JOHN BATES SERVICES

1 Pickler's Hill. ABINGDON OX14 2BA Tel: Abingdon (01235)-526702

Peer review of multi-modal appraisal methodology

Final Report for Department for Transport
May 2023

1 Introduction

The logit model, in various forms, is extensively used in transport models underlying forecasts, in order to represent transport choices (such as mode and destination). In most cases, its output can be directly used in conjunction with the standard appraisal guidance using the so-called Rule of a Half [RoH], which is a linear approximation to the true demand curve. When large changes in generalised cost are envisaged, it is recommended to use a piece-wise linear approximation which is referred to in guidance as "numerical integration" [NI].

Significant problems are encountered when new alternatives are introduced to the model. The general problem as to how best to introduce them has been discussed in Bates (1991, 1992). But there is also a major problem for appraisal in that the Do-Minimum [DM] costs are not available, and, in terms of the logit model, are effectively infinite. The same problem, in reverse, applies when alternatives are removed in the Do-Something [DS]. The appraisal guidance (Department for Transport (2020)) refers to this issue as "New Modes" [Chapter 12] but this can be viewed as a convenient shorthand for the more general problem of changes to the choice set.

Recently Leeds ITS have led research, funded by the Rail Safety and Standards Board [RSSB] and in part overseen by DfT, to develop appraisal guidance for new railways and rail stations. This proposes an alternative to the TUBA approach which they have termed "multi-modal" (though it may be objected that transport appraisal is in general multi-modal, so this is not ideal terminology).

As a result of some concern about the validity and appropriateness of the method, DfT issued a Brief for a Peer Review in December 2022, and I was awarded the contract towards the end of January. The main document for the review was the November 2021 Final Report on "Rail Openings Appraisal" [ROA] (Ojeda-Cabral et al, 2021). I made it clear that the limited time available for my review meant that I would concentrate my effort on Chapter 4 "User Benefits".

In the next Section I summarise my conclusions on the ITS method and make some general comments on the TUBA approach to "New Modes". I then go on in the following Section to discuss a general way forward for rail models and station choice in particular. The Final Section sets out my recommendations both for practical applications and for further supporting research.

2 The "multi-modal" approach

Chapter 4 of ROA begins with a thorough assessment of the origins of the recommended approach in TUBA, which stems from work by ITS in 2001for Geoff Hyman of DfT. As noted above, the key issue with "new modes" is the absence of a DM generalised cost for the new mode. The TUBA approach is to define "pseudo-DM" costs "which, when passed through the demand model, give very low (but non-zero) OD flows using the new mode, the aim being to find a point on the demand curve that is close to (but not on) the cost axis.

"The appropriate pseudo-DM cost may vary between OD pairs and purposes, depending on the modelling structure." The TUBA proposal is to find the cost which will ensure that "the pseudo-DM has 0.02 times the DS number of trips". It is admitted that this is "potentially quite difficult" and "partly a matter of trial and error." Very little guidance is given, though "3 times the DS cost is recommended as the starting point". We may note in passing that, uncontroversially, NI is recommended, given the significant difference in cost between DM and DS. Nothing is said about separate components of generalised cost, which are an important aspect of TAG appraisal guidance.

In the light of the difficulties of following the TUBA Guidance, the ROA goes back to the original ITS work (Nellthorp & Mackie, 2001) [N&M] where other possibilities were discussed. When working in the context of a uni-modal rail model, and applying PDFH elasticities, the results can be extremely sensitive to the composition (and definition) of generalised cost (essentially because the ratio of cost and time elasticities is unlikely to be compatible with the assumed value of time in the generalised cost formula). At the very least, this suggests that the TUBA Guidance needs to be expanded.

It is also noted in ROA that one of the options proposed by N&M was to apply the RoH to groups of **switchers**, "applying the rule of a half to specific groups switching to a new mode from various other choices, including existing modes, alter[n]ative destinations and pure generation [with]… the first switcher gaining the maximum reduction in generalised cost, and the last (marginal) switcher gaining an amount only marginally greater than zero." This was not followed up at the time because, it appears, the DfT (or DETR, as it then was) was looking for approaches where only own-mode costs and trips were required. N&M considered this unfortunate because the option was "the only one of the five which can easily be disaggregated into components".

In my view, the ROA has made much more of this than is really justified. The authors have taken the idea that the costs of the alternatives from which people switch can make a major contribution to the benefit calculation – hence the "multi-modal" terminology. However, they have not developed the theory in line with general discrete choice analysis, and the approach they propose suffers both from deficiencies and from requiring considerable (unnecessary) effort.

In an attempt to explain the approach, the ROA sets out an example (in §4.5.4) which is an unfortunate over-simplification – in subsequent correspondence they have proposed an alternative version which is now valid, but still requires major assumptions which are largely arbitrary.

I take the view that any method proposed to deal with "new modes" must also have validity in the more general case. In brief, for the general case the ROA proposal is that existing rail users get the full benefit, while entirely new rail users (who were not previously travelling at all) get half the benefit: this is in line with general practice. The remaining rail users in the DS will have switched from other modes. The original version was to give switchers a benefit equal to the difference between their previous mode mean generalised cost and the new rail mean generalised cost, but this fails to allow for the fact that – unless the choice is deterministic – switching can still occur even if the DS mean generalised cost is greater than that for their previous mode. In the alternative version, it becomes necessary to segment the travelling population into those who, in the DS, will use rail, and those who will not, and to construct alternative values of generalised cost.

In the "new modes" case, there are of course no existing rail users. As a simplification, it is proposed that the bus mode is used as a proxy, and that while switchers from bus get the full saving, newly generated trips and switchers from car get ½ the benefit of bus switchers.

In line with the ROA example, consider that there are three modes: car, bus and rail. Following the principles of discrete choice analysis, each mode m has a utility which can be written as $U_m = V_m + \varepsilon_m$, where V is the measurable part and ε is a random term with assumed distribution. We consider an example where only rail changes between the DM and DS: $V^0_{\text{rail}} \rightarrow V'_{\text{rail}}$.

The fact that rail is chosen by some travellers in the DS implies that for those travellers:

$$V'_{\text{rail}} + \varepsilon_{\text{rail}} > \max\,(V_{\text{car}} + \varepsilon_{\text{car}},\, V_{\text{bus}} + \varepsilon_{\text{bus}})$$

Some of these DS rail users will already have been using rail, and they clearly receive the full benefit of $V'_{\text{rail}} - V^0_{\text{rail}}$. But others will have switched from car or bus.

For bus switchers, it was the case in the DM that $V_{\text{bus}} + \varepsilon_{\text{bus}} > \max\,(V^0_{\text{rail}} + \varepsilon_{\text{rail}}, V_{\text{car}} + \varepsilon_{\text{car}})$. For this group, therefore, we must have $V'_{\text{rail}} + \varepsilon_{\text{rail}} > V_{\text{bus}} + \varepsilon_{\text{bus}} > V^0_{\text{rail}} + \varepsilon_{\text{rail}}$. This implies that for the bus switchers:

$$V'_{\text{rail}} - V_{\text{bus}} \;>\; \varepsilon_{\text{bus}} - \varepsilon_{\text{rail}} > V^0_{\text{rail}} - V_{\text{bus}}$$

Hence in switching, bus users obtain benefit of $V'_{\text{rail}} - V_{\text{bus}} - (\varepsilon_{\text{bus}} - \varepsilon_{\text{rail}})$ where the random term is bounded as in the inequality just derived. At its maximum, they obtain no benefit, and at its minimum they obtain $V'_{\text{rail}} - V_{\text{bus}} - (V^0_{\text{rail}} - V_{\text{bus}}) = V'_{\text{rail}} - V^0_{\text{rail}}$. The actual amount will depend on the distribution of the random term, but for relatively small changes, the average benefit will be equal to ½ $(V'_{\text{rail}} - V^0_{\text{rail}})$, which is the standard result from the RoH.

As usual, the approximation becomes worse as $(V'_{\text{rail}} - V^0_{\text{rail}})$ increases (requiring the use of NI). Only when the random term tends to zero or $(V'_{\text{rail}} - V^0_{\text{rail}})$ becomes large is the previous utility (here, V_{bus}) of any relevance. Even then, the benefit is only equal to $V'_{\text{rail}} - V_{\text{bus}}$ when there is **no** random term.

For the general case, the maths becomes difficult, but for the simple binary logit it is tractable, and it is set out in Appendix A. This provides a formula for the average benefit obtained by switchers. It is shown that even in the case of a "new mode", the average benefit obtained by switchers is not equal to $V'_{\text{rail}} - V_{\text{bus}}$: only when there is no random term – so that the choice is deterministic – is the formula valid.

From this, we learn a number of important things. Firstly, it is certainly possible to investigate the benefits obtained by "switchers", though a general treatment is quite difficult, and for relatively small improvements, the benefit is as given by the RoH. Secondly, as long as we are dealing with a logit choice model, it is incorrect to claim that the benefit to switchers is equal to the difference in the measurable component of the utility of their DS and DM choices, even when we are in a "new mode" situation. Thirdly, while it is true that for a deterministic choice model, the benefit **is** equal to the difference in the utility of the DS and DM choices, we need to consider the potential importance of the ASCs, particularly in a mode choice situation. While this is discussed in ROA (§4.3.2), no convincing solution is offered. The RoH, by contrast, does not depend on ASCs, as the benefit calculations are entirely within-mode, and there is no reason to suppose that there is any change in the ASCs between DM and DS.

Based on these remarks, I do not consider that the multi-modal approach offers a useful solution to the "new modes" problem. Rather, it is necessary to improve the Guidance relating to the "pseudo-DM".

A more restrictive definition of the problem

Given the general problems posed by the introduction of "new modes", it is best avoided as much as possible! Careful thought is required as to whether the new opportunity is really to be regarded as "new": for example, should a high speed rail link be regarded as something fundamentally different from existing rail services? Most journeys where there is no obvious rail alternative could still be made by rail in principle, even though the cost and inconvenience make it a wholly unlikely option.

The clearest case where the problem of "new modes" seems unavoidable is in relation to new rail stations, and this is the focus of the next section. However, we will also discuss briefly those cases where the rail alternative is very unlikely.

3 Station Choice and the development of rail demand models

In conformity with work I have done elsewhere, I am going to use "i" and "j" for **zonal** origins and destinations, and "R" and "S" for stations at the start and end of the journey: this is compatible with that used in HS2's Planet Framework Model [PFM] and in the NoRMS model for Transport for the North [TfN]. I am aware that much of the rail literature has used different notation, typically with "i" and "j" for stations.

In stark contrast to other modes, the rail industry has a major data advantage in obtaining regular estimates of station-to-station travel from LENNON. This has led to numerous attempts to model the distribution of demand by means of "gravity models", with the general form of an origin weight, a destination weight, and a separation component. It is natural, and in line with general PDFH recommendations, to construct the separation component out of GJT and Fare, together with their recommended elasticities.

The origin and destination weights are more problematic. Clearly we would like them to reflect the volume of "opportunities" as measured by variables like population and employment. When a well-defined area (like a free-standing town) has a single rail station, this is straightforward. But it is much more difficult when there are multiple stations and interacting urban areas. Conventionally, the approach has been to define a "catchment area", which, in standard terminology, is the area from which the station derives potential demand, and the literature displays various assumptions made in this regard, typically using a radius of influence (within x minutes etc.). In many cases, however, this is difficult to define: there may be competing stations, and there may be features relative to the journey being made (eg fast services versus stopping services).

On this basis, a typical early model for the station-to-station demand Q would have the form

$$Q_{RSk} = F_{RSk}^{\varepsilon_F}.GJT_{RS}^{\varepsilon_{GJT}}.Y_R^{\varepsilon_Y}.O_R^{\varepsilon_O} D_S^{\varepsilon_D} \quad \text{(k = ticket type[1])}$$

where F is fare, Y is an income variable relating to the origin station catchment, and O and D are features relating to the land-use (population/employment etc.) of the origin and destination station catchments respectively.

[1] this is also a feature of the standard LENNON data, but for convenience we will omit it from subsequent discussion. Its main determinant, of course, is fare.

It can be expected that demand will decrease with distance, *ceteris paribus*, so that the exponents (elasticities) on fare and GJT will be negative. Some models seek to combine the two variables into a generalised cost measure. The more problematic item is the appropriate definition of the station catchment areas.

As noted, the catchment area of a station is defined as the set of zones from which it derives potential demand. From a demand modelling point of view, however, it is more helpful to use it in the inverse sense – so that for any zone, it is the set of stations that might be accessed (on the production side), or which might serve as the alighting station for the final destination (on the attraction side). It is reasonable for the production and attraction sets to be different.

In principle, with a model defined on station-to-station movements, a new station could be introduced (or one removed) and the model applied. The problem is that this would not deal with underlying issues of station choice/competition. There has therefore been a (sensible) tendency to redefine the model on a zonal basis, and to introduce access and egress. This leads to models of the type suggested by Wardman *et al* (2007)[2]:

$$Q_{RS} = GC(F_{RS}.GJT_{RS})^{\varepsilon_G}\left(.\sum_i Pop_i^{\varepsilon_{PO}}.Acc_{iR}^{\varepsilon_A}\right)\left(\sum_j Pop_j^{\varepsilon_{PD}}.Egr_{Sj}^{\varepsilon_E}\right) \quad \text{(Eq 5)}$$

$$Q_{iRSj} = K.GJTAE_{iRSj}^{\varepsilon_G}.F_{RS}^{\varepsilon_F}.Pop_i^{\varepsilon_{PO}}.Pop_j^{\varepsilon_{PD}} \quad \text{(Eq 10)}$$

In (Eq 5) the population in potential catchment area zones (raised to a power) is multiplied by Access/Egress times (also raised to a power) and then summed over all potential areas. By contrast, in (Eq 10), access and egress are combined with GJT. Clearly alternative formulations can be proposed, and there is also the issue of whether separate GJT and fare elasticities are appropriate or whether some definition of GC can be used (as implied in (Eq 5)). But dominating both types of models is the definition of zones within the catchment area. This will also depend on whether the focus is on general travel or only on long-distance travel.

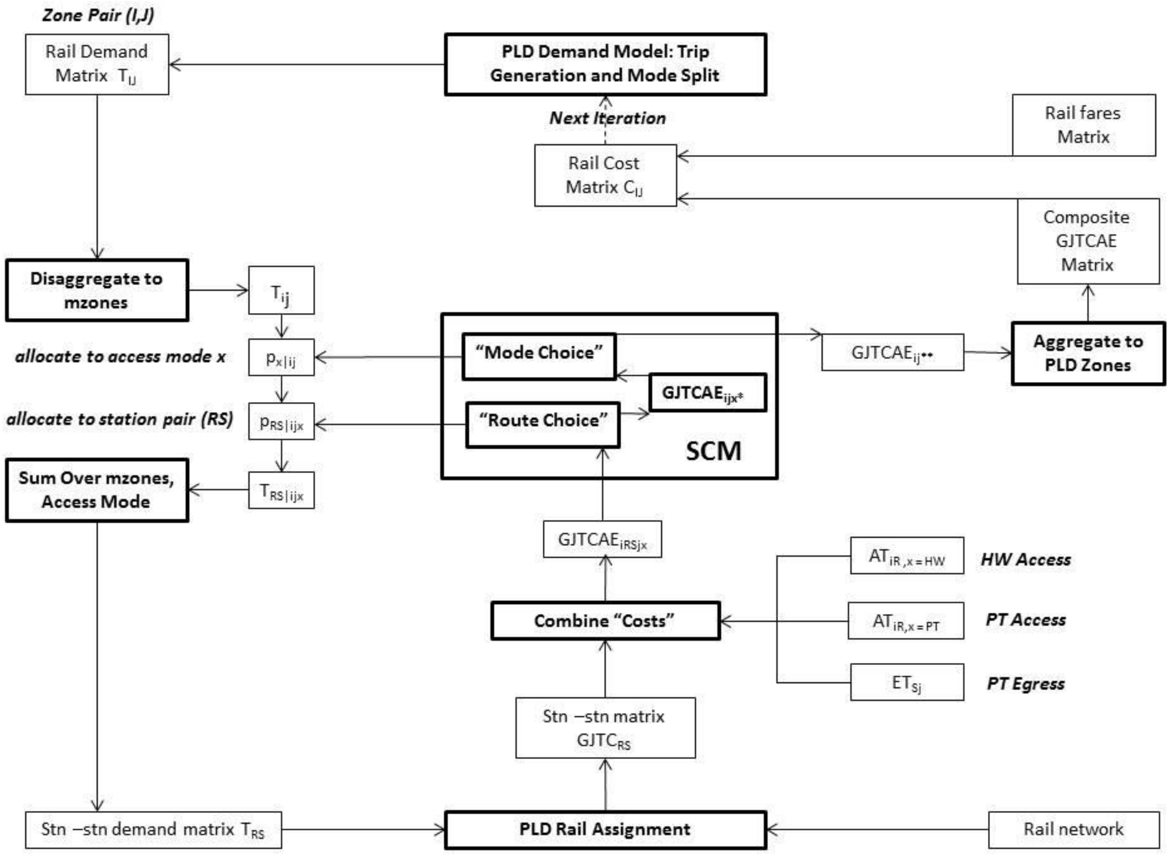
Wardman *et al* report (p 138): "For the subsequent station choice aspect of the work, potential competing stations were defined as those within 20km of at least one origin zone. Candidate competitor stations are ordered by criteria calculated from the product of the total number of journeys originating at the candidate station and the population of a zone, divided by the distance to the center of population for that zone, then summing across all zones for the origin station. The candidate stations are sorted in decreasing order of these criteria, and the top 15 form the competing stations. At an early stage, tests were conducted on whether 5, 10, 15, or 20 competitor stations should be specified, and 15 provided the best fit." (it was noted that some stations do not have the full 15 competing stations).

There is a problem, however, with any definition based on an arbitrary number, particularly if stations are to be introduced or removed. One of the issues is that there is little data available as to what stations are actually used for journeys originating or destinating in particular zones. An important exception is the 2004 National Rail Travel Survey [NRTS] which contains the required detail for a large sample of journeys. Although it is now very old (it is understood that a new survey is planned), it is the best source available for modelling station

[2] The equation numbers that follow are aligned with those in Wardman *et al*, though minor changes have been made in the notation.

choice, and has been used for the station choice models [SCMs] in HS2's PFM and (to some extent) TfN's NoRMS.

These SCMs are based on discrete choice principles. In both cases, they are nested within a larger transport model including mode choice (and, in the case of NoRMS, destination choice): the PFM SCM has the structure shown below[3]:



The diagram should be read from the bottom up, starting with the rail assignment. When the demand model (for mode split and generation)[4] is reached, the direction changes to come down the left-hand side. Note that some of the detail (relating to the disaggregation of PLD zones to "mzones") is unnecessary to the discussion here.

Both models make some allowance for the choice of access/egress mode as well as station choice. The vexed question of catchment area is dealt with in different ways.

If we write x for the access mode and y for the egress mode, then the overall cost for the movement i-R-S-j can be described as:

$$C_{iRSj,xy} = Acc\mathrm{Cost}_{\mathrm{iR,x}} + \bar{C}_{RS} + \mathrm{E}gr\mathrm{Cost}_{Sj,y} \qquad (1)$$

and the choice of R, S, x and y can be based on the relative costs of the available options.

[3] The figure is taken from Figure 5-3 of HS2 Ltd (2022). "PLD" refers to the Planet Long Distance model, which is the main modelling component of the PFM

[4] Note that PFM, unlike NoRMS, contains no destination choice (distribution) component

There are various options for structuring the choice model over these four choices. The choice of station can be carried out independently at the two ends, or simultaneously {RS}: the access/egress mode choices {xy} can be at the same level or at a higher or lower level in a nested model.

In HS2's PFM, there are two separate SCMs, one for choice of station pair [RS] for non-London residents (where, for car available travellers, there is also a choice of access mode between car and a generic public transport mode: no account is taken of egress mode choice) and one for station choice within London, where neither access nor egress mode choice was taken account of. The non-London model had a nested structure with the RS choice at the bottom with a scale parameter (on GJT) of about –0.06, and the choice of access mode having a structural parameter θ in the range (0.25, 0.62) varying with purpose.

Probably the most complete estimation of an SCM is that carried out for the Dutch National Model, where specific data (similar to NRTS) was available. Again, a nested structure was used, with access/egress mode choice above station pair choice. Values for the scale parameter for train IVT were in the range –0.03 to –0.07, and structural parameters in the range 0.6 to 0.8. In this model, station choice depends not only on access, egress and on train costs but also on station characteristics such as the number of departing trains and parking arrangements for cars and bicycles. Considerable tests were carried out on potential catchment areas.

All these models can be considered to have the following general structure for the probability of station choice [RS] and the associated access/egress modes [xy], given zone-to-zone demand [ij]:

$$\Pr[RS,xy|ij] = \Pr[RS|xy,ij] *.\Pr[xy|ij] \qquad (2)$$

They are applied to estimates Q_{ij} of rail demand (segmented by purpose and other characteristics) between zones i and j. In principle, therefore, this kind of model could be applied to independently derived estimates of rail demand (eg from "gravity models").

For such discrete choice SCM models, we are back in the general problem of "new modes" if we introduce new stations. I am not aware of what is done in the Dutch model, but both UK models described above adopt the position of a "pseudo-DM", currently defined as follows:

[PFM]: new stations will, of course, not have any direct services associated with them in the DM, but it must nonetheless still be possible to reach the desired destination, so they need to be connected into the DM rail network. In cases where the new station is close to an existing station, this could be done by means of a walk link, but in other cases a transit link will be required. These links should be realistic rather than merely notional. It is expected that the proportion of passengers allocated to these 'new stations' in the DM will be very small. [*HS2 Ltd (2022)*]

[NoRMS[5]]: To make the new station available as an option in the Do-Minimum (even though it is not open) we connect a walk link to the nearest station so that it is

[5] This is based on unpublished material [*NoRMS Model Development Report T3a It2e*]

possible to calculate the generalised cost of making a rail journey via the new station.

This approach has the advantage of making the pseudo-DM realistic in terms of how one might access the rail network having gone to the (non-existent) station. Unfortunately, this does not guarantee that the DM demand for the new station will be negligible, and in particular it has been found that when the new station is close to an existing station it may be allocated too much use in the DM.

An alternative is to follow the TUBA Manual Guidance, and to choose the DM link from the new station to the rest of the existing network such that the resulting demand at the new station is 2% of that in the DS. For a single new station, this is probably feasible, though it would involve considerable iteration, given the different sources of demand contributing to the new station. It would also, of course, depend on the specification of the DS, which is not very convenient.

A further possibility which is worth considering is to make the rail **route choice** (ie RS, **given** access/egress mode) deterministic. In other words, for each possible {xy} combination, given ij, choose RS|xy,ij so as to minimise $C_{iRSj,xy}$ over all possible/relevant {RS} pairs.The choice of access/egress modes (which, as noted, has been found to be more random than that for the station pair) could then be made by means of a logit model. This would remove the need for a pseudo-DM cost since, unlike the case of the logit model, for a deterministic choice the benefit can be taken as the difference in (generalised) cost between the stations chosen in the DM and DS. It must be acknowledged, however, that such a choice is somewhat at odds with the actual evidence on station choice from PFM and the Dutch National Model, where the scale parameters were only slightly larger than those recommended in TAG for public transport destination choice (and of the same order as those for the car mode).

4 Recommendations

Some compromise has to be made in the case of new stations, and whatever is proposed will need to be tested. In my current view, the deterministic approach to station pair choice, given access/egress modes, offers the best solution. This would also sidestep some of the issues relating to catchment areas.

Current rail network models (in which we can include DfT's MOIRA) provide estimates of best route between each station pair RS. In principle, we could connect every zone to every station by available (non-rail) modes[6]. For each combination (xy) of access and egress modes, given a zone-to-zone movement (ij), we can find the station pair combination $(RS)^*_{xy,ij}$ which minimises $C_{iRSj,xy} = Acc\text{Cost}_{\text{iR,x}} + \bar{C}_{RS} + \text{E}gr\text{Cost}_{Sj,y}$: this is a straightforward network problem which merely needs to repeated for each (xy) * (ij) combination.

In practice, most of the possible station pairs will be irrelevant. If we begin by taking the nearest[7] station at each end (thus, R* and S*), and calculate the total cost $C_{iR^*S^*j,xy}$, any station R or S where the access or egress cost already exceeds this total cost can clearly be

[6] as discussed below, practical rules could be designed to avoid this
[7] this could be defined either in distance or in time terms

ignored, so that rules could be developed as to how many stations actually need to be connected to each zone. Unlike in the non-deterministic case, the results are not sensitive to arbitrary decisions about catchment areas and the number of stations to include.

In considering this, it would be useful to have some idea of how many different stations are indeed used from any particular zone (obviously the zone size will be relevant here). A small exercise using the NRTS (or its successor, if available) would be helpful here.

Given the station choice (RS)* for each (xy, ij), the choice of access/egress mode combination could be obtained from a standard logit model. The logsum from such a model could then be used to influence changes in demand, whether due to destination choice, mode choice, or "generation". Within a pure rail model, the "generation" effect would subsume other types of choice, such as shifting from other modes, or changing destinations.

The cost formula in Eq (1) is the most straightforward case, where access and egress costs are added to the station-to-station rail costs. More generally, we could define

$$GC_{RSxy|ij} = f\left(Acc\mathrm{Cost}_{\mathrm{iR,x}}, \bar{C}_{RS}, \mathrm{E}gr\mathrm{Cost}_{Sj,y}\right) \qquad (3)$$

and, as before, for each {ij, xy} find (RS)* which minimises $GC_{RSxy|ij} \to GC_{**xy|ij}$, In principle, other variables relating to stations (eg classification, parking constraints/availability) could be introduced.

We then use a logit model to choose the {xy} probability for each {ij}:

$$p_{xy|ij} = \frac{\exp(-\lambda GC_{**xy|ij})}{\sum_{x'y'} \exp(-\lambda GC_{**x'y'|ij})}$$ giving composite GC

$$GC_{****|ij} = -\frac{1}{\lambda} \ln \sum_{x'y'} \exp(-\lambda GC_{**x'y'|ij}) \qquad (4)$$

for the rail demand model.

Note that a potential problem is that this composite cost cannot be decomposed – in particular to Fare and GJT. However, it is possible to approximate the **change** in the components, as noted in Bates (2015), inasfar as the components are additive. So, for example, if we have the simplest form as given in Eq (1) above:

$$\Delta GC_{****|ij} \approx ½ \sum_{x'y'} \left[\left(p'_{x'y'} + p_{x'y'} \right) . \left(\Delta Acc\mathrm{Cost}_{\mathrm{iR*,x'}} + \Delta \bar{C}_{R*S*} + \Delta \mathrm{E}gr\mathrm{Cost}_{S*j,y'} \right) \right] \qquad (5)$$

It is important to note that R* and S* are potentially different between the DM and DS.

In the base year, we require as good as possible an estimation of the rail matrix at the chosen zone level. Note that this would be a useful asset if it could be made available at a national level. For convenience, we will assume the model operates at the MSOA level. NTEM provides a TAG-recommended resource for providing zonal trip ends (Production and Attractions for home-based purposes, and Origins and Destinations for non-home-based). From these, using some appropriate measure of separation (distance, or cost) a first estimate of a matrix can be generated: it may be possible to use output from the National Travel Model [NTM].

NTEM also provides rail mode shares by urban area type, though these are geographically coarse, and do not take account of the actual supply of rail services. So while these shares should be achieved at the aggregate urban area type level, they can be expected to diverge significantly for particular zones. Here, the LENNON matrix can be very useful, using a variant on Matrix Estimation [ME]. It is understood that DfT's TAG unit M3.2 is currently being updated with advice on how this might be done.

ME is normally associated with the highway, where the object is typically to modify a prior matrix t_{ij} to be consistent with "observed" vehicle flows. This can be briefly described as follows:

Suppose we have a set of counts V_a which may be written as $V_a = \sum_{ij} T_{ij}\, p_{ija}$ where p_{ija} is the probability that a movement from i to j will use link a. Given certain assumptions, the solution for the most likely matrix can be written $T_{ij} = t_{ij} \prod_a X_a^{p_{ija}}$ where {X} is a set of constants to be determined (related to the Lagrangean multipliers on the constraints for each count). By analogy, we can aim to modify the implied zonal matrix of rail travel to be consistent with the LENNON matrix, using the station choice probabilities $p_{RS|ij}$ implied by the SCM (summing over access/egress modes and purpose), with the aim of ensuring that $M_{RS} = \sum_{ij} T_{ij}\, p_{RS|ij}$, where M_{RS} is the LENNON station-to-station matrix. Some work along these lines has been carried out in connection with the NoRMS model, referred to as "MOIRA matching".

The general TAG advice for using ME in the highway case stresses that the changes brought about to the "prior" matrix should not be "significant". However, such advice may be questioned in the rail case as just described, since the LENNON matrix is likely to be far more reliable than any "prior" zone-to-zone matrix which could be developed for rail.

While it should always be borne in mind that introducing new alternatives presents challenges both for modelling and appraisal, the above proposals for dealing with new stations appear to offer a practical way forward. A number of tests will need to be done, and it is recommended that further analysis of NRTS (or its successor, if available) should be carried out with a view to a better understanding of the stations actually used from any particular zone and the potential error in assuming a deterministic choice of station pair, given access and egress costs.

It will also be useful to crystallise the best definition of rail generalised cost including access and egress, as it appears that a number of variants are currently in use.

Finally, appropriate software will need to be set up to implement the various procedures, including those for "MOIRA matching", and the general recommendations will need to be tested against existing implementations.

New rail opportunities

The above discussion was specifically in the context of station choice. However, it may be objected that in some cases, even without new stations, an existing journey by rail may be so unlikely that a new connection has the characteristics of a "new mode". Consider, for example, the journey between Bicester Village and Bletchley, a distance of about 22 miles which could be served by the proposed East-West Rail route and take around 25 minutes. Currently, the journey involves travelling into London Marylebone, taking the Underground to Euston, and then travelling out to Bletchley, taking about two and a half hours (with an off-peak journey costing £66), whereas a bus journey would take just over an hour and cost £9. Clearly, no-one would currently make this journey by rail. It is also likely that much of the new rail demand will come from existing modes.

In such a case, we require a reliable mode choice model to allocate the potential overall demand to individual modes, for both the DM and the DS. Such a model would need to have been estimated so as to provide plausible DM choices for journeys in cases when all specified modes are available, and in doing so, deal appropriately with possible correlations between the error terms for different modes, as well as the possibility of ASCs. Provided this is done, we should be able to have confidence in the DS rail forecast.

It will, of course, be the case that the benefit to (new) rail travellers is high: however, although when using the standard RoH this would be half the change in generalised cost between the (unlikely) DM and the DS, this will be substantially reduced by the (recommended) use of NI. While this may **seem** unreasonable when considering the generalised cost for existing modes, the only way to avoid this is to assume that the choice between rail and at least one of the other modes is purely deterministic.

Note that the PDFH-based rail elasticities for GJT and Fare are agnostic as to the source of the changed demand (and in particular how much might be transfer from other modes, how much transfer from other rail journeys, and how much from "pure" generation – increased frequency of travel). While the balance is likely to vary according to the generalised cost of other modes, it would be helpful to increase the understanding of the likely amount of transfer from other modes, as this could help to validate mode choice models where required.

5 Suggestions for further work

The key proposals emanating from this review may be summarised as follows:

- Develop station choice models, using 2004 NRTS and its possible successor surveys
- Investigate the range of stations actually used for given zone pairs (ij), using same source material, and assess impact of the compromise proposal to use deterministic station pair choice, given access/egress modes.
- Develop a national rail matrix at a **zonal** level, using NTEM methodology, with Matrix Estimation [ME] applied to LENNON matrices.
- Crystallise the best definition of rail generalised cost including access and egress
- Investigate the possibilities for determining the source of the changed demand implied by PDFH-based rail elasticities (in particular how much might be transfer from other modes, how much transfer from other rail journeys, and how much from "pure" generation – increased frequency of travel).

References

Bates J J (1991), Introducing New Alternatives to a Choice Model, Traffic Engineering and Control, Vol 32 no. 19, pp 468-471

Bates J J (1992), Developing Choice Models to Deal with New Transport Alternatives, 6th World Conference on Transport Research, Lyons, France

Bates J J (2015), New Issues in Transport Appraisal, paper presented to 14th International Conference on Travel Behavior Research (IATBR), Windsor, UK

Department for Transport (2020), TUBA: General Guidance and Advice Version 1.9.14 August 2020,

HS2 Ltd (2022), Planet Framework Model version 10a: model description report, available at https://www.gov.uk/government/publications/hs2-planet-framework-model-version-10a-model-description

de Jong, G, Daly, A J, Pieters, M & van der Hoorn, A I J M (2005), The logsum as an evaluation measure: review of the literature and new results, Transportation Research A (41), pp 878-889

Nellthorp J & Mackie P J (2001), Alternatives to the rule of a half in matrix-based appraisal. Institute for Transport Studies, University of Leeds, Final report to DETR.

Ojeda-Cabral M, Batley R and Johnson D (2021), Rail Openings Appraisal: Review and development of appraisal practice for new railway lines, stations and services. Final Report (v2.0), Institute for Transport Studies, University of Leeds, November 2021

Pearce D W and Nash C A (1981), The Social Appraisal of Projects - A Text in Cost-Benefit Analysis, MacMillan.

Rohr C, Fox J, Daly A, Patruni B, Patil S & Tsang F (2010), Modelling Long-Distance Travel in the UK, Association for European Transport

Wardman M, Lythgoe W & Whelan G (2007), Rail Passenger Demand Forecasting: Cross-Sectional Models Revisited, Railroad Economics, Research in Transportation Economics, Volume 20, 119–152

Appendix A: Mathematics for Integrals of logistic distribution *f*(δ)

With a choice between two modes 1 and 2 based on the logit model, the basic specification is:

$$U_i = V_i + \varepsilon_i = -\lambda.G_i + \varepsilon_i$$

where G_i is generalised cost (which could in principle include ASCs, α_i), and ε is a random term with the extreme value distribution. Mode 1 will be chosen if $U_1 > U_2$ so that $V_1 + \varepsilon_1 > V_2 + \varepsilon_2$.

Because the difference of two Gumbel terms has the logistic distribution, writing $\delta = \varepsilon_2 - \varepsilon_1$, this implies that $\delta < V_1 - V_2$. Hence, the probability of choosing mode 1 = $F(V_1 – V_2)$, where F is the cumulative distribution function.

The standardised logistic distribution, with mean zero and standard deviation $\pi/\sqrt{3}$, has the density function $f(\delta) = \dfrac{\exp(-\delta)}{\left(1+\exp(-\delta)\right)^2} = ¼\,\mathrm{sech}^2(½\delta)$ and the cumulative distribution function is $F(\delta) = \dfrac{1}{1+\exp(-\delta)} = ½ + ½\tanh(½\delta)$. Hence, the proportion p_1 choosing 1 is given by

$$p_1 = \int_{-\infty}^{V_1-V_2} f(\delta).d\delta = F(V_1 - V_2) = ½ + ½\tanh(½(V_1 - V_2)) \qquad \text{(A1)}$$

Now consider a transport improvement whereby the cost of mode 2 reduces from C^0_2 to C'_2. There is no reason to expect any changes in the average mode-specific utilities α_1, α_2. As a result of the change $V_2 \rightarrow V'_2 > V^0_2$.

This therefore affects the proportion choosing mode 1, which becomes $p'_1 < p^0_1$.

The total population can now be segmented into three groups:

a) those who remain with mode 1 (p'_1). They receive no benefit

b) those who were on mode 2 anyway (p^0_2): they receive the full benefit of the improvement [= $V'_2 – V^0_2 = \lambda\,(C^0_2 – C'_2)$]

c) those who switch from mode 1 to 2

Since group c) was previously on mode 1, it must be the case that for them

$$\delta < V_1 – V^0_2$$

Since they are now on mode 2, it must also be the case that

$$\delta > V_1 – V'_2.$$

Hence $V_1 – V'_2 < \delta < V_1 – V^0_2$ (A2)

The utility gain that they obtain from the switch is given by

$$U_{12} = (V'_2 + \varepsilon_2) - (V_1 + \varepsilon_1) = V'_2 - V_1 + \delta \qquad \text{(A3)}$$

(Note again that by definition, the difference between V'_2 and V_1 includes not merely the difference in travel costs but also the difference in the average mode-specific disutilities.)

Hence, **if** $\delta \equiv 0$ (ie no random effects), then the benefit of transferring from period 1 to period 2 is given by

$$U_{12} = -\lambda(C'_2 - C_1) + (\alpha_2 - \alpha_1) \qquad \text{(A4)}$$

in other words, it depends on the change in cost experienced **and** the change in mode-specific utility.

However, as soon as we allow for random effects, we must integrate over the relevant range of δ to give the average $\bar{U}_{12}$:

$$\bar{U}_{12} = V'_2 - V_1 + \frac{\int_{V_1 - V'_2}^{V_1 - V^0{}_2} \delta . f(\delta)\, d\delta}{\int_{V_1 - V'_2}^{V_1 - V^0{}_2} f(\delta)\, d\delta} \qquad \text{(A5)}$$

As noted above, $f(\delta) = \dfrac{\exp(-\delta)}{\left(1 + \exp(-\delta)\right)^2} = ¼\,\text{sech}^2(½\delta)$

For the denominator of the integral quotient in Eq (A5) we need the integral of $f(\delta)$. We can evaluate this since it is known that $\int \text{sech}^2(x)\, dx = \tanh(x) + C$.

To integrate $f(\delta) = ¼\,\text{sech}^2(½\delta)$ we change the variable of integration, putting z= ½ x, dx = 2 dz so

$$¼\int \text{sech}^2(½x)\, dx = ½\int \text{sech}^2(z)\, dz = ½\tanh(½x) + C \qquad \text{(A6)}$$

For the numerator of the integral quotient in Eq (A5) we need the integral of $\delta . f(\delta)$.We can integrate $\int x.\text{sech}^2(x)\, dx$ by parts, using the formula $\int u.\frac{dv}{dx}.dx = uv - