Reply to the ANNEX "Errors and issues with Kaul and Wolf's two working papers on tobacco plain packaging in Australia"

Prof. Dr. Ashok Kaul[*]
Institute for Policy Evaluation (IPE), Saarland
and Department of Economics, Saarland University

Prof. Michael Wolf, Ph.D.[†]
Department of Economics
University of Zurich

February 11, 2015

Abstract

The annex to the letter with subject matter "Request for retraction of two papers published on UZH website" by Pascal A. Diethelm lists seven (so-called) "errors" and seven "issues"[1] with our two working papers Kaul and Wolf (2014a,b). In doing so, this annex is supposed to provide the detailed reasons for the request to retract the two working papers from the UZH website. As we show in this reply, no such reasons exist. Although there are some (minor) points of debate, there is not a single *"extremely serious"* error in our two working papers.[2] The only serious errors to be found are in the annex instead. Finally, the annex makes numerous false statements regarding our two working papers that fall well outside the bounds of credible scholarly debate and instead appear to be nothing more than advocacy.

1 Lack of Authorship

We briefly point out that the annex to the letter by Pascal A. Diethelm is signed only by "OxyRomandie". We wonder why the authors of the annex are not listed? What are their names and their qualifications? Why do they hide in anonymity? Who assumes the responsibility for the claims and allegations in this annex?

[*] Saarland University, Department of Economics, Campus Building C3.1, D-66123 Saarbrücken, Germany. Email: a.kaul@ipe-saarland.de.

[†] University of Zurich, Department of Economics, Wilfriedstrasse 6, CH-8032 Zurich, Switzerland. Email: michael.wolf@econ.uzh.ch. Michael Wolf is corresponding author.

[1] The annex contains seven "issues" in the body of the annex; however, the summary page of the annex omits one of the "issues" and lists only six of them.

[2] Mr. Diethelm requests the retraction our working papers from the University of Zurich website because of seven *"extremely serious"* errors.

2 So-Called Errors

We now respond to the seven allegations of "errors" listed in the annex. They all refer to two working papers Kaul and Wolf (2014a,b). In a nutshell, there is no substance to these claims. Furthermore, some of the claims reveal a surprising lack of basic statistical knowledge. We accept some of the arguments of the authors as (equally) legitimate standpoints in a controversial debate. But we are flabbergasted by the conclusion that the letter of Mr. Diethelm draws, namely that the University of Zurich should retract our two working papers. We appreciate OxyRomandie's contribution (even though their view is opposite to ours) to the academic debate. Instead of urging the retraction of papers with an opposite view, they should recall the principles of academic freedom, namely the belief that the freedom of inquiry by faculty members is essential to the mission of academia.

2.1 So-Called Error # 1: Erroneous and misleading reporting of study results

OxyRomandie criticizes the *"erroneous and misleading reporting of study"*. This critique is not justified for the following reasons.

1. We (the authors) have always accurately described the study results including in the statements cited by OxyRomandie. There is nothing in these statements that is not supported by our research conducted in the two UZH working papers under consideration.
2. PMI in its statements, from which the OxyRomandie critique cites selectively, has also provided a fair characterization of the results, in the context of the overall debate and state of evidence (which is anyhow not limited to our two working papers). In particular, this is also true for the PMI media release of March 2014 cited by the authors of the annex. Importantly, that media release has the headline *"Researchers Find No Evidence Plain Packaging 'Experiment' Has Cut Smoking"*. The same media release also accurately quotes the study as saying that *"no such evidence [for an effect of plain packaging] has been found."* It is correct that the media release also contains the selectively quoted language that the *"plain packaging experiment in Australia has not deterred young smokers."* However, from the context provided by both the headline and the details of the study it is clear for any reasonable reader that what the study found was "no evidence for an effect" as opposed to "evidence for no effect".

   The same is true for the discussion of the study in PMI's response to the UK consultation. Contrary to what the authors of the annex claim, that response was clear in saying that the study found no evidence for an effect.

   In any event, the results of our working papers speak for themselves, and what PMI (or any other stakeholder for that matter) says about the study when engaged in the political debate cannot be attributed to us.

