there is a study taking place in New York City that aims to determine if a particular intervention intended to help those who are anxious about flying in an airplane overcome their fear. Following a pre-test/post-test pairing protocol, a dependent means *t*-test is used to analyze data in a one-tailed improvement versus no-improvement hypothesis. A significant test result is obtained, the null hypothesis is rejected and the researcher concludes that the intervention is efficacious. As noted, however, two interpretations are possible. The first (the parallel universe view) says, all things being equal and if all members of the population of interest had received the intervention at the time the study was being carried out, there would have been a mean improvement on the dependent variable (fear of flying). Such an interpretation, although not especially useful because it is purely hypothetical, controls best for potential historical influences; indeed, its linguistic formulation makes it invulnerable to competing explanations arising from time-displacement.

The alternative interpretation (the proven future inference) is misrepresentative in the case described. On September 11, 2001, there were high-

profile terrorist attacks directed towards targets in New York and elsewhere in the United States involving hijacked commercial airplanes. A significant result for the efficacy of a fear of flying initiative obtained on September 10 would presumably not have been produced if the pre-test measurement was made on September 10 and the intervention (and measurement protocol) instituted on the afternoon of September 11.

Conclusion

Using the case of the *t*-test as an exemplar of other repeated measures designs such as ANOVA, the emphasis here is on interpreting matched-pair hypothesis test results. Questions concerning the strengths and weaknesses of between- versus within-subjects designs have not been addressed. Similarly, the focus has not been on what to do about time-displacement confounding effects, when they are likely to occur, or on techniques for controlling for them. Rather, the focus here has been merely on interpretation. In studies that investigate, by way of a dependent means protocol, the value of a parameter which does not exist as part of the natural order, two types of interpretations are possible when the *t*-test result is significant: parallel universe or proven future. Although each of these is equivalent in its description of what is being observed, each differs in its practical utility.

The parallel universe view is best suited to circumstances where it is reasonable to believe that time displacement effects are credible alternatives to an observed change in the dependent variable following manipulation of the independent variable. Such cases typically exist in (but are not limited to) the social sciences, even when researchers believe that the reality they study is unaffected by their investigations. Although within-subjects designs have the potential to reduce the standard error of the mean and hence increase the likelihood of statistical significance, they control for a more limited range of confounding influences than between-subjects designs. This disadvantage is not insurmountable unless and until time displacement effects can offset or compound the changes introduced when an independent variable is manipulated. In such instances, theorists typically propose the idea of a control group as the solution (Lewin & Somekh, 2011; Adams et al., 2011; Tajfel & Fraser, 1986). However, the creation of a control group is not always practicable.

The alternative is the proven future interpretation. In the physical sciences, the possibility of time-displacement effects are perhaps lesser than in the social sciences. In these kinds of situations, 'proven future' views are often easier to justify; intuitively this seems reasonable because the past is always the best

predictor of the future in the same sense that gravity, for example, can always be relied upon to occur. It is therefore in the social sciences where the choice between the two interpretations of a within-subjects finding becomes especially relevant.

Irrespective of their field of study, researchers, because of the nature of what they do (in particular when engaging in funded research where significant results are typically those that are rewarded), are inclined to favor the proven future interpretation of significant dependent-mean *t*-test results. Two possible factors influence this phenomenon. First (the focus here), not enough consideration is given to the protocol of interpretation. The second is more psychological and arises from researchers' desires to be consequential in their endeavors. Where parallel universe interpretations exist in the literature, they tend to be made implicitly as if researchers lacked confidence to make an inference about their population of interest.

Whatever the case, it is noteworthy that the choice between a parallel universe and a proven future interpretation represents a trade-off: the former is less prone to confounding influences but is less practical (no one can go back in time), the latter is more practical but more prone to confounding influences from time displacement effects. In reporting results, this compromise should be acknowledged and its consequences with respect to the research problem emphasized. When explaining procedures and depicting the elements of an analysis in diagrammatic form, the trade-off perspective can be highlighted through using broken-lines to indicate those frequency distributions which can only occur in the future. Such a convention could be used to signal that certain distributions exist only in a particular future, one which is contingent on the ubiquitous presence of the second level of the independent variable (i.e. the post-intervention state). Two such contingent within-subjects future distributions (which could be depicted with broken lines) would be the population distribution after all elements of the population have received an intervention (μ_2; σ_2) and the difference between each element of a population before and after an intervention (μ_d; σ_d). Use of such a nomenclature would flag the distinctiveness of the repeated measures procedure and serve as a reminder that certain parameters (e.g. μ_2; σ_2; μ_d; σ_d) are missing for the moment and contingent on a forthcoming population-wide intervention.

References

Adams, J., Khan, H., Raeside, R., & White, D. (2011). *Research methods for graduate business and social science students* (7th ed.). Los Angeles, CA: Sage. doi: 10.4135/9788132108498

Ainsworth, S., & Hardy, C. (2012). Subjects of inquiry: Statistics, stories, and the production of knowledge. *Organization Studies, 33*(12), 1693-1714. doi: 10.1177/0170840612457616

Alasuutari, P., Bickman, L., & Brannen, J. (Eds.). (2009). *The SAGE handbook of social research methods*. London, UK: Sage. doi: 10.4135/9781446212165

Azarbarzin, M., Malekian, A., & Taleghani, F. (2015). Effects of supportive ‑educative program on quality of life of adolescents living with a parent with cancer. *Iranian Journal of Nursing and Midwifery Research. 20*(5), 577-581. Retrieved from http://ijnmr.mui.ac.ir/index.php/ijnmr/article/view/1236

Baillargeon, G. (2012). *Méthodologie et techniques statistiques avec applications en management et relations industrielles* [Statistical methodology and techniques with applications in management and industrial relations] (3rd ed.). Trois-Rivières, QC: Les Éditions SMG.

Baykara, Z. G., Demir, S. G., & Yaman, S. (2015). The effect of ethics training on students recognizing ethical violations and developing moral sensitivity. *Nursing Ethics, 22*(6), 661-675. doi: 10.1177/0969733014542673

Cousineau, D. (2009). *Panorama des statistiques pour psychologies: Introductions aux méthodes quantitatives* [Survey of statistics for psychology: Introduction to quantitative methods]. Brussels, Belgium: De Boeck.

Demetriou, A. (1998). Cognitive development. In A. Demetriou, W. Doise, & C. F. M. van Lieshout (Eds.), *Life-span developmental psychology*. Chichester, NY: J. Wiley & Sons.

Elliott, A. C., & Woodward, W. (2007). *Statistical analysis quick reference guidebook*. Thousand Oaks, CA: SAGE. doi: 10.4135/9781412985949

Fee, A., Gray, S. J., & Lu, S. (2013). Developing cognitive complexity from the expatriate experience: Evidence from a longitudinal field study. *International Journal of Cross Cultural Management, 13*(3), 299-318. doi: 10.1177/1470595813484310

Fortin, M.-F., Côte, J., & Filion, F. (2006). *Fondements et étapes du processus de recherche* [Foundations and stages of the research process]. Montréal, QC: Chenelière Éducation.

Garst, A. B., & Ozier, L. W. (2015). Enhancing youth outcomes and organizational practices through a camp-based reading program. *Journal of Experiential Education, 38*(4), 324-338. doi: 10.1177/1053825915578914

Gravetter, F. J., & Wallnau, L. B. (1995). *Essentials of statistics for the behavioral sciences* (2nd ed.). New York: West Publishing Company.

Kutcher, S., Wei, Y., & Morgan, C. (2015). Successful application of a Canadian mental health curriculum resource by usual classroom teachers in significantly and sustainably improving student mental health literacy. *The Canadian Journal of Psychiatry, 60*(12), 580-586. doi: 10.1177/070674371506001209

Lau, E. Y. Y., Li, E. K. W., Mak, C. W. Y., & Chung, I. C. P. (2004). Effectiveness of conflict management training for traffic police officers in Hong Kong. *International Journal of Police Science & Management, 6*(2), 97-109. doi: 10.1350/ijps.6.2.97.34468

Levin, J., & Fox, J. A. (2000). *Elementary statistics in social research* (8th ed.). Boston, MA: Allyn and Bacon.

Levin, R. I., & Rubin, D. S. (1998). *Statistics for management* (7th ed.). Upper Saddle River, NJ: Prentice-Hall.

Lewin, C., & Somekh, B. (2011). *Theory and methods in social research* (2nd ed.). Los Angeles, CA: SAGE.

Mason, R. D., Lind, D. A., & Marchal, W. G. (1999). *Statistical techniques in business and economics* (10th ed.). New York: Irwin/McGraw Hill.

Pritchard, T., Hansen, A., Scarboro, S., & Melnic, I. (2015). Effectiveness of the sport education fitness model on fitness levels, knowledge, and physical activity. *The Physical Educator, 72*(4), 577-600. doi: 10.18666/tpe-2015-v72-i4-6568

Salkind, N. (2011). *Statistics for people who (think they) hate statistics* (4th ed.). Thousand Oaks, CA: SAGE.

Schacter, D. L., Gilbert, D. T., & Wegner, D. M. (2011). *Psychology* (2nd ed.). New York, NY: Worth Publishers.

Scott, K. E., & Graham, J. A. (2015). Service-learning: Implications for empathy and community engagement in elementary school children. *Journal of Experiential Education, 38*(4), 354-372. doi: 10.1177/1053825915592889

Tajfel, H., & Fraser, C. (1986) *Introducing social psychology*. London, UK: Penguin Books.

Upton, P. (2011). *Developmental psychology*. Exeter, UK: Learning Matters.

Wright, D. B. (1997). *Understanding statistics: An introduction for the social sciences*. Thousand Oaks, CA: Sage Publications.