Methodology Review and Concept Assessment – Project Origin

A report prepared for:

ISBA

20th October 2020

Restricted Distribution: Circulation prohibited without the express consent of ISBA

RSMB Limited, Savoy Hill House, 7-10 Savoy Hill, London WC2R 0BU
Tel: + 44 (0)20 7808 3600 Email: contact@rsmb.co.uk Web: www.rsmb.co.uk
Registered office as above. Registered in England and Wales No. 2173860

Executive Summary

The WFA blueprint for cross-media audience measurement has undergone extensive industry briefing and peer review of its components. In particular, the Virtual ID solution is a hybrid model which leverages the power of the census to accurately and precisely measure impressions and the power of a single source panel to accurately measure cross-media duplications and campaign demographic reach and frequency.

The VID solution has two very important attributes:

- It is actionable, which means that whenever and wherever impressions are added to a campaign, there will never be a reported decrease in campaign reach.
- It is accessible, which means that the necessary access to a publisher's data and the complexity of the production system are both minimized.

The RSMB report:

- Provides a conceptual assessment of the statistical modelling which underpins the VID solution and considers important trade-offs amongst conflicting objectives.
- Identifies sources of potential loss of accuracy consequent on partial failure of underlying model assumptions.
- Considers how TV data can be integrated into the VID solution.
- Proposes a framework for a proof of concept study which will provide a transparent evaluation of model performance.

There is always more than one way to execute a data integration, often dependent upon the priority between accuracy, actionability and accessibility. RSMB have outlined some alternative approaches but note that these are likely to sacrifice actionability and/or accessibility.

Overall, RSMB recommends that the VID model provides an elegant solution with a sensible trade-off between the likely accuracy in the prediction of cross-media duplications and the

practicality of the process. In our opinion it merits a proof of concept evaluation and cannot be approved or rejected without. The impact of the following issues should be addressed:

- The underlying probability model is grounded in sound statistical theory, but the estimation of its parameters is somewhat abstract. In our opinion this provides a best fit to the single source panel training dataset, but over-fitting may produce unstable parameter estimates and compromise their subsequent application.
- Are the dimensions controlled in the model sufficient to predict systematic differences between individuals in terms of publisher and cross-publisher demographic profiles and reach and frequency? This is the standard conditional independence assumption which applies to all probabilistic data integration exercises. To an extent nuanced differences between campaigns will regress to the average and cross-media duplications will regress to random. Regression-to-the mean diagnostics quantify this effect and must be evaluated for a large number of diverse campaigns. This is essential input for consensus agreement on acceptable accuracy.
- For retention of an individual publisher's campaign frequency distribution, it is best to assign the complete set of campaign impressions for a cookie (or more persistent personal identifier) to the same virtual person (VID). This is a key principle of the allocation algorithm.
- However, for the optimum prediction of demographic profiles and cross-media duplications, it may be necessary to split cookies in order to introduce other control dimensions (such as time of day) into the model. This is a key trade-off in the VID model.
- We believe that the presence of targeting and frequency capping will challenge the assumptions of conditional independence, but this must be assessed against specific cross-media planning options.

The VID framework can be extended to embrace the BARB TAM panel output. RSMB have made recommendations on how BARB respondent level data can be configured into an impressions-led dataset, which replicates the input data required from a website and is therefore a natural input to the VID model. We note the following key issues from our report:

- The BARB panel has a very persistent personal identifier which transcends all TV channels, platforms and devices. The key trade-off between “cookie” retention and optimum prediction of cross-media duplications is more acute.
- Through sampling error, the single source panel may estimate cross-TV channel duplications which are different to BARB. Post reach and frequency analysis calibration may be required, therefore compromising accessibility.

In principle broadcaster video on demand (BVOD) census data can be available in the same format as website census data. However, for each BVOD commercial impression, there is often more than one person viewing, at a level which cannot be ignored. RSMB have recommended an extension to the VID methodology to accommodate viewer-per-view factors. We note the following key issues from our report:

- Estimation and application of viewer-per-view factors introduce elements of regression-to-the-mean and sampling error into the census-based campaign impressions data.
- In the allocation algorithm, each cookie must be linked to multiple VIDs with different demographics, but only a probabilistic selection of the linked VIDs are allocated each impression.
- This will increase regression-to-the-mean effects in all metrics.

As required for a proof of concept study, RSMB have been assiduous in attempting to identify all points of potential weakness in the VID model. However, these considerations would apply to any data integration solution, accepting that there may be different trade-offs between accuracy, actionability and accessibility. We reiterate our recommendation that the VID model presents a plausible solution which may produce an acceptable trade-off.

1. Introduction

The WFA blueprint for cross-media measurement has undergone extensive industry briefing and peer review of its components. This review identified two immediate areas for further assessment:

- The role and strength of the Virtual ID model for estimating and explaining publisher and cross-publisher reach and frequency.
- How TV data, in its current form, can be integrated within the blueprint, preserving its current outputs and without introducing systematic bias for or against the media.

Project Origin (UK) intends to conduct an independent assessment of these two areas. The end-goal purpose is enabling local build, execution and evaluation of a proof of concept test.

This RSMB report:

- Is a conceptual assessment of the statistical modelling which underpins the VID model and identifies the trade-offs between accuracy and accessibility.
- Does not address the issues of the definition or quality of impressions between media.

The WFA blueprint is a framework rather than a completely tied-down solution. Therefore, it has been important for RSMB to:

- Make a fair and valid interpretation of the VID model architects' intentions. This has involved an iterative question and answer process with Google, Facebook and ISBA, to tie down any areas of ambiguity or to correct RSMB's misunderstandings (see Appendix A).
- Make methodological choices where necessary (some relevant to the specifics of the UK market) and to suggest variations which should be tested in a proof of concept in the interests of improved understanding of the trade-offs.

The VID solution has two very important attributes:

- It is actionable, which means that whenever and wherever impressions are added to a campaign, there will usually be an increase, but never a decrease, in campaign reach. Maybe surprisingly, this is often quite difficult to achieve. However, it is a worthy goal and it may be worth sacrificing some accuracy in favour of logical results when slicing and dicing campaigns.
- It is accessible, which means that the necessary access to a publisher's data is minimized, the data extraction algorithm is easy to apply and the integrated dataset for a campaign is transparent/easy to analyse.

The VID solution is a hybrid model of census and single source panel data. We make the following general comments about performance compared to raw panel data:

- Metrics which are highly dependent on census data, such as impressions, will show improvement in both accuracy (truth) and precision (volatility).
- Metrics which are highly dependent on panel data, such as cross-media duplications, are likely to suffer a loss of accuracy (model smoothing effects) or only limited improvement in precision.

Whilst the report does consider alternatives to the VID solution, it is likely that these will sacrifice actionability and/or accessibility. Therefore, we have not treated these alternatives as a priority. Throughout our description of the methodology, we have identified where/how the model is sensitive to the underlying assumptions and performance should be evaluated and diagnosed.

It is expected that this report will be an essential input to the specification of the proof of concept test in an invitation to tender, in particular to recommend criteria and approaches to validation.

2. Conceptual Assessment – Virtual ID Model

The VID framework is an outcome of considerable development work in the measurement of reach and frequency within and across websites, where the publisher collects census-based advertising impressions data over time. In this sense, the VID model is a re-purposing of models which have a track record of high performance in the online sector. It makes sense to start by evaluating the models in this context before considering the somewhat different challenge of TV.

A fundamental principle of the VID model is that all the heavy lifting is done by one or more specialist organisations who provide a central hub for single source panel operations, model development/implementation, provision of data interrogation algorithms to individual publishers, collation of publisher output and delivery of cross-media reach and frequency. An individual publisher is required only to run a simple algorithm (maybe daily) which inspects each impression, accesses limited metadata and probabilistically assigns each impression to a virtual person.

2.1 The VID Model in Outline

Essentially, the VID solution is hybrid model which leverages the power of the census to accurately and precisely measure impressions and the power of the panel to accurately (if not so precisely) measure cross-media duplications and demographic reach and frequency. Therefore, in the first instance we deviate slightly from the well-developed WFA blueprint and describe the model in terms of reach and impressions. There is a step in the process where the single source value of cookies is leveraged – at the end of this section we explain how the model must be re-formulated to correctly accommodate this procedure.

The first phase is operated by the Central Hub:

- The basic premise is that the population can be split into a (large) number of small groups which are systematically different to each other in terms of their exposure to advertising impressions, within and across individual media channels. Crucially, within each group, there are no systematic variations between individuals in terms

of exposure whatever the mix of impressions, for example by media channel, time of day, media content or any other planning criteria.

- Each group will have a population size and a relative rate of exposure to each **type** of impression. In the first instance we assume that a **type** is defined by a media channel and a device, for example a website accessed through a tablet. The rates of exposure are assumed to be the same for each person in the group.
- An underlying probability model has been chosen to describe the relationship between campaign reach and numbers of impressions by **type.** The functional form of this model is the same for each group, but the relative rates of exposure will be different between groups.
- A single source panel is used to calculate observed reach and numbers of impressions by **type**, for a large number of campaigns. These metrics inform an optimisation algorithm which determines the model parameters: these are the number of groups required plus the population size and relative rates of exposure by **type** for each group. The idea is that one set of parameters will work for any campaign. The key optimization objective is to minimise the differences between modelled and observed campaign reach.
- The last step is to create a population of virtual people (VIDs). Each group has a number of VIDs equal to its population. Each VID in a group has a unique identifier and a contact rate for each **type** of impression.

The second phase is operated by each publisher, for a particular live campaign:

- The publisher has access to the VID dataset and an impression allocation algorithm developed, tailored and provided by the Central Hub.
- The algorithm operates on each **type** (channel/device) separately.
- As each impression is registered, it is allocated to a single VID, at random with probability proportional to the relative rate of exposure for that type.
- Unless the impression has a cookie or other unique identifier which has already been allocated to a particular VID, in which case the new impression is allocated to that same VID. This is designed to mitigate unnecessary assumptions of randomness. We refer to it as the cookie hash procedure.
- At any point in the campaign, the publisher can return to the Central Hub the VID dataset, updated with a list of impressions allocated to each VID, together with any

privacy allowed metadata for campaign slicing and dicing (e.g. time of day, media content).

The third phase is operated by the Central Hub:

- Collate the updated VID impressions data across all publishers and count the cross-media reach and frequency analysis.

There is a demographic overlay:

- The first phase is operated within each of a set of mutually exclusive and exhaustive demographics cells.
- Each VID also has a demographic classification.
- In the second phase, demographic information in an impression's metadata is used to restrict allocation to those VIDs with a matching demographic.