Indeed, the fact that the papers are publicly available allows any interested reader to refer to our original work to verify claims made in the public policy debate. OxyRomandie's efforts to censor our working papers would rob the public of the opportunity to review our actual research and conclusions.

3. Finally, it is worthwhile recalling the public policy context. Supporters of the plain packaging policy, in Australia and elsewhere, have claimed that the policy will lead to a sizeable reduction of smoking rates within a relatively short period of time, especially among children; for example, see Pechey et al. (2013). This, in essence, is the hypothesis that we have been testing with our research. The results of the research did not find evidence supporting the validity of the hypothesis.[3]

   It can hardly be surprising that non-statisticians involved in the political debate about plain packaging would occasionally characterize our research as showing that "plain packaging doesn't work", although in reality there is only an absence of evidence that it does work. The authors of the annex are free to criticize such statements, but the twisted conclusion that such statements would render our two working papers themselves flawed has obviously no basis whatsoever.

In sum, labeling this point an "error" is factually wrong.

2.2 So-Called Error #2: Power is obtained by sacrificing significance

The authors of the annex claim we made an "error" by sacrificing statistical significance in order to gain power. They claim that our inference method is defective and that our conclusions are invalid. These statments either deliberately misrepresent our approach, analysis, and conclusions or represent a fundamental lack of expertise in the field of statistics.

Our approach is to give most leeway to finding a plain packaging effect over the 13 months in question (that is, from December 2012 through December 2013). We therefore accept the lowest (pointwise) confidence level of the three levels most widely used in applied research. These three levels are 90%, 95%, and 99%; so we use the level 90%. As a consequence, the probability of falsely establishing a plain packaging effect in the data when none exists in reality is actually quite high at 10%; this probability is called the significance level.[4]

No chosen significance level is wrong *per se*. There is well-known trade-off between the significance level of a test (that is, the probability falsely establishing an effect in the data

[3] We are not the only researchers who conclude there is no evidence for the effectiveness of plain packaging; for example, see Davidson and de Silva (2014).

[4] To be fair, this probability is for falsely detecting a plain packaging effect in either direction: negative or positive. Here, a negative effect corresponds to a reduction in smoking prevalence and is thus an intended effect; on the other hand, a positive effect corresponds to an increase in smoking prevalence and is thus an unintended effect.

when none exists in reality) and the power of a test (that is, the probability of failing to establish an effect in the data when one does exist in reality): *Ceteris paribus*, the lower the significance level, the lower power. So if one chooses a lower significance level (in order to safeguard against the probability of falsely establishing an effect in the data), one necessarily 'sacrifices' power. On the other hand, if one wishes more power, one necessarily has to 'sacrifice' in terms of the significance level (where here a 'sacrifice' means an increase). In sum, the higher the significance level, the more likely it is to establish a plain packaging effect by mistake. High significance levels should therefore be a concern to those parties interested in *not* establishing a plain packaging effect. On the other hand, tobacco control advocates should welcome high significance levels (and thus equivalently low confidence levels), at least if they are convinced that plain packaging has the intended effect of reducing smoking prevalence.

The main analysis and findings of both our working papers are based on pointwise confidence intervals; we include the analysis based on uniform confidence intervals 'only' as a variation (among many other variations) and this is cleary stated in the papers. No errors and no unusual choices (of the confidence level, say) have been made at any point in the paper.

It is then entirely consistent to base the formal power analysis also on pointwise confidence intervals. It is true, as the authors of the annex point out, that the power against a "zero" effect ($\Delta = 0$) in the working paper Kaul and Wolf (2014b) is 0.50 (after rounding). But this does not make our power values irrelevant.[5] Power is high against effects that could be reasonably expected from studying the tobacco control literature; for example, see Pechey et al. (2013).

The noteworthy fact and main finding is that if one searches for a plain packaging effect over the 13 pointwise confidence intervals, one still does not find evidence for any effect at all even though the 'global' significance level accounting for such data mining is around 0.5. The issue is similar in the working paper Kaul and Wolf (2014a), where if one searches for a plain packaging effect over the 13 pointwise confidence intervals, one only finds evidence for a single effect for December 2012 (where the 'global' significance level accounting for such data mining is again around 0.5). Our conclusions are therfore are not affected by the critique.

If we had based the main findings of the working papers on uniform confidence intervals (which to this date are still rarely used in practice), then the power analysis should also have been based on the uniform confidence intervals (which would have resulted in lower power values). It would have been reasonable for the authors of the annex to request an additional power analysis based on uniform confidence intervals as a robustness check. But labeling the existing power analysis as an "error" is factually wrong.

[5] Also, a significance level of 0.5, once data mining is properly taken into account, against a "zero" effect is not an error. Indeed, many analyses in top scientific studies have a significance level of similar magnitude, once data mining is properly taken into account. For example, this point is made by Heckman et al. (2010).

2.3 So-Called Error #3: Inadequate model for calculating power which introduces a bias towards exceedingly large power values

We make it very clear which model we use for our power analysis, namely an immediate and permanent reduction of smoking prevalence by Δ percentage points.

Of course, alternative models could have been used as well. It is often the case in applied statistics that competing models can explain a given set of data in a reasonable way. The problem is that nobody can claim to know what the unique 'right' model should be, since nobody can claim to know how plain package should affect smoking prevalence (if at all).

The authors of the annex promote a model of "gradual reduction". One can possibly think of plausible reasons for (and also against) such a model. On the other hand, one could also promote a model of an immediate effect that dies out over time (as plain packaging might have an initial, 'shock' effect to which people get used over time, resulting in the effect dying out over time).

Our model of an immediate and permanent reduction is in line with claims of prominent experts in the field. Professor David Hill, Director of Cancer Council Victoria (Australia), for example, predicted that *"plain packaging will slash smoking rates ... make significant inroads into reducing rates of smoking initiation and consumption ... [and] has enormous potential to cut smoking rates."*[6]. As such, our model is certainly a reasonable one.[7]

The authors of the annex could have requested additional power analyses based on alternative models of a plain packaging effect. But labeling the model we use as *a priori* inadequte or even as an "error" is factually wrong.

2.4 So-Called Error #4: Ignorance of the fact that disjunctive grouping of two tests results in a significance level higher than the significance level of the individual tests

This point is very much related to the claimed Error #2. Our main findings are based on the combined approach and so it is entirely consistent to base the power analysis on the combined approach. At no point do we make any false claims about the significance level of the combined approach, contrary to what is stated by the authors of the annex.

(At any rate, if one were to base the power analysis on approach IM-2 only rather than on the combined approach IM-3, then the power values would go down but not by much. Consequently, this point is not really of great practical relevance to begin with.)

The authors of the annex state that we claim that the overall result satisfies the 5% significance level. Such a claim would indeed be erroneous. Since we make no such claim, the

[6] Source: Sidney Herald, 8 April 2011. Similarly, Pechey et al. (2013) predicted that in Australia plain packaging would be reduced by 1.5 percentage points (median of experts' estimates) within two years after introduction. By any standard, such a reduction would result in a substantial effect in the short term.

[7] Unless the authors of the annex want to claim that Professor David Hill is not to be taken seriously.

statement of the authors of the annex, namely labeling a non-existing statement as an "error" clearly falls outside the bounds of acceptable scholarly debate and suggests that the annex is really a non-scientific, advocacy piece.

Remark 2.1 In a follow-up email to Prof. Michael Hengartner (the Rector of the University of Zurich) dated 4 February 2015, Mr. Diethelm writes:

> *Please also note the following corrigendum in the same Annex: The title of Error#4 should read:*
>
> *"Ignorance of the fact that disjunctive grouping of two tests results in a significance level lower than the significance level of the individual tests"*
>
> *"Lower" not "higher", as indeed higher alpha values correspond to lower significance levels. This should be obvious in the text describing the error (the paper copy of the letter includes the correction).*

This corrigendum begs belief, since lower alpha values correspond to lower significance levels, rather than higher alpha values. Indeed, in the context of hypothesis testing, "alpha" and "significance level" are one and the same. Consequently, the original title was correct (even though the allegation was still wrong, as explained above).

2.5 So-Called Error #5: Failure to take into account the difference between pointwise and uniform confidence intervals

The authors of the annex state that we did the best of both worlds: use uniform confidence intervals to assess plain packaging effects (which makes it less likely to find such an affect) and use pointwise confidence intervals for the power analysis (which results in larger power values).

But this is clearly not true. In both working papers, we use pointwise confidence intervals as a main analysis to assess plain packaging effects and use uniform confidence intervals as a variation 'only'. It is true that in Kaul and Wolf (2014b), both approaches yield the same result (that is, all intervals contain zero) but that is not an "error". In Kaul and Wolf (2014a), the two approaches do not yield the same result: Using pointwise confidence intervals, we find a plain packaging effect for December 2012; but not using uniform confidence intervals. Since the main analysis for the assessment of plain packaging effects is based on pointwise confidence intervals, it is entirely consistent to base the power analysis on pointwise confidence intervals as well. (It would have been reasonable to point out that, as a variation, an additional power analysis could have been carried out based on uniform confidence intervals.)

Labeling this point as an "error" is factually wrong. In addition, the statement of the anonymous authors of the annex is a misrepresentation of the facts and again goes beyond the bounds of credible academic debate.

2.6 So-Called Error #6: Invalid significance level due to confusion about one-tail vs. two- tail test

The authors of the annex state that we make a one-tailed use of a confidence interval via our rule that evidence for a plain packaging effect is established if the confidence interval lies entirely below zero. It is stated in the annex, that a 90% confidence interval leads to a 5% significance level rather than a 10% significance level. These statements are wrong. In making their statements, the authors exhibit a lack of basic statistical knowledge: They seem not aware of a Type III error of a hypothesis test. Allow us to explain.

For a given month, we consider the testing problem

$$H_0 : \mu = 0 \quad \text{vs.} \quad H_1 : \mu \neq 0 \ ,$$

where μ denotes the plain packaging effect. This effect is defined in a way such that negative values ($\mu < 0$) correspond to an intended effect (that is, a reduction in smoking prevalence) while positive values ($\mu > 0$) correspond to an unintended effect (that is, an increase in smoking prevalence). Unless one is 'blind' on one eye, one has to allow for the possibility of an unintended effect as well. (For example, it could be argued that minors might be more likely to take up smoking due to a 'rebel' effect.) In any case, an unintended effect cannot be ruled out *categorically a priori.*

Our rule is a follows. We compute a 90% confidence interval for μ. If the interval is entirely below zero, μ is established to be negative at the 10% significance level. Conversely, if the interval is entirely above zero, μ is established to be positive at the 10% significance level.

A Type III error occurs if a plain packaging effect exists (that is, $\mu \neq 0$) but the wrong sign is established by the test. This can happen in two ways. First, if the true effect is positive ($\mu > 0$) but the test establishes it as negative ($\mu < 0$). Second, if the true effect is negative ($\mu < 0$) but the test establishes it as positive ($\mu > 0$). We consider both directional mistakes equally grave. As the authors point out, the probability of a mistake in the first direction is bounded by 5%. But they fail to take into consideration the probability of a mistake in the opposite direction, which is also bounded by 5%. Since both mistakes cannot happen at the same time, the probability of a mistake in either direction is bounded by 10%. Hence, our analysis is not an "error".

Note that John W. Tukey, one of the most eminent applied statisticians of all time, argued that establishing the direction of a treatment effect is the most important task of statistical inference, which places foremost emphasis on the control of a Type III error; see Tukey (1991). The following recommendation from IARC (2008, pages 29–30) makes the same point: *"It is also important to consider the direction of effects. Some interventions might prove counter-productive."*

Here is another way to look at the alleged "error". Those researchers who are 'blind' on one eye and only allow for an intended (namely, negative) plain packaging effect *a priori* might be tempted to carry out a one-sided hypothesis test (which could be achieved by inverting a one-sided confidence interval); this would indeed make it more likely to establish an intended effect. But those researchers who are willing to look at both sides of the coin and allow for a plain packaging effect in either direction *a priori* must carry out a two-sided hypothesis test, which is what we did.[8] Perhaps the authors of the annex implicitly argue for the use of a one-sided test. But then they would have to justify that *a priori*, without a single shred of doubt, plain packaging can only have an intended effect (if any); that is, it can *only* happen that $\mu \leq 0$. Lacking divine insight, it seems impossible to make such an argument. Hence, we carry out a two-sided test.

In sum, labeling this point as an "error" is factually wrong.

2.7 So-Called Error #7: Invalid assumption of long term linearity

The authors of the annex state that our assumption of long-term linearity of the smoking prevalence trend is contradicted by actual data. In particular, they compute 95% uniform confidence intervals around (estimated) expected annual prevalence based on the regression line. Since there is one data point outside the corresponding confidence interval (namely for the year 2011), they refute the assumption of long-term linearity. In their argument, the authors of the annex seem to fail to understand the difference between a confidence interval (for an expected value) and a prediction interval (for a random variable). The fact that an observed value falls outside the confidence interval cannot serve as a basis for refuting long-term linearity.

To make an analogy, the confidence interval for the expected value of the number of heads in ten tosses of a coin is $[5, 5]$, that is, the single number five.[9] It would be unjustified to 'contradict' the number five because one tossed a coin ten times and obtained seven heads, which is outside the confidence interval.[10] The authors of the annex seem to have made a similar argument, though.

But even if the intervals in Figure 5 of the annex were prediction intervals rather than confidence intervals, the test for linearity would not be entirely correct. This is because a prediction interval only applies to 'new' data, that is, data that have not been used in the computation of the interval itself. Instead, the data of the year 2011 (where observed prevalence falls outside the interval) have been used in computing the interval itself.

Furthermore, why the aggregation to annual data? Aggregating data from monthly frequency to annual frequency leads to a loss of information and there is no (good) reason for

[8] Descriptive evidence in fact suggests that unintended plain packaging effects cannot be ruled out for certain subpopulations. There is in any case no *a priori* justification to disallow an unintended plain packaging effect.

[9] This is an extreme case of a confidence interval, since in this instance we know perfectly well what the expected value is; as a result the coverage probability is actually 100%.

[10] Note that the 95% prediction interval for the number of heads in ten tosses of a coin is $[2, 8]$.

doing so.[11]

There are many standard tests for a linear relation to be found in basic textbooks.[12] We suggest the authors of the annex use such a standard test (applied to the original monthly data instead of the artificially aggregated yearly data) in their future work.

In sum, labeling our assumption of long-term linearity as an "error" based on erroneous reasoning is factually wrong (apart from ironic).

3 Issues

We next respond to the seven issues listed in the annex.

3.1 Issue #1: Avoiding evidence by post-hoc change to the method

In both working papers, the main analysis is based on pointwise intervals. Doing so, in Kaul and Wolf (2014a), we establish a plain packaging effect for December 2012. On the other hand, no such effect could be established if uniform confidence intervals were used instead. As we clearly state, December 2012 is of special interest, since it was the first month when plain packaging was in effect (in full force). Therefore, a special focus on December 2012 (that is, the use of pointwise confidence interval) is well justified. If one takes this point of view, then one is equally justified in focusing on the year 2013 'only' for the power analysis (that is, only on 12 months instead of 13 months). Namely, allowing for an initial plain packaging effect due to a 'shock' effect, say, what is the probability of detecting any surviving, long-term effect over the following year (12 months)?

If no special focus on December 2012 is allowed (which is not our point of view), then the alternative analysis based on uniform confidence intervals fails to find any effect at all. To be fair, a power analysis based on uniform confidence intervals would yield lower power values.

3.2 Issue #2: Unnecessary technicality of the method, hiding the methodological flaws of the papers

It is true that our criterion for establishing an (intended) plain packaging effect — in terms of yes or no — is equivalent to the criterion stated in the box at the bottom of the page. However, our "algorithm" is more informative in that it gives a confidence interval for the true effect, that is, a range of plausible values for the true effect (at the 90% confidence level). More informative is better in our opinion, even if it comes at the cost of a somewhat more complex

[11] Perhaps the authors of the annex tried their method with the original monthly data first, which would have been the natural thing to do, but did not obtain the desired result (that is, a 'contradiction' of the assumption of long-term linearity).

[12] In fact, we have used some of those tests in our two working papers.

"algorithm". At any rate, this issue is clearly not about whether our "algorithm" is right or wrong.

In addition, the authors of the annex call our test *"very crude and naïve"* without providing any specific reason or explanation. Unfortunately, they fail to enlighten us how to derive a "refined" and "sophisticated" test instead.

3.3 Issue #3: Very ineffective and crude analytic method

The authors again refer to *"the crude nature"* of our test by saying *"see Issue #2 above"*. But in Issue #2, they only make this accusation without giving any kind of reason or explanation. These types of unsupported statements are beneath the standards of responsible scientists; instead, such statements suggest that the anonymous authors have embraced the role of advocates.

Next, the authors of the annex state that they have developed a simple t-test that is more powerful than the test we propose. This statement is 'substantiated' by a power plot which shows higher power for their test. Unfortunately, the authors fail to provide any details of their t-test whatsoever. Therefore, one would have to accept their plot on faith alone, since it is impossible to guess how their test is supposed to be carried out.

Last but not least, what is the outcome of the authors' t-test when applied to the two data sets analyzed in our working papers? It stands to reason that if their test had established a plain packaging effect (in the desired direction), they would have reported this finding.

3.4 Issue # 4: Non standard, ad-hoc method

With one exception, all the techniques that we use are standard and implemented in any decent statistical software. The various algorithms call for a number of individual techniques used sequentially, so it does not come as a surprise that any of such algorithms is not found in a textbook in its entirety. But nothing is non-standard in our algorithms. The point is that any such algorithm is built only from textbook methods.[13] The term *"ad-hoc"* is therefore unwarranted.

The one exception mentioned previously is the formal power analysis. Apart from the most trivial settings, any formal power analysis requires some programming on the part of the applied researcher(s). Our aim was to provide a clear and transparent description of what we did. Apparently, the authors of the annex were able to reproduce the results of our power analyses, so all is well in this regard too.

[13] To use an analogy, consider a new recipe for cooking a food dish that only requires standard techniques (such as pan-frying and baking). Since the recipe is new, it cannot be found in any previous cookbook in its entirety, but that does not mean that it cannot be replicated by a cook without any special, professional skills or machinery. This is opposed to a recipe that requires professional skills (such as many recipes of the famous restaurant *El Buli*) or professional equipment (such as an expensive *sous-vide* oven).

3.5 Issue #5: Contradiction and lack of transparency about the way the data was obtained

It is difficult to understand precisely what “issue” the authors of the annex are raising regarding the data set analyzed in our two working papers. The Roy Morgan Single Source data set is a high-quality data set that has been used by other researchers studying the effectiveness of tobacco control measures in Australia, including some of the researchers engaged in the same policy advocacy as OxyRomandie. For instance, Melanie Wakefield, a prominent advocate for plain packaging, has utilized the very same data set in previous research on smoking prevalence rates in Australia; for example, see Siapush et al. (2009) and Wakefield et al. (2008). Of all the available, empirical data sets regarding tobacco prevalence in Australia, we found the Roy Morgan Single Source data set to be the most reliable and promising. This conclusion certainly seems to be consistent with other prominent scientists in the field of tobacco research.

The Single Source data set is compiled by Roy Morgan Research, a highly-reputable market research firm in Australia that has also been engaged by the government of Australia to conduct its National Drug Strategy Household Survey. The Single Source data set is made available for purchase like any other commercial data set. We were able to obtain the raw data set from Roy Morgan for the adult population through PMI’s ongoing subscription with Roy Morgan. However, we understand that PMI has a strict policy against the collection of any data on minors. Consequently, PMI does not subscribe to Roy Morgan’s data for minors (ages 14–17). Therefore, an additional purchase of data for minors was required. We disclosed that PMI provided the funding for the research, but also made clear that PMI was never given access to the underlying data on minors.

Last but not least, the authors of the annex claim that our statement of the data being publicly available is *“simply false”*, since the data *“need to be bought”*. This is really a matter of semantics. We never said the data were publicly available *at no cost*. The point is that they are not proprietary.

3.6 Issue #6: Conflict of interest not fully declared?

Both working papers clearly state that *“Philip Morris International provided the funding for this research.”* The IPE press release dated 1 July 2014 further states that the paper *“was commissioned by Philip Morris International.”*

These disclosures are consistent with standard academic practice for disclosing interests. It is not necessary and would be highly unusual to provide details of the contracts underlying the funding. (If OxyRomandie believes that it would be useful if the working papers made this fact more explicit, then that may be a topic worthy of discussion, but clearly there is nothing in its complaint about the scope of the disclosure language that warrants retracting the two working papers.)

The participation of the IPE Institute for Policy Evaluation was likewise no secret. On page 1 of both working papers, the lead author's affiliation with IPE is fully disclosed as *"Prof. Dr. Ashok Kaul,* ***Institute for Policy Evaluation (IPE)****, Saarland, and Department of Economics, Saarland University"* (bold face added).

3.7 Issue #7: Lack of Peer Review

At no time has anyone – especially the authors – ever suggested that our two working papers were peer-reviewed. Indeed, the very nature of publication in a working paper series makes this obvious to any fair-minded, objective reader of either working paper.[14] As is customary in the field of economics, the presentation of research findings often begins with the publication of a working paper as part of a working paper series, which is then typically followed by submission for publication in a peer-reviewed journal. Notwithstanding the overly sensational and unscientifically personal nature of the OxyRomandie correspondence, publication of our research findings has clearly resulted in relevant methodological debate.

4 Conclusion

The authors of the annex have set out to discredit our research by providing a list of seven (so-called) errors and a list of seven issues. But they have clearly failed in their mission. Although there are some (minor) points of debate, there is not a single *"extremely serious"* error in our two working papers, as we have explained in detail in this reply.

Instead, the authors of the annex (i) have shown a surprising lack of basic statistical knowledge and (ii) have made false statements about the content of our working papers. What they have achieved to discredit, therefore, is only themselves. Perhaps this serves to explain the anonymous nature of the annex.

We welcome a constructive debate based on substance and comporting to scientific standards and decorum. We regret that the authors of the annex have repeatedly overstepped the limits of scientific debate (i) by misrepresenting our approaches, methods, and findings; (ii) by engaging in personal attacks; and (iii) by hiding behind the cloak of anonymity. Although we firmly believe that constructive criticism will further scientific discovery and support good policy-making, we cannot accept defamatory statements and unsubstantiated attacks both on our academic institutions and on us as individuals.

[14] Specifically, on the title page of Kaul and Wolf (2014b), it says *"Working Paper No. 149"* and on the title page of Kaul and Wolf (2014a), it says *"Working Paper No. 165"*.

References

Davidson, S. and de Silva, A. (2014). The plain truth about plain packaging: an econometric analysis of the Australian 2011 tobacco plain packaging act. *Agenda: A Journal of Policy Analysis and Reform*, 21(1):27–43.

Heckman, J., Moon, S. H., Pinto, R., Savelyev, P., and Yavitz, A. (2010). Analyzing social experiments as implemented: A reexamination of the evidence from the HighScope Perry Preschool Program. *Quantitative Economics*, 1(1):1–46.

IARC (2008). *Methods for Evaluating Tobacco Control Policies*. Handbook of Cancer Prevention Volume 12, IARC Press, Lyon.

Kaul, A. and Wolf, M. (2014a). The (possible) effect of plain packaging on smoking prevalence in Australia: A trend analysis. Working Paper ECON 165, Department of Economcis, University of Zurich.

Kaul, A. and Wolf, M. (2014b). The (possible) effect of plain packaging on the smoking prevalence of Australian minors: A trend analysis. Working Paper ECON 149, Department of Economcis, University of Zurich.

Pechey, R., Spiegelhalter, D., and Marteau, T. M. (2013). Impact of plain packaging of tobacco products on smoking in adults and children: An elicitation of international experts' estimates. *BMC Public Health*, 13:18.

Siapush, M., Wakefield, M. A., Spittal, M. J., Durkin, S. J., and Scollo, M. M. (2009). Taxation reduces social disparities in adult smoking prevalence. *American Journal of Preventive Medicine*, 36(4):285–291.

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science*, 6:100–116.

Wakefield, M. A., Durkin, S., Spittal, M., Siapush, M., Scollo, M. M., Simpson, J. A., Chapman, S., White, V., and Hill, D. (2008). Impact of tobacco control policies and mass media campaigns on monthly adult smoking prevalence. *American Journal of Public Health*, 98(8):1443–1450.