And a further demographic modelling requirement to correct errors in a publisher's demographic labels or to make an informed demographic assignment when the publisher has no demographic labels, demographic labels for only a subset of impressions or there is a different demographic granularity:

- For a large set of campaigns, each publisher has to count the number of impressions in each demographic group defined in terms of their demographic labels. This includes the number of impressions without a demographic label.
- For the same campaigns, the Central Hub will calculate from the single source panel an equivalent profile defined in terms of panel demographics.
- The Central Hub then calculates how many impressions need to move from one demographic cell to another to another to match the panel target profile.
- This matrix of changes is then converted to a set of transition probabilities.
- The allocation algorithm identifies the demographic label in an impression's metadata then allocates the impression to the target demographic cell with probability proportional to the transition probability.
- Unless the impression has a cookie or other unique identifier which has already been allocated to a particular demographic, in which case the new impression is allocated to that same demographic cell.

- The allocation to VID phase happens before moving on to the next impression.

In practice, this interpretation of the process is not quite correct. Whilst we believe that it is important to consider a model based upon the relationship between reach and numbers of campaign impressions, this does not support the application of a cookie hash procedure in the allocation and demographic correction/assignment algorithms. To correctly accommodate the cookie hash procedure, the model needs to be re-based on the relationship between reach and numbers of campaign unique cookies. A campaign unique cookie is a cookie with at least one campaign impression. We recommend that both variations of the model are evaluated in the proof of concept, to help demonstrate important trade-offs in estimating cross-media duplications.

2.2 The Basic VID Premise

The basic premise is that the population can be split into a (large) number of small groups which are systematically different to each other in terms of their exposure to advertising impressions, within and across individual media channels. Crucially, within each group, there are no systematic variations between individuals in terms of exposure whatever the mix of impressions, for example by media channel, time of day, media content or any other planning criteria.

This is the conditional independence assumption required for all data fusion methodologies. In the VID model, it means that within a homogeneous group, a person's probability of exposure to a particular impression is statistically independent of their exposure to any other impressions in a campaign. This is irrespective of any characteristic which classifies the type of impression (e.g. media channel, device type, time of day, media content). It also means that each person in the group has an equal chance of being the one exposed to that impression.

In the VID model, determination of the model groups is based upon all cookie types (media channel/device) and the overall propensity to view each cookie type. This means that for natural laydown campaigns, the groups are expected to be homogeneous in terms of media duplications. However, they may not be homogeneous for campaigns which vary from a natural laydown, for example with a deliberate skew by time of day or media content. It may be an issue for targeted campaigns.

All data fusions have a trade-off between the granularity in the definition of homogeneity and limitations of sample size in the single source panels which provide the data source to find the homogeneous groups. It is not a disaster if the dimensions not considered are second order to those that are controlled, but it may put a marker down for a limitation on the application of the final solution, particularly for campaign designs which are outside the domain of study. The other consideration is that there is no point increasing the dimensions of the definition of homogeneity if the extra criteria are not actionable in the overall model. For example, if these criteria are not available in the publisher metadata or their application frustrates or creates a trade-off against another feature of the overall model (for example allocation of whole cookies to VIDs); this is considered in section 2.11 and 2.12.

Homogeneity can be explored using an analysis of variance technique (e.g. CHAID) applied to the single source panel. We see this as either:

- An exploratory tool to identify the most effective criteria prior to the model fitting exercise.
- A diagnostic tool to potentially explain a visible loss of accuracy in cross-media campaign reach and frequency results.

In either case we recommend that the proof of concept is first run on the construct described due to the complicating impact on other model components.

2.3 Types of Impression

*Each group will have a population size and a relative rate of exposure to each **type** of impression. In the first instance we assume that a **type** is defined by a media channel and a device, for example a website accessed through a tablet. The rates of exposure are assumed to be the same for each person in the group.*

Relative rates of exposure are best explained by example:

	Population Proportion	Relative Rate of Exposure	
		Type 1	Type 2
Group 1	0.5	0.4	1.0
Group 2	0.5	1.6	1.0
Total	1.0	1.0	1.0

The population proportions and relative rates of exposure are the parameters of both the campaign reach model and the campaign impression allocation algorithm.

The probability of a person in Group 1 viewing a Type 1 impression is 0.5 x 0.4 = 0.2.

If we have a population of 100 people and a campaign with 100 Type 1 impressions and 50 Type 2 impressions, then we can apply the model parameters to calculate the expected number of impressions for each person:

	Population	Impressions		Impressions per Person		
		Type 1	Type 2	Type 1	Type 2	Total
Group 1	50	20	25	0.4	0.5	0.9
Group 2	50	80	25	1.6	0.5	2.1
Total	100	100	50			

On average, each person in Group 1 is expected to view 0.9 campaign impressions.

2.4 Underlying Probability Model

*An underlying probability model has been chosen to describe the relationship between numbers of impressions by **type** and campaign reach. The functional form of this model is the same for each group but the relative rates of exposure will be different between groups.*

For each person in a group, it is assumed that exposures to a campaign will follow a Poisson process with a rate parameter equal to the average number of impressions per person. The Poisson process is widely used as a model of campaign reach and frequency and we endorse this application. As long as there is no systematic variation in a person's rates of exposure, then the Poisson will accurately predict the probability of that person having 0,1,2,3,etc. exposures to a campaign. It fits with the basic premise of conditional independence. Of course, in practice that premise is never perfectly achieved; subsequent steps in the VID process seek to mitigate such "failure" and it will be important to quantify the effect.

The Poisson model generates a simple formula for prediction of campaign reach given the number of impressions per person:

Personal reach probability = $1 - \exp(-x)$

Where x = number of impressions per person.

These personal reach probabilities are summed across the whole population to generate the prediction of campaign reach.

In our example, 50 people have an x of 0.9 and 50 have an x of 2.1.

Campaign reach = $50 \times (1-\exp(-0.9)) + 50 \times (1-\exp(-2.1)) = 73.5$

If we hadn't bothered to split the population into homogeneous groups, then the model would have delivered an overall reach of 77.7.

Note that in theory, this Poisson process is theoretically equivalent to allocating impressions to people, one by one, at random, with probability proportional to the relative rates of exposure. This is important for tying the Poisson model to the allocation algorithm.

2.5 Estimation of Model Parameters

*A single source panel is used to calculate observed numbers of impressions by **type** and reach for a large number of campaigns. These metrics inform an optimisation algorithm which determines the model parameters: these are the number of groups required plus the population size and contact rates by **type** for each group. The idea is that one set of parameters will work for any campaign. The key optimization objective is to minimise the differences between modelled and observed campaign reach.*

RSMB's approach to generating a set of population groups, with maximum between group variance and minimum within group variance, would be via multivariate analysis of variance, maybe involving CHAID and factor analysis. In this case, the input would be the observed exposure rate, for each channel/device, for a large number of impressions (accumulated over a large number of campaigns), for each single source panel member. Having determined the segmentation, we would then calculate the population proportions and the average relative contact rates for each segment (group). The advantage of this approach is that when these parameter estimates are the input to a Poisson model, the modelled frequency distribution will sum to the right number of impressions for a campaign. Further, the transparent definition of these parameters fits with the theoretical link which justifies their use as selection probabilities in the allocation algorithm; the parameter estimates are robust and will not take extreme values when the observed data does not match the model assumptions (for example if the reach to cookie relationship is linear). However, it will not be optimum for estimating the campaign reach observed in the panel. For this reason, our products sometimes need a post-analysis calibration routine to match input data trading currencies.

The VID model approach is to treat the "population proportions" and "relative contact rates" as theoretical parameters which do not have values which can be observed in the SSP data. Instead they take whatever values are necessary to match the observed panel total impressions by type and overall cross-media reach, for a large number of campaigns. This training dataset can be maximized by altering the duration and selection of types within each campaign, i.e. slicing and dicing. We understand that iterative, least squares algorithms exist to provide a solution.

We are generally comfortable with abstract models which don't have a complete theoretical justification, but we do have a residual concern that there is not a unique set of parameters,

rather a large number of solutions which all work equally well but deliver a diversity of population groups/contact rate parameters. The concern is that the model fitting may not be consistent with the application of its parameters in the micro-level allocation algorithm. Maybe this is equivalent to over-fitting, in which case a diversity of campaign results for model training could be the solution.

A related concern is that when the underlying probability model is not perfect (or input data is volatile) then theoretical parameters are not well-behaved. This may be the root cause of the special case in the allocation algorithm, in which some impressions are not allocated to a VID. In the proof of concept, it may be necessary to develop a bespoke model fitting algorithm, possibly with constraints to produce logical parameters, rather than to rely on an off-the-shelf product.

Whichever approach to parameter estimation is followed, we will very quickly run out of single source panel sample size – there are potentially a lot of Dirac groups and certainly a lot of types. It will be necessary to aggregate the input data before or during model fitting by forcing similar (smaller) websites to have the same relative rates of exposure. This will effectively reduce the number of parameters.

In this step, we are trying to strike the right balance between:

- Tying the model to the single source panel and suffering the consequent sample size problems.
- Looking for a relatively small set of systematic factors, which are robust yet explain all systematic variations in behaviour.
- Logically convert to model parameters.

We recommend an alternative method for consideration, analogous to a methodology which is a feature of some of RSMB's reach and frequency products. For each individual single source panel member, it is easy to calculate their personal relative rate of exposure for every media type, based upon their accumulated impressions for the training campaigns. Each individual panel member then defines a homogeneous sub-group of the population and their weight defines this sub-group's proportion of the population.

This is a very simple way to define the groups and calculate the model parameters. It is unashamedly tied to the panel and directly retains its sampling error. However, it is based on an accurate measurement of normative cross-media duplications for a set of real people, with no regression-to-the-mean. If consequent analysis applications are granular, then this sampling error "lumpiness" may show through. However, it will be informative as a convenient and extreme alternative to the abstract model.

We recommend:

- That the proof of concept is run on the construct described.
- Repeated on the one group per panel member basis.
- That the analysis of variance approach is reserved as a diagnostic tool to explain potential loss of accuracy in the overall VID solution reach and frequency.

2.6 Virtual People - VIDs

*The last step is to create a population of virtual people (VIDs). Each group has a number of VIDs equal to its population. Each VID in a group has a unique identifier and a contact rate for each **type** of impression.*

There are no statistical issues to consider in this step. It is worth noting that the information attached to the VID will be increased to facilitate other steps in the allocation process (e.g. demographics) and/or develop the model (e.g. to control other dimensions).

2.7 Individual Publisher Algorithm

The publisher has access to the VID dataset and an impression allocation algorithm developed, tailored and provided by the Central Hub.

The model parameters (group population proportions and relative contact rates) drive an algorithm which operates, in private, on a publisher's own impressions delivery database. The algorithm is designed and written by the Central Hub. To an extent there will be an onus for the publisher to create a raw output from the database which has a format which is consistent with the requirements of the algorithm. The Central Hub needs to tailor the detail of the algorithm to accommodate specific variations in the publisher's metadata, for example, inconsistent demographic classifications or registration identifiers rather than cookies. The algorithm has to be operated by the publisher to avoid the Central Hub having direct access to raw data in the publisher's database.

2.8 Operated by Type

*The algorithm operates on each **type** (channel/device) separately.*

Each publisher may have several channels/websites operated on several devices. However, in our interpretation of the VID model, each channel/website/device combination is treated as a separate silo. The allocation algorithm operates on each type separately even though cookies or unique identifiers may be common across types. This is necessary to respect the fact that different types have different model parameters (rates of exposure) – this means that if two impressions have the same cookie, they are likely to be allocated to different VIDs if they are different types.

Sub-dividing cookies in this way results in loss of single source connections in cookie records but this is a trade-off against appropriate allocation of impressions by channel/device type. Hopefully, conditional independence will mean that the synthetic allocation of sub-divided cookies to VIDs will create a dataset which, in aggregate, is representative of the original duplications between types.

Again, we recommend that the proof of concept is first run on the construct described, but recognizing that losing cookie integrity may result in a loss of accuracy in cross-media duplication estimates, if the basic model assumptions of conditional independence do not hold. This may be a particular issue for TV where the BARB panel "cookie" transcends all measured channels and devices.

2.9 Impression Allocation

As each impression is registered, it is allocated to a single VID, at random with probability proportional to the relative rate of exposure for that type.

Returning to our example, every VID has the following parameters attached to their data record, depending upon which group they were allocated to:

	Population Proportion	Relative Rate of Exposure	
		Type 1	Type 2
Group 1	0.5	0.4	1.0
Group 2	0.5	1.6	1.0
Total	1.0	1.0	1.0

The population proportions and relative rates of exposure are the parameters of both the campaign reach model and the campaign impression allocation algorithm. In practice there will be many more groups with a spread of exposure rates.

If a publisher owns only the Type 1 impressions, then they don't need to consider or even see the Type 2 relative rates of exposure. The publisher then runs their version of the allocation algorithm on their Type 1 impressions, for the nominated campaign.

The probability of a person in Group 1 viewing a Type 1 impression is 0.5 x 0.4 = 0.2.

The probability of a particular VID in Group 1 viewing a Type 1 impression is 0.2 divided by the number of VIDs in Group 1. Then Type 1 impressions are allocated one by one to VIDs, independently of previous allocations, at random and in proportion to this probability.

This is very close to the underlying process on which the Poisson model is derived, so we have theoretical consistency between the panel trained reach model and the allocation algorithm.

Another publisher will run their version of the allocation algorithm for the campaign's Type 2 impressions. This allocation will be independent of the Type 1 allocation of impressions to the common set of VIDs.

If the independence assumptions are met perfectly, then this will form an accurate estimate of the population campaign reach and frequency distribution for:

- Type 1 impressions
- Type 2 impressions
- Type 1 and Type 2 combined

2.10 Cookie Hash Procedure

Unless the impression has a cookie or other unique identifier which has already been allocated to a particular VID, in which case the new impression is allocated to that same VID. This is designed to mitigate unnecessary assumptions of randomness.

In practice, the conditional independence assumptions will not be perfectly met. This will compromise the campaign reach and frequency relationship within each type. In this situation there will be an advantage if the single source attributes of the cookie can be retained. Hence the above modification.

As discussed at the end of section 2.1, this cookie hash procedure cannot be inserted without changing the configuration of the underlying probability model and the resulting allocation parameters. The architects of the blueprint have confirmed that this reflects their design. The fundamental issue is that:

- The probability of a person viewing a particular impression (the rate of exposure parameter) is different to the probability of a person viewing at least one impression in the campaign (the adjusted allocation unit).

- Heavier viewers are likely to be allocated too many impressions and this will drive down reach and distort demographic profiles.

Essentially, we need to change the model from:

Reach = f (campaign impressions)

To:

Reach = f (campaign unique cookies).

We believe that this is the basis of the tried and trusted cookie deletion models. The Poisson model is again a sensible basis for the reach curve, although the model assumptions have more chance of failing for campaigns (or parts of campaigns) where the number of unique cookies per campaign person viewing is close to 1. It may be a bit more difficult to generate robust parameter estimates than for the impressions-based model.

Of course, this assumes that cookies are available in the single source panel data and that cookie deletion within a panel is representative of cookie deletion in the population.

We acknowledge that the model must also embrace people using multiple devices/browsers in the same type and multiple people sharing the same devices/browsers. As long as all this information is picked up and reported in the single source panel, then the model should naturally still work. We guess that there will be an issue if the average number of campaign unique cookies per campaign viewer is less than 1.

We recommend that the proof of concept is run:

- With the impressions-based model and no cookie hash routine.
- With the impressions-based model and the cookie hash routine (illogical but informative).
- With the campaign unique cookie-based model and the cookie hash routine.

2.11 Publisher Output: VID Impressions by Type

At any point in the campaign, the publisher can return to the Central Hub the VID dataset, updated with a list of impressions allocated to each VID, together with any privacy allowed metadata for campaign slicing and dicing (e.g. time of day, media content).

Obviously, this step is straightforward. But at this point it is appropriate to note the potential loss in accuracy of reach and frequency, for the individual types (i.e. not yet considering cross-media), from a failure in the conditional independence assumptions; regression-to-the-mean. The allocation process is only controlling for people's overall propensity to view impressions (of that type) so is likely to be accurate for an average (maybe natural laydown) campaign, but may not pick up variations in the reach and frequency relationship for campaigns which are skewed, by design, in terms of other planning criteria (e.g. time of day, media content, targeting).

The protype study needs to evaluate and quantify this regression-to-the mean. We recommend that this can be achieved by ensuring that:

- A representative set of diverse campaigns are used for model fitting and RTM evaluation.
- Planning criteria like time of day are included in the impression metadata.
- Test campaigns are sliced and diced by these criteria.
- Planning criteria are available in the single source panel data so that benchmark reach and frequency results can be calculated.

We recognize that a macro calibration model will be needed to adjust panel reach and frequency to website actual/census impressions, to create a valid benchmark. Hopefully, this required calibration is small, otherwise we just end up arguing the merits of two alternative model constructs, rather than testing the VID model against a source of truth!

2.12 Cross-Media Reach and Frequency

Collate the updated VID impressions data across all publishers and count the cross-media reach and frequency analysis.

All publishers are mapping impressions to a common set of VIDs. This means that it is straightforward for the Central Hub to count impressions across all publishers/channels/websites/devices.

At this point it is appropriate to note another potential loss in accuracy, this time in terms of the duplications between types, again from potential failure in the conditional independence assumptions. The VID groups are homogeneous in terms of propensity to view permutations of impression types, but only at the highest level. The issue is that a group that is heavy viewing for both Type 1 and Type2, may actually be heavy/light for impressions served in the morning. Then for a campaign that is only delivered in the morning, all-time allocation probabilities will be appropriate for Type 1 but not for Type2 who will be allocated too many impressions. Without analysing the single source panel, we don't know if heavy viewers are heavy viewers or not at all times of the day.

Once again, we recommend that the proof of concept is first run on the existing construct. Repeat the regression-to-the-mean evaluation recommended in 2.11, but this time in cross-media mode.

If there is damaging evidence of RTM (we are not assuming that this will be the case) then it may be necessary to introduce other criteria into the definitions of types. This is a follow-up exercise. In order to avoid a proliferation of types, we would recommend a multivariate analysis of variance in the single source panel to identify a parsimonious set of criteria for sub-division of types.

We recognize that the model needs to be adjusted to accommodate these extra dimensions:

- Single source panel sizes will constrain the number of extra dimensions that can be accommodated in the model fitting exercise. Types may need to be grouped, i.e. given the same relative exposure rates, because time of day may be a more important discriminator.

- In the cookie hash procedure (section 2.10), a cookie may need to be sub-divided so that it's morning impressions can be allocated to one VID but other day-part impressions can be allocated to a different VID.
- Or we just accept that the dimension of the first impression determines the "campaign permanent" link to a VID. Probabilistically this is OK but may be volatile.

2.13 Demographic Overlay

The first phase is actually operated within each of a set of mutually exclusive and exhaustive demographics cells.

The first phase of model fitting and parameter estimation is done within each demographic cell. As before, this is entirely a single source panel-based analysis with the demographic classifications and populations determined by the panel (assuming the panel is controlled to census in the usual way).

Each VID also has a demographic classification.

The list of VIDs is now partitioned by homogeneous group within demographic cell. The demographic classification is added to the VID record.

In the second phase, demographic information in an impression's metadata is used to restrict allocation to those VIDs with a matching demographic.

As well as providing important planning/analysis functionality, this stratification by demographics improves conditional independence and the accuracy of cross-media duplications (but not to the extent that all regression-to-the-mean will go away!). So far, our description assumes that:

- Every publisher has demographic classifications in the metadata.
- The declared demographics are correct.
- They are complete for every cookie.

2.14 Demographic Correction/Assignment

For a number of large campaigns, each publisher has to count the number of impressions in each demographic group defined in terms of their demographic labels. This includes the number of impressions without a demographic label.
For the same campaigns, the Central Hub will calculate from the single source panel an equivalent profile defined in terms of panel demographics.

These data are required separately for each type that the publisher operates. We might expect the proportion of impressions with demographic labels and the correctness of the demographic labels to vary by campaign. Therefore, it is again important to have a diversity of campaigns for model training. In practice we would use percentage profiles because we expect sampling error in the panel to deliver a random difference in absolute numbers of impressions. Here's a simple example for one campaign and one type:

Campaign Impressions Profile		
Age Group	Publisher	Panel
16-24	20	25
25-34	30	38
35+	10	37
Missing	40	0
Total	100	100

Missing demographics are a feature of publisher metadata, where no effort is made to collect or allocate demographics to impressions, or users do not comply. Single source panel demographics will be complete.

2.15 Creating a Transition Matrix

The Central Hub then calculates how many impressions need to move from one demographic cell to another to another to match the panel target profile.

For this campaign/type example, we can calculate how many publisher impressions need to move from one demographic cell to another to match the panel profile truth:

Publisher		Assigned To		
Age Group	Impressions	16-24	25-34	35+
16-24	20	15	3	2
25-34	30	0	25	5
35+	10	0	0	10
Missing	40	10	10	20
Total	100	25	38	37

If we assign the impressions to “panel quality” demographics as in the table, then we match the panel observed profile for this campaign/type. There is no justification for the transitions chosen in this hypothetical example, they just deliver the right number. Many other permutations would work and in practice a model fitting algorithm would search for the smallest number of changes.

2.16 Transition Probabilities

This matrix of changes is then converted to a set of transition probabilities.

Following our example, the assignments are simply converted to row proportions:

Publisher		Assigned To		
Age Group	Impressions	16-24	25-34	35+
16-24	1.00	0.75	0.15	0.10
25-34	1.00	0.00	0.83	0.17
35+	1.00	0.00	0.00	1.00
Missing	1.00	0.25	0.25	0.50
Total				

The objective is to find a single set of transition probabilities for each type (channel/device) that successfully matches the panel profile for all the campaigns used in the model training/fitting exercise.

We are concerned that there may be a number of quite different solutions that all work equally well but we do not have a coherent argument to justify this concern – it may well be that the demands of a diverse set of campaigns lead to a unique solution. We have considered the possibility of calculating transition probabilities directly from the single source panel data, but this depends upon having access to the demographic cookie labels for every panel member. We also run the risk of the panel not being representative of the number of people who withhold or falsify their demographic labels.

There is an obvious source of potential regression-to-the-mean in this process, because some cookies/impressions will switch demographic even though the publisher labels are correct. The extent will be evident in the diagnostics from the model fitting exercise as a failure to predict each campaign's demographic profile. It may be necessary to consider increasing the granularity of demographic correction to embrace other planning criteria, but this collides with the allocation procedures, as discussed in section 2.12.

Our inclination is to press ahead with the proof of concept in its current construct and consider the impact on the overall solution's performance as part of the diagnostic evaluation - because there may be too many other model components affected by a change to this component.

2.17 Transition Probabilities in the Allocation Algorithm

The allocation algorithm identifies the demographic label in an impression's metadata then allocates the impression to the target demographic cell with probability proportional to the transition probability.
Unless the impression has a cookie or other unique identifier which has already been allocated to a particular demographic, in which case the new impression is allocated to that same demographic cell (the cookie hash procedure).
Next, the allocation to VID phase happens before moving on to the next impression.

The key point here is that an impression/cookie is assigned to a demographic cell before it is assigned to a VID.

Following the argument in section 2.10, the underlying model must be re-based to campaign unique cookies instead of impressions in order to accommodate the cookie hash procedure.

The same permutations must be tested in the proof of concept.

3. TV

The priority is to consider how the BARB panel audience measurement can fit into the VID model. This means configuring the respondent level data to look like the equivalent output from a website.

We can think of Broadcast TV in the same way as online video in the sense that we have:

- A collection of channels which are equivalent to websites.
- A number of platforms/devices which are equivalent to online devices.
- Each person receives a time series of impressions in a campaign.

For viewing measured by the BARB panel, a key consideration is that there is a unique "cookie" (the panel id) which transcends all channels, platforms and devices. A simplistic interpretation of the VID methodology would be to:

- Retain the single source advantage of the panel id within each permutation of channel/device.
- Ignore the connection between permutations of channel/device.

This is an allocation algorithm trade-off in the relative accuracy of duplications across TV channels versus duplications across TV and websites. In a sense this is a move away from the WFA's ideal solution which is identification of the same people across all websites and TV channels, but we're not there yet and this may be the right trade-off at present. However, in this report and within the VID construct, we will consider the extreme alternatives of:

- Ignoring the panel id connection between impressions completely (i.e. no cookie hash)
- Retaining the single source connection of the panel id completely (i.e. across all permutations of channel/device).

Other considerations are:

- The BARB measurement universe does not include all broadcaster delivered impressions.

- The BARB panel is only a sample and needs to be scaled to the VID population.
- Panel turnover and non-reporting, which means we have cessation or a break in a person's measurement rather than a continuous measurement with changing cookie ids.
- Guest viewing in panel member households.
- Theoretical consistency with BARB calculation conventions (gold standard).
- Calibration.

3.1 The TV-VID Model in Outline

Again, we'll start with a simplistic description of RSMB's conversion of the VID model to TV.

With access to the published BARB data (Database 1 for respondent level demographics and viewing data, Database 2 for commercial spot metadata [transmission logs] and audiences [impressions]), all phases can be operated by the Central Hub.

In the first phase:

- The underlying probability model is extended from the website only version to embrace all TV channels/devices and platforms which are reported by BARB.
- Using the same single source panel (not the BARB panel), we calculate the observed number of impressions by **type** (website/TV channel/device/platform) and the overall cross-media reach for the same, large number of campaigns. The same optimization algorithm is used to determine an extended set of model parameters: again, these are the number of groups required plus the population size and contact rates by **type** (now extended to TV types) for each group.
- In the creation of a population of virtual people, each group has a number of VIDs equal to its population. Each VID has a unique identifier and a contact rate for each (extended) **type** of impression.

In the second phase for TV:

- The BARB panel data is converted to an impression led dataset, to be consistent with expected website output. There is one dataset for each **type.**
- Impressions are listed in chronological order according to time of viewing. An impression is defined to be one panel member viewing one spot in the campaign. If a panel member has a weight of (say) 5000 then that impression will be listed 5000 times. The panel id will be extended to define the replicates as different individuals.
- Guest viewer impressions will be included in the list as separate events.
- "Metadata" for each impression will include panel member id, demographics, information for campaign slicing and dicing.
- As for website impressions, as each impression is inspected, it is allocated to a single VID, at random with probability proportional to the relative rate of exposure for that **type.**
- Unless it has a panel/replicate id which has already been allocated to a particular VID, in which case the new impression is allocated to that same VID.

In the third phase:

- Collate the VID impressions data across all **types**, now including websites and TV channels, and count the cross-media reach and frequency analysis.

The demographic overlay:

- Operates in the same way as for website **types**.

The demographic correction/assignment procedure:

- Is required to convert limited guest demographics to a different demographic granularity.

In principle, BARB panel data fits neatly into the VID process. In practice, single source attributes of the panel may be compromised in order to optimise modelling of cross-media duplications.

3.2 Extending the Underlying Probability Model

The underlying probability model is extended from the website only version to embrace all TV channels/devices and platforms which are reported by BARB.

The same probability model is assumed to apply across websites and TV channel viewing. This is a reasonable assumption for model formulation. Indeed, the Poisson model for exposures within homogeneous groups is also a fundamental component of the Negative Binomial Distribution, which is used in BARB's reach and frequency calibration routines. The operational difference is that in the BARB calculation procedures the model is used as an on-the-fly, post reach and frequency analysis, macro model, calibration routine; whereas in the VID model it is used in its deconstructed personal probability mode to inform the allocation algorithm which creates the reach and frequency analysis. Theoretically the VID model is a generalized version of the NBD, with many parameters designed to explain a multiplicity of exposure differences between homogeneous groups, necessary and appropriate for the task at hand. Hold the thought that a version of the generalized model may form a sophisticated, on-the-fly, calibration routine designed to force consistency with BARB gold standard reach and frequency results.

All the comments in sections 2.2, 2.3 and 2.4 apply equally to TV.

3.3 Estimation of Model Parameters

*Using the same single source panel (not the BARB panel), we calculate the observed number of impressions by **type** (website/TV channel/device/platform) and the overall cross-media reach for the same, large number of campaigns. The same optimization algorithm is used to determine an extended set of model parameters: again, these are the number of groups required plus the population size and contact rates by **type** (now extended to TV types) for each group.*

We reiterate the comments in section 2.4 but have no further issues related specifically to TV.

3.4 Virtual People – VIDs

In the creation of a population of virtual people, each group has a number of VIDs equal to its population. Each VID has a unique identifier and a contact rate for each (extended) ***type*** *of impression.*

We reiterate the comments in section 2.4 but have no further issues related specifically to TV.

3.5 Individual TV Channel Algorithm

The BARB panel data is converted to an impression led dataset, to be consistent with expected website output. There is one dataset for each ***type****.*

The individual publisher algorithm used for websites requires relatively minor modification, but otherwise the same principles are used to allocate BARB panel impressions to VIDs. However, the BARB panel data has to be summarized and reconfigured to generate an appropriate input dataset.

As for websites, the algorithm is operated separately for each TV type (channel/device/platform). This is discussed in section 2.8 for websites. The same comments apply and are more significant because we are sub-dividing the panel id (the "perfect cookie") into many types in order to ensure the optimum allocation of impressions to types. So, each VID is likely to comprise impressions from many different BARB panel members. As for websites, this is the considerable challenge for the assumptions of conditional independence. However, we stick to our recommendation that the proof of concept is first run on this basis.

3.6 Conversion of BARB Panel Data

Impressions are listed in chronological order according to time of viewing. An impression is defined to be one panel member viewing one spot in the campaign. If a panel member has a weight of (say) 5000 then that impression will be listed 5000 times. The panel id will be extended to define the replicates as different individuals.

BARB panel respondent level data is published daily. The reporting sample varies from day to day and is weighted to meet target population profiles. Hopefully the target population profiles match those of the single source panel, but this is not a major concern for the proof of concept.

The data is reported on a household basis, in terms of chronological viewing sessions, channel/device/platform classifications, persons viewing and their demographics. The separate broadcast transmission logs are used to attach commercial events to panel member viewing sessions. Then for each commercial event, we can create a list of event viewing panel members – these are the BARB panel commercial impressions.

The BARB panel is only a representative sample of all impressions. Therefore, each impression is replicated a number of times according to the panel weight. The sampling fraction of the BARB panel is roughly 1:5000, so we recognize that there will be a large number of replicates.

Each replicate has to be assigned a unique "cookie", otherwise all replicates would be allocated to the same VID! We recognize that a panel member's weight may change from day to day, therefore the number of replicates will change from day to day. So, if the weight goes up, it will look like there are new "cookies", even if the impression is for a panel member who has already seen an impression in the campaign. This introduces a bit of noise into the process, but it is no worse than real cookie deletion/replacement.

3.7 Guest Viewing

Guest viewer impressions will be included in the list as separate events.

For each BARB panel viewing session we know how many guests are present in a number of demographic groups. The list of impressions includes a separate record for each guest impression.

BARB consider these guest viewing records to be a surrogate measurement of panel members' viewing in other homes represented by the panel. There is no reach type connection between a guest viewing impression and any panel members' impressions in the same home. Therefore, each guest impression will have a unique "cookie" which can't be related to any

other cookie. Note that in some countries' TAM operations, guest viewing is already fused to panel members' viewing records.

We believe the consequent allocation routines for guest viewing will reflect the principles of the BARB macro model NBD routine for adding guest impressions to a reach and frequency analysis. Both are based upon the same probability theory. However, we recommend that the proof of concept is run with and without guests to isolate any regression-to-the-mean differences.

3.8 Impression "Metadata"

"Metadata" for each impression will include panel member id, demographics, information for campaign slicing and dicing.

The BARB panel "metadata" for each impression will include all the information necessary for interaction with the allocation algorithm or for subsequent slicing and dicing of reach and frequency analyses by other planning criteria (time of day, programme content). The panel member id is the extended version which identifies separate panel member replicates.

3.9 Impression Allocation

As for website impressions, as each impression is inspected, it is allocated to a single VID, at random with probability proportional to the relative rate of exposure for that ***type****.*

This exactly follows the procedure for websites described in section 2.8 and transparently treats guest impressions in the same way as panel member impressions. For the procedure to generate an accurate estimate of the population campaign reach and frequency, there is an additional, implicit assumption: within a homogeneous group, a person's probability of exposure as a guest impression is statistically independent of their exposure to any other at-home or guest impressions. And one final assumption: it is assumed that a person's relative exposure rates by type are the same for at-home viewing and guest viewing – this is similar to the underlying assumptions in the BARB reach and frequency calculation conventions.

3.10 Cookie Hash Procedure

Unless it has a panel/replicate id which has already been allocated to a particular VID, in which case the new impression is allocated to that same VID.

In principle this follows the procedure described for websites in section 2.10 and the same comments/reservations apply.

In addition, we recommend some adjustments to the procedure which benefit from panel management information. If a VID has already been allocated a BARB panel "cookie" which is still live (we know if a panel member has left the panel), then they cannot be allocated an additional cookie. If the "cookie" has ceased to be live, then it is entirely appropriate to give that panel member a chance of receiving an additional "cookie".
Again, this cookie hash procedure cannot be inserted without changing the configuration of the underlying probability model and the resulting allocation parameters.

This time, we need to change the model from:

Reach = f (campaign impressions)

To:

Reach = f (campaign unique panel ids)

The Poisson model is again a sensible basis for the reach curve, although the model assumptions have more chance of failing for campaigns (or parts of campaigns) where the number of unique panel ids per campaign person viewing is close to 1. It may be a bit more difficult to generate robust parameter estimates than for the impressions-based model. Within a campaign, panel non-reporting and turnover rates are low, therefore this is likely to be more of an issue for single type TV campaigns than for single type websites. Again, we need to consider the relative merits of empirical rather than theoretical estimation of parameters.

3.11 Cross-Media Reach and Frequency

*Collate the VID impressions data across all **types**, now including websites and TV channels, and count the cross-media reach and frequency analysis. In theory this procedure has an underlying consistency with BARB conventions.*

All the issues and recommendations discussed in section 2.12 are now relevant for the website/TV cross-media reach and frequency analyses. We recommend that the proof of concept is run separately for online vs TV, as well as overall.

As mentioned before, we are concerned that we lose a lot of the single source value of the BARB panel by sub-dividing the panel id into "cookies" which are not connected across types. We may also need to further sub-divide cookies to ensure a representative allocation of impressions by other planning criteria (time of day, media content) to homogeneous groups. Apart from anything, this moves further away from the aspiration of unique, privacy-safe identifiers which transcend all media components and provide perfect, single-source duplication measurements. In the meantime, we could take the view that we should not throw away the value of any cross-media identifiers that we already have, in particular the BARB panel id which transcends all TV channels/platforms/devices. Therefore, we repeat the suggestion that the first impression of a cookie determines a "campaign permanent" link to a VID.

This modification could embrace fragmentation of types to recognize differences in allocation probabilities by other criteria such as time of day or programme content. The potential downside is that the VID destination of a person's first campaign impression is also the destination of all their subsequent campaign impressions. It is unlikely that this is optimum for controlling duplications with websites that the common identifier does not transcend, because we may well need the flexibility to allocate every impression to an appropriate group. In fact, this is the start of the argument which leads us down the more traditional data fusion model which depends upon matching people across media channels in terms of their normative behaviour. However, such solutions are less accessible because it would be necessary to make a deeper dive into private website data. Therefore, we still see this (a panel id "cookie" hash procedure which transcends types) as a potential follow-up to an unsatisfactory outcome from the proof of concept.

3.12 Demographic Overlay

*Operates in the same way as for website **types**.*

This assumes that BARB demographic classifications are complete and match the demographic profiles of the single source panel. This is not an issue for the proof of concept which is designed to establish the accuracy of the underlying VID model. However, there is a guest demographic consideration.

3.13 Guest Demographic Expansion

Is required to convert limited guest demographics to a different demographic granularity.

We only need this procedure if the required reporting demographics are inconsistent or more granular than the guest demographics collected in the BARB panel. It follows the demographic correction/assignment procedures for websites described in sections 2.14 to 2.17. The key difference is that the transition matrix must be trained using the BARB panel because:

- Guests are unlikely to be collected in the single-source panel.
- Even if they were, training in the BARB panel is statistically sound because the demographic correction algorithm for one type is independent of behaviour in all other types.
- It may sound circular, but it's not. We are converting a BARB macro model factoring to a micro model person level equivalent.

The performance of this part of the model is evaluated in terms of goodness-of-fit for the large number of campaigns used for training. We recommend that the proof of concept is operated for a set of demographics which are consistent with BARB guest demographics and for a set which are not.

3.14 On-the-Fly Calibration

If TV regression-to-the mean is unacceptable then within the VID construct, we should consider:

- Increasing the dimensions by which type is defined to embrace other planning criteria (as discussed before).

However, it may be necessary to develop an on-the-fly calibration routine:

- To be operated for each campaign (or slice of it) after VID impressions have been collated, counted and configured into a reach and frequency analysis. As mentioned in section 3.2, the underlying probability model in the VID solution may provide the basis for this calibration.
- Calibration is likely to compromise actionability, for example to generate instances of negative reach build.
- Requires calculation of BARB gold standard reach and frequency analysis for each campaign and slice of it – this is possible but could not be described as accessible.
- May compromise otherwise representative online/TV duplications.

We recommend that calibration should not be a feature of the first proof of concept run.

4. Broadcaster Video-on-Demand (BVOD)

BARB has its own hybrid census/panel model, Dovetail Fusion, for reporting programme audiences on broadcaster video on demand platforms. Published data includes:

- Web-DB2: aggregated programme audiences by player, device, activity and demographic.
- PVX: fused census and panel respondent level viewing records, mainly used for reach and broadcast/player duplication-based analysis.

However, BARB does not currently report audiences to BVOD campaigns.

Therefore, Project Origin would be based on direct access to broadcaster BVOD census data, with a methodology broadly equivalent to that used for websites. In principle, the single source panel would provide sufficient input to the VID model, but there may be a calibration advantage in using the Dovetail Fusion programme-based data.

There are some issues to consider:

- Linear commercial impressions generated through catch-up TV, within a BVOD service, on a TV set, are captured as timeshift viewing and reported by BARB as part of the currency. We note that these impressions must not be double counted, but this is not a key issue for a proof of concept study.
- For each BVOD commercial impression, there is often more than one person viewing, at a level which cannot be ignored.
- Consequent adjustments to the algorithm must not distort consistency between the VID model parameters and the allocation algorithm.
- Estimation and application of viewer-per-view factors introduce elements of regression-to-the-mean and sampling error into census-based campaign impressions.

4.1 The BVOD-VID Model in Outline

In the first phase:

- The single source panel provides a representative measurement of BVOD campaign person impressions, naturally picking up the fact that each machine impression is often viewed by more than one person.
- The same model and process is therefore extended to create an additional set of parameters for BVOD **types** (players/devices, i.e. each census silo).
- In the population of virtual people, each VID has a unique identifier and a relative rate of exposure for each (further extended) type of impression.

An additional phase is required:

- To calculate viewer-per-view factors, one set for each **type** (player/device) of BVOD impression.
- For the large number of training campaigns, the single source panel is used to calculate the observed number of machine impressions (consistent with the census data) viewed by 1, 2, 3 etc people.
- When this frequency distribution is converted to proportions (which sum to 1.0) it forms the viewer-per-view factors for each **type**.
- This assumes that the panel is a household sample which measures all individuals in each household, ideally but unlikely to include guest viewing.
- The result is an additional set of parameters which are not linked to the VIDs nor their parameters.

In the second phase, operated by each broadcaster for a particular, live campaign:

- As each impression is registered and before it is allocated, we need to determine how many different people viewed that impression.
- This is according to a random selection with probability proportional to the viewer-per-view factors.
- A number of replicates of the impression are made.

- Then allocate each replicated impression to different VIDs, at random with probability proportional to the relative rate of exposure for that type. We need to keep track of the cookie-VID links for the campaign.
- Unless the impression has a cookie, which has already been allocated to a number of different VIDs, in which case the new impression is allocated to each of those same VIDs.
- Unless the required number is more than the number of previously cookie-linked VIDs, in which case an additional VID is selected for impression allocation.
- Or the required number is less than the number of previously cookie-linked VIDs, in which case the impression is allocated to only a random selection of the previously cookie-linked VIDs.

The publisher metadata is likely to have a demographic label for only one potential user in a household. Therefore, the demographic overlay will not work without the demographic correction/assignment routine. However, this must be adjusted to incorporate the viewer-per-view factors and the fact that one demographic label converts to a number of different demographic cells.

- Since the calculation of transition probabilities is based upon comparison of percentage profiles for both the panel and census, we do not need to consider viewer-per-view factors in this calculation.
- We need to keep track of the maximum number of times an impression has been replicated for each cookie and the cookie-demographic links for the campaign.
- For each replicated impression, we need to identify the demographic label in the metadata and allocate the impression to a target demographic cell with probability proportional to the transition probability.
- Unless the replicated impression (a) has a cookie which already has a number of cookie-linked demographic cells and (b) does not exceed its previous maximum number of replicates, in which case the allocation is restricted to those demographic cells.

We believe that by determining the number of replicates independently for each impression, our suggested approach has not further confounded the issue of consistency between the calculation of relative rates of exposure (impressions vs campaign unique cookies) and their application in the allocation and demographic correction algorithms.

As already stated, the integration of BVOD census data closely follows the methodology for websites, therefore all the issues discussed in section 2 are relevant. In the following sections, we only discuss issues and proof of concept requirements which are specific to BVOD.

4.2 Calculation of Model Parameters

The single source panel provides a representative measurement of BVOD campaign person impressions, naturally picking up the fact that each machine impression is often viewed by more than one person.
The same model and process is therefore extended to create an additional set of parameters for BVOD ***types*** *(players/devices, i.e. for each census silo).*
In the population of virtual people, each VID has a unique identifier and a relative rate of exposure for each (further extended) type of impression.

Remember that this phase operates within each cell of the demographic overlay. There are no BVOD specific issues to consider.

4.3 Viewer-per-View Factors by Type

To calculate viewer-per-view factors, one set for each ***type*** *(player/device) of BVOD impression.*
For the large number of training campaigns, the single source panel is used to calculate the observed number of machine impressions (consistent with the census data) viewed by 1, 2, 3 etc people.
When this frequency distribution is converted to proportions (which sum to 1.0) it forms the viewer-per-view factors for each ***type****.*

In reality, viewer-per-view factors are highly dependent on time of day, day of week, programme content and size of household. Because our suggested approach applies these factors independently for each impression, if these additional criteria are available in the metadata, then it would be possible to refine the factors without confounding the overall VID process.

We recommend that regression-to-the-mean in the viewer-per-view factors is evaluated as a diagnostic tool when they are calculated. It will also be instructive to calculate their sampling errors to demonstrate the variability that they introduce into census campaign impressions. Never-the-less, we recommend that the proof of concept is first run with factors which vary by type only. It is possible that this source of regression-to-the-mean will not be visible in whole campaigns.

4.4 The Single Source Panel

This assumes that the panel is a household sample which measures all individuals in each household, ideally but unlikely to include guest viewing.

A single source panel which does not measure all individuals in each household will not show how many people are viewing each machine impression. Guest viewing may also be missing. However, viewer-per-view factors do not require a single source dataset and can be sourced from an independent operation.

4.5 VPV Parameters

The result is an additional set of parameters which are not linked to the VIDs nor their parameters.

This configuration is relatively easy to conceive and it may be useful to avoid such a link.

We acknowledge that there may be a more elegant solution in which the viewer-per-view factors are incorporated into the VID parameters and/or the demographic transition matrices. So far, we have been unable to untangle the differing demands of the allocation algorithm in order to facilitate such a configuration. We suspect that it would require us to sacrifice the cookie hash procedure – we don't recommend that.

It may be that there is some value in considering a configuration of the VIDs into household structures, but:

- This is not essential for cross-media reach and frequency applications, where co-viewing targets are unlikely to be analysed.
- It may improve the logicality of the cookie hash procedure in each media type, but is unlikely to be optimum for prediction of cross-media duplications because it creates an additional constraint on allocation to the most appropriate demographic/VID.
- To be effective it really requires full household composition to be part of the broadcaster's demographic label.

4.6 Replicating Impressions to Match People Viewing.

As each impression is registered and before it is allocated, we need to determine how many different people viewed that impression.
This is according to a random selection with probability proportional to the viewer-per-view factors.
A number of replicates of the impression are made.

If the VPV for a type says that 20% of impressions are viewed by 2 people, then each impression has a 20% chance of being replicated twice.

Each replicate must retain the metadata, the cookie id and an additional replicate id.

4.7 Impression Allocation

- *Then allocate each replicated impression to different VIDs, at random with probability proportional to the relative rate of exposure for that type. We need to keep track of the cookie-VID links for the campaign.*

So far this follows the website methodology. But we do need to keep track of all the VIDs that a cookie has previously been linked to for the campaign. We assume (probably naively) that this can be incorporated in a variation of the cookie hash procedure without causing a computational nightmare.

4.8 Cookie Hash Procedure

Unless the impression has a cookie, which has already been allocated to a number of different VIDs, in which case the new impression is allocated to each of those same VIDs.
Unless the required number is more than the number of previously cookie-linked VIDs, in which case an additional VID is selected for impression allocation.
Or the required number is less than the number of previously cookie-linked VIDs, in which case the impression is allocated to only a random selection of the previously cookie-linked VIDs.

The cookie hash procedure is more complicated because each cookie has impressions which are each replicated a different number of times. The number of cookie-linked VIDs may well increase as the campaign progresses and we must not use all of them if they exceed the number of replicates for a particular impression. It would be simpler and possibly more stable to fix the number of replicates based on the random outcome for the first impression, but this is not how co-viewing works and would distort the frequency distribution.

We do have a residual concern that our approach has introduced an inconsistency between the definition of the relative exposure rates and their application in the allocation algorithm. Therefore, we recommend that it is even more important to run the proof of concept with and without the cookie hash procedure.

4.9 Demographic Overlay and Correction/Assignment

Since the calculation of transition probabilities is based upon comparison of percentage profiles for both the panel and census, we do not need to consider viewer-per-view factors in this calculation.

In the construct described, within each type, there is only one set of viewer-per-view factors. Therefore, applying the VPVs will increase the absolute number of census impressions to match the panel definition. However, this multiplication cancels out if we work in terms of percentage profiles.

If we develop VPVs to vary by time of day or other criteria, within each type, then a correction of census data would be required in this step.

The calculation of transition probabilities follows the website procedure.

4.10 Allocation of Impressions to Demographic Cells

We need to keep track of the maximum number of times an impression has been replicated for each cookie and the cookie-demographic links for the campaign.

In the algorithm, impressions are assigned to demographic cells before they are assigned to VIDs. As the campaign progresses, the number of cookie-linked demographic cells will increase. The relevant number to keep track of is the maximum number of times an impression has been replicated, even if some replicates have been allocated to the same demographic cell.

It is actually more complicated than this, in that we need to also keep track of the number of times each demographic cell has been linked to each cookie. We are again assuming (probably naively) that this can be accommodated in a variation of the cookie hash procedure without creating a computational nightmare.

Allocation to demographic cells takes place before allocation to VIDs.

4.11 Cookie Hash Procedure

For each replicated impression, we need to identify the demographic label in the metadata and allocate the impression to a target demographic cell with probability proportional to the transition probability.

So far this follows the website procedure.

4.12 Cookie Hash procedure

Unless the replicated impression (a) has a cookie which already has a number of cookie-linked demographic cells and (b) does not exceed its previous maximum number of replicates, in which case the allocation is restricted to those demographic cells.

The cookie hash procedure has to be adjusted to recognize that while the number of demographic cells linked to a cookie is likely to increase as the campaign progresses, a new impression might be replicated fewer times. Therefore, each replicated impression is allocated to only a random selection of the cookie-linked demographic cells, with probability proportional to the transition probability. Note that if a demographic cell has previously been linked twice to a cookie, then its transition probability must be doubled.

Again, we recommend that by running the proof of concept both with and without the cookie hash procedures, we will check that there is no inconsistency between the definition of the demographic transition probabilities and their application in the allocation algorithm.

5. TV Set-Top-Box Return Path Data (RPD)

In principle, STB census RPD provides an input to the VID model which is equivalent to BVOD in that there will be a stream of impressions and metadata which could be used in the same way. However, the metadata is richer and RPD sits somewhere between a BVOD dataset and the BARB panel:

- There is usually a perfect “cookie” in that there is a permanent link between the individual STB and the RPD records. This is like a panel id in that we only have to worry about new and lapsed customers, the equivalent of panel turnover.
- The metadata may contain information on household composition, i.e. demographic labels for each person in the individual STB household. Therefore, each impression can be assigned to one or more people in the household before going being allocated to a VID – basically RPD can be changed from a household to a person-based construct, consistent with the simpler one-to-one matching in the VID website and TV versions.
- Leverage of the household composition data requires that the single source panel is household-based. Alternatively, person assignment probabilities can be informed by BARB panel data.

We recommend that RPD does not need to be in a first run of the proof of concept.

6. Alternatives to the VID Model

It is important to recognize that the VID model provides an actionable and accessible solution. Hopefully RSMB's variations to accommodate BARB panel and BVOD census data have not compromised this accessibility.

There is also a sound statistical theory which underpins and links together all phases of the solution.

All-in-all, this is an elegant solution with a sensible trade-off between likely accuracy in the prediction of cross-media duplications (reach and frequency) and the practicality of the process. It's a trade-off that RSMB are used to making, even agonizing over, every time we approach a data fusion project.

Inevitably there are other criteria that we would wish to take account of when matching impressions with the most suitable VIDs or allocating demographics, ideally to further reduce conditional dependence. We've mentioned time of day and media content – to be fair the architects have by no means ruled out such enhancements. But they do potentially compromise both actionability and accessibility and may not significantly improve accuracy for the intended applications.

In short, the VID solution deserves a proof of concept evaluation in a protype study. Without wishing to "pass the buck", RSMB's interest is in the accuracy of the solution and we have identified many places where the solution may compromise such accuracy. We've been on the giving and receiving end of such evaluations and everyone realizes that this is the correct thing to do. Then all parties who have to agree to use the VID solution, or to explore alternatives, will be as informed as possible about the compromises they are making – how accurate is acceptably accurate?

Before moving on to our suggestions for the proof of concept, it is worth outlining some alternative methodologies to get a feel for alternative trade-offs.

6.1 Calibrating the Single Source Panel

This approach is attractive because it makes maximum use of the single source panel. We don't like to pigeon-hole any technique because components of any methodology inevitably leak into other methodologies, but this is probably best described as an on-the-fly calibration.

The idea is that for any campaign, the single source panel can be used directly to calculate a people based, cross-media reach and frequency analysis, with demographic profiles. It's derived from a panel; therefore, we have to deal with issues of panel turnover and non-reporting but the research industry has long-standing calculation conventions for this purpose. The panel is unbiased and therefore provides an accurate measurement of impressions, demographic profiles, cross-media duplications and reach and frequency. However, there are weaknesses:

- There is an increasingly long tail of channels/websites and the panel does not have the granularity to measure or report this detail.
- Even for relatively robust measurements, there is enough sampling error to create noticeable divergence from census impressions or panel-based currencies such as BARB.

One solution is to calibrate the cross-media reach and frequency analysis to:

- Website census and BARB panel impressions.
- Or to a full currency status reach and frequency analysis for each website or TV channel, or indeed all TV channels combined from the BARB panel.

This can be thought of as a complicated rim-weighting analysis, where the cells are defined to be permutations of cross-media frequencies, then cell counts are iteratively adjusted until all marginal impressions or reach and frequency targets are achieved. The smaller the adjustments required, the more the single source panel cross-media duplications are preserved. As discussed in section 3.14, a more sophisticated calibration would be based on the underlying reach model, then the model parameters are adjusted rather than direct adjustment of the cell counts; this reflects the calibration routines used in the BARB calculation procedures to deal with panel turnover and guest viewing.

This solution is a significant development and it has some drawbacks:

- As the campaign builds, the calibration must be re-run. This is bound to generate inconsistencies such as negative reach build.
- This is aggravated by the likely calibration gap.
- The ability to deal with small channels/websites is limited by single source panel sample size.
- Heavy lifting is transferred to the bureaux systems which collate component impressions and/or reach and frequency analyses and provide the cross-media reach and frequency estimators.

In all, this on-the-fly methodology is likely to be more accurate but there are compromises to actionability and accessibility.

It is worth noting that this methodology may be necessary to provide benchmarks on accuracy for the proof of concept.

6.2 Macro Model

A currency status reach and frequency analysis is run for the components of a particular campaign:

- Each website/device
- All channels combined from the BARB panel
- Each BVOD/platform/device

Elements of the VID methodology or their macro model equivalents are used in the construction of reach and frequency for individual websites and BVOD services.

Based on a set of training campaigns, the single source panel is used to estimate the duplications between cross-media components. These inform a model which predicts cross-media reach and frequency, given the campaign's component reach and frequency inputs.

We have the following comments:

- The process is reasonably accessible although some heavy lifting sits with the reach and frequency bureaux.
- It is automatically calibrated to census/currency data.
- We believe that accuracy of campaign cross-media duplications will be better on average, but potentially less nuanced, than the VID model.
- Actionability is likely to be compromised in campaign slicing and dicing.

6.3 Respondent level Data Fusion

We identify two potential sources of regression-to-the-mean in the VID model:

- Without excessive splitting of cookies, the information used to create a behavioural link between impressions and virtual people is limited.
- Allocation of the first impression determines the allocation of all other impressions for that cookie – probabilistically OK but likely to be volatile.

In theory this can be improved with a normative data fusion:

- Based upon a set of training campaigns and the single source panel, a normative cross-media viewing signature can be created for each VID. This signature can be extended to incorporate other criteria such as time of day and media content.
- Each component of the overall signature will relate to an individual media type.
- Based upon the same set of training campaigns and an individual media type, the equivalent normative viewing signature is created for each cookie or BARB panel member.
- Then a data fusion is applied to permanently assign each website cookie or BARB panel member to a VID.
- As impressions are registered in the donor datasets, they are assigned to the cookie-linked VIDs.
- The advantage is that normative, multi-faceted, behaviour within each media type is likely to correlate better with cross-media behaviour.

In theory, this would provide the VID model with enhanced accuracy, but there are a number of issues:

- We believe that cookies have a lifespan of days rather than weeks – there is a danger that most will have been deleted by the time the fusion cohort is constructed. Therefore, this enhancement is only relevant if websites have more persistent ids.
- The BARB TV panel is a better prospect – it is possible to operate a mixed approach for TV vs online. However, we would need to consider how to deal with panel turnover.
- It may be that this respondent level fusion solution is only realistic for post-campaign evaluation. As such, this would not match the requirements of actionability, in particular in-flight planning.

6.4 A Simple Probability Solution

If the underlying probability model fits well, then given the number of impressions for each media type, it is possible to predict a campaign frequency distribution for each homogeneous VID group using the parameters of the model. This avoids the allocation algorithm. It is both accessible and actionable, but we must consider the following:

- Unlike the VID process, this solution is totally dependent on the goodness-of-fit of the underlying probability model. The VID allocation algorithm transfers across whole cookies or BARB panel members, to an extent preserving real reach to frequency relationships.

However, an evaluation of this approach can be a natural feature of the prototype.

7. The Prototype – Proof of Concept

RSMB's focus is on the proof of concept component of the prototype. This means that we will concentrate on the statistical performance of the models rather than the operability, flow of information through the process or the validity of the input data.

We do flag the potential importance of the cookie hash procedure. We understand that this is (perhaps paradoxically) a deterministic element in a random process, which avoids storing large look-up tables of VID to cookie links. We will assume that the linkages we require will not be obscured in live application.

Throughout our description and interpretation of the VID methodology, we have noted sources of potential failure in model assumptions and performance. It is important to recognize that we should not cancel development if any particular step has poor performance, because the overall outcome may still be robust. However, consideration of the individual steps is important for diagnosis of overall performance.

When benchmarking VID cross-media reach and frequency against the "truth", for example the raw single source panel results, it is important to mitigate the confounding effects of sampling error. This means that definitive conclusions must be based on robust benchmarking. In other cases we have to use our judgement and a consensus view on acceptable accuracy.

It will also be important to understand which specific applications are supported by the VID solution. For example, it is conceivable that the model works for reach and average frequency, but not each individual frequency. We expect to identify a set of analysis capabilities. It is often the case that replacing a traditional survey with a modelled solution requires a more focused set of analysis requirements.

7.1 Regression-to-the-Mean Evaluation

Essentially this measures the accuracy of the VID model.

Regression-to-the-mean may occur when there is a violation of the assumption of conditional independence. In the context of cross-media reach, it could be that general cross-media weight of viewing permutations (the viewing signatures controlled in the allocation of cookies to VIDs) are not predictive of those for a specific campaign. For example, if a campaign is only served in the morning when all-time light viewers may be relatively heavy viewers of one media segment. It is also important to recognize that RTM can occur across dimensions which are controlled in the VID model. This is more subtle and arises because cookies/impressions are assigned to VIDs on a probability basis.

To be fair, all RSMB's fusion products suffer from regression-to-the-mean. The important action is to make every effort to quantify this.

RTM will be evident in the following application metrics:

- Reach duplications between permutations of media types (websites/channels/platforms/devices).
- Consequently, reach and frequency analysis.
- Sliced and diced by criteria such as time of day or campaign length.
- By demographic group.
- Demographic profiles of campaign impressions in situations where demographics are modelled or corrected.

The bottom line is that we need to start with a large number of cross-media campaigns which can be measured robustly by a single source panel and compare the metrics with those produced by the model. An important benchmark is to construct the metrics in each media type separately and then compound across types at random. Then regression-to-the-mean is the degree to which the difference between truth and random has been eroded in the model results.

For small types or other small components of campaigns, it will not be possible to produce robust estimates of the truth. Therefore, we are limited to assessing robust averages of small components within a class, or making a pragmatic assessment of the difference between model and random.

7.2 Model Performance

While regression-to-the-mean is the ultimate statement of performance, we need to do our best to diagnose and identify sources of RTM. We need to measure the performance of the steps in the model process. For example, the model depends upon breaking down the population into a set of homogeneous sub-groups, but what proportion of the overall variance amongst the population is explained by variance within groups, compared to variance between groups?

Core model components need to be tested:

- Creation of homogeneous groups. Between vs within group variance in rate of exposure parameters. Measure lack of homogeneity across other dimensions such as time of day.
- Underlying probability model. Goodness-of-fit in terms of reach and frequency, for training campaigns. Ability to deal with outliers, e.g. targeted or extreme campaign plans.
- Demographic correction. Compare modelled transition matrix with actual single source panel equivalent (assumes publisher demographic labels are available in the panel).
- BARB panel guest demographic correction. Goodness-of-fit of transition matrix compared to BARB guest factoring conventions.
- Calculation of BVOD viewer-per-view factors. Measure variance around the average and systematic variation across other dimensions such as time of day.

All the above are potential sources of RTM. Additional sources are:

- Restricted set of control dimensions. Ambition restricted by conflict with cookie hash procedures and actionability (from first campaign day).
- Frequency capping. Defies assumptions of conditional independence. Beneficial effect of cookie hash procedure is limited by cookie deletion.
- Small websites/channels. Smoothing effects of generic rates of exposure.
- Cookie hash procedure for the BARB panel. Independent operation within each channel/platform/device loses single source value of panel id within TV, but may be

optimum for TV website duplication. However, if the procedure consolidates all TV channels (types), then all a panel member's subsequent impressions will be allocated according to the first impression type, irrespective of the subsequent types.

- Demographic correction. We "know" how many cookie labels are incorrect but not which. Therefore, some will change erroneously.
- BARB panel guest viewing cannot benefit from the cookie hash procedure.
- Allocation algorithm for BVOD. Each cookie is linked to multiple VIDs, but only a random selection of the linked VIDs are allocated each impression. Cookie hash has reduced control.
- Demographic correction for BVOD. Similarly, cookie hash has less control.

At various points in sections 2 and 3 we have recommended that more than one variation of the VID model is tested in a proof of concept. This will help to demonstrate trade-offs in accuracy between the different application metrics. In some cases, it will provide reassurance (or otherwise) over the residual concerns we have expressed in the consistency between the calculation of model parameters and their application in the demographic correction and allocation algorithms. Note that this relates to RSMB's application of the WFA framework to BVOD, so an issue entirely of our own making! We can summarise the variations as follows:

- Abstract (as described) vs empirical reach model parameters. (2.5)
- With and without the cookie hash routine. Trade-offs may be different for single media types vs cross-media duplications. Tests the residual concern above. (2.10)
- With and without BARB panel guest viewing. (3.7)
- For demographics which do vs don't require guest demographic correction. (3.13)

We also have some suggestions for developmental or diagnostic analyses which we do not consider a top priority for the proof of concept.:

- Analysis of variance to determine the optimum criteria for identification of homogeneous groups. (2.5, 2.12)
- On-the-fly calibration to force consistency with individual media components (e.g. BARB currency reach and frequency). (3.14)
- Test a cookie hash procedure which transcends all BARB channels. (3.11)

As discussed in section 3.11, with the BARB panel input we have the ideal opportunity to test a cookie hash procedure which transcends all TV channels and platforms. We do understand that this is a very important feature of the WFA blueprint. However, we recognise that this is highly dependent on other components and outcomes of the VID model. For example, can we extend the definition of types to embrace dimensions which split cookies but still put cookies back together? Therefore, we believe that the design and execution of this part of the proof of concept should be informed by a first phase.

7.3 Proof of Concept Design

We recommended that the proof of concept is most efficient if it is based upon a simulation within a single source panel, rather than introducing the complication of census data at this stage. Ideally, this single source panel must:

- Be household based and measure all individuals in the household.
- Collect TV data equivalent to the BARB panel.
- Measure website impressions and metadata.
- Including cookies, publisher demographics, other planning dimensions.

Compromises should be accepted in the interests of getting an early statement of performance or dealbreakers.

In the first instance we recommend limiting the study to (say) 6 websites plus 6 TV channels and their BVOD players.

It will be important to evaluate performance separately and in combination for websites, TV and BVOD.

The most convincing experimental design is a split-sample or fold-over study:

- The panel is divided into two representative halves.
- Comparison of the two halves gives an indication of sampling error.
- One half is used to train the VID model.
- This same half also provides the benchmark truth for the analyses described in sections 7.1 and 7.2.

- Information is stripped from the other half to simulate website, BVOD player and BARB panel data.
- The VID model is applied to this second half and compared to the benchmark truth.

There is likely to be a strain on sample size, in which case a goodness-of-fit test in the whole panel may be more effective.

We recommend that this design will provide sufficient content for a proof of concept which:

- Evaluates performance of the basic WFA blueprint.
- Identifies areas for development which can address shortcomings.
- Facilitates agreement to move forward with a full prototype study of the VID model and/or demands that alternatives are considered.

Appendix A: Virtual People: Actionable Reach Modelling – RSMB Questions

Following the WFA peer review sessions and a more thorough reading of the technical documents referenced in the WFA Draft Proposal, our understanding of the methodology had improved. In order to remove any further confusion and precisely determine further details, the following list of questions was compiled and shared with Google and Facebook. The questions are based on Google's paper, Virtual People: Actionable Reach Modelling; the numbering reflects this. Responses from Google and further back and forth are included; beyond this, the meeting of 17th September led to the resolution of some remaining queries.

General Questions

Q1. This is a general issue rather than a question. It manifests in some of the specific questions below. It is not always clear to us if we are modelling/allocating individual impressions (one cookie with one impression) or unique campaign cookies (one cookie with at least one impression in a particular campaign) or unique browsers (one cookie with any browser behaviour in a defined period).

Response: We are mapping impressions to people. We do this based on impression data and use cookie (generally user_id) hash as a source of randomness. Thus if a cookie did not change relevant information from event to event (inferred demo, geo etc) then the cookie is mapped to the same person across all impressions.

2.1 Demographic Correction Model

Q1. We recognize that this is a macro model version provided for background understanding. Can you confirm that the metrics are campaign impressions rather than unique cookies?

Response: The model can be used for two metrics: estimating the corrected demographic distribution of impressions and campaign reach estimates by corrected person demographic buckets. If the redistribution matrix is calculated via a confusion

matrix then impressions should be used for the confusion matrix, as it's important to take user activity into account. If the matrix is trained by fitting audience distributions then cookie counts should be used, as they are less noisy and using campaigns already accounts for user activity.

2.2 Mapping Cookies to Users

Q1. The reach function maps the cookie reach vector to the total number of unique people reached (across all cookie types). So, in this case are the metrics unique campaign cookies? It is difficult to see how this is consistent with the VID allocation algorithm which we think operates at an impression level.

Response: Cookie is used as a source (seed) of (deterministic) randomness when impressions are mapped to people. Thus the model essentially operates on bucketized cookie count vectors.

Q2. Do the alpha parameters in the Dirac Mixture have a meaning? We think they may be the proportion of the universe in each sub-population/pool and everyone in that pool has the same Poisson model to generate reach.

Response: Yes, alpha parameters are proportions in which the population is broken by pools. People in one pool have identical activity.

Q3. The Dirac mixtures are trained to the observed SSP data. Does this mean:

- Minimising variance within pools and maximising variance between pools in terms of the activity vector x?
- Maximising variance between pools in terms of the overall people reach R(t)?
- Counting the panel universe in each pool or estimating a theoretical universe (alpha) for each pool which has the best fit to campaign reach for a lot of campaigns?
- Similarly, calculating the elements of vector x from the SSP or estimating theoretical values for each pool which have the best fit to campaign reach for a lot of campaigns?

Response: It means generating such pools and activities that minimize the error between panel measurement and ADF-based estimates.

Q4. In the modelling and assignment processes (as opposed to arbitrary slicing of campaigns for planning purposes), is it assumed that a cookie type cannot be classified by any criteria (e.g. time of day) which sub-divides cookies)?

Response: Criteria that subdivide cookies, like time of day, or topic of the video, can be used in modelling, but in moderation. Using such features makes different impressions of the cookie assigned to different people, which may cause reach inflation if the model overfits. Using them is particularly useful when modelling co-viewing on CTV. Larger panel size may be required to train coefficients for these features.

4. Virtual people Assignment

Q1. V is allowed to be undefined for some events, then these events bring no incremental reach. This suggests that an activity (x) cannot be calculated for these events. How does this relate to the reach curve? It worries us that we're not really understanding how the reach curve fits into the allocation algorithm?

Response: If the derivative at 0 of the reach curve is equal to K then it is required that (1-K) portion of events have no virtual people assigned to them. This is the primary motivation for assigning no VIDs to some events.

RSMB Follow-up: I think I'm failing to understand a mathematical principle and/or the notation in the algorithm. The following is what I think. The correspondence functions C(e) and T(c) just ring-fence the user ids and cookie types to which event e could be served. K is then the sum of (population x probability of viewing that cookie type) across all Dirac deltas. I suppose I don't understand why K might not be 1?

Response: I believe your understanding of details are correct. K might be less than 1 because indeed probabilities over the deltas sum to a number less than 1. The

remaining volume of probability is the probability of the event to have no virtual person assigned.

As if we have 3 bins, and each has probability of 0.3. The sum of these probabilities is 0.3. And then with probability 0.1 the ball just doesn't go to any bin.

Events to which no virtual person is assigned indeed bring no incremental reach, but should be counted to overall event count. This artificial mechanism is required for modelling reach curves that have derivatives at 0 less than 1. In practice such curves are indeed observed.

17th September Meeting Outcome:
Google explained that parameters for the best fitting reach curve may have relative exposure rates (x values) which don't sum to 1 within each type. RSMB may propose that for the ISBA/UK prototype, exposure rates are constrained to sum to 1 in the model fitting. Then all impressions/cookies will be transparently allocated to a VID.

4.1 Assignment for Dirac Mixture

Q1. Is the D(x) function/formula itself a feature of the assignment algorithm or only it's alpha and vector x parameters?

Response: **clarification needed**

RSMB Follow-up: This may be a trivial question unless there is an answer I'm not expecting! I'm really just checking that there are no functions in Algorithm 1 which are delivering a derivative at zero of the reach curve equal to $K < 1$.

Response: Technically alpha vector and x activities are sufficient for the algorithm execution. We are mentioning D because we think of alphas and activities to be defining D. Derivative at zero less than 1 is possible, as discussed in the previous section.

Q2. Are the components of the assignment probabilities based upon impressions or unique cookies?

Response: These are features of the impression, which are almost consistent for a cookie. E.g. declared age/gender are technically features of the impression, as the user may change it, but they rarely change.

RSMB Follow-up: This is not really what we're asking. The SSP is used to determine the probabilities in the activity vector x. Then for example, if each VID in one Dirac group has twice the probability of each VID in another group, then we expect them to get twice the number of impressions. However, there is not a linear relationship between impressions and unique cookies. So, in terms of unique cookies, each VID in the first group may have only 1.5 times the probability of each VID in the other group. If we calculate probabilities in terms of impressions but allocate whole cookies, will we distort the reach? I'm having trouble getting my head around this.

Response: I don't quite agree with two premises here.
First:
"*Then for example, if each VID in one Dirac group has twice the probability of each VID in another group, then we expect them to get twice the number of impressions*."
In actuality I believe we have:
"If each VID in one Dirac group has twice the probability of each VID in another group, then we expect them to get twice the number of **cookies**."
And second is subtle:
"*However, there is not a linear relationship between impressions and unique cookies.*"
If cookies are sampled uniformly at random regardless of impressions then the relationship will on average be linear. On the other hand, for real people there can be a skew and indeed the proportions are different.
This question has many moving pieces, it would be useful to discuss it at a meeting.

17th September Meeting Outcome:
RSMB's interpretation of the model is not correct. Throughout model fitting and application in the allocation algorithm, relative rates of exposure are in terms of unique cookies rather than individual impressions. Most digital panels collect cookies so model fitting is achievable.

Q3. Each cookie is mapped to the same virtual person for all events. So only events with new cookies are assigned at random to VIDs. Does this demand that the answer to Q2 is “unique cookies”?

Response: Each cookie is ***often*** mapped to the same person for all events. For instance, if declared age and declared gender are used as features and the cookie did not change their declared age and gender over the course of the reporting period then all impressions of the cookie are mapped to the same virtual person id. But when user_id changes age or gender then the virtual person id is likely to change as well.

Q4. Is each cookie mapped to the same virtual person for different campaigns?

Response: Usually, but not always. E.g. if campaign A occurs before a change in declared age/gender and campaign B occurs after the change then the cookie will be mapped to different virtual people for these campaigns.

Q5. The output is described as a mapping from unique cookies to VIDs. In terms of the data that is passed on outside the publisher, is this only visible as a mapping of events (impressions) to VIDs?

Response: The model technically works by mapping impressions to people, but because of the hashing mechanisms all impressions of a cookie are often mapped to the same person. When we reason about the model, we often use a simplified mind-model where cookies are mapped to people directly.

Q6. We assume that for a particular campaign, every publisher will be running the algorithm against a common set of VIDs. Is that fair?

Response: Yes, all pubs must map impressions to the common set of VIDs.

4.2 Assignment for Dirac Mixtures into People Categories

Q1. Are the components of the conditional demographic probabilities based upon impressions or unique cookies?

Response: They are based on features of the impressions that rarely change for a cookie.

RSMB Follow-up: Again, I'm concerned that the demographic share of unique cookies is different to the share of impressions.

Response: This is a valid concern. The principled way to address it that we know is described in section 5.1. Cookie level demographic correction.

17th September Meeting Outcome:

RSMB's interpretation of the model is not correct. Demographic transition matrices are also based upon unique cookies throughout.

Q2. Is each cookie mapped to the same population category for all events?

Response: It's often the same, but not necessarily. E.g. declared age/gender may change from impression to impression, but we assume that it happens rarely.

Q3. Each publisher may have different granularity of demographic labels (e.g. broader age groups). Is it envisaged that the conditional probability matrix will be different for each publisher?

Response: Yes, a separate demographic probability matrix is assigned to each publisher. It could be a modelling decision to have some publishers share the same matrix.

4.3 Generic Framework: Dirac Mixtures of people Categories

Q4. The output from Algorithm 3 includes a mapping of virtual people to people categories. Does this mean that a unique VID will have different demographics for each publisher?

Response: Mapping of a virtual person to people categories must be consistent across all publishers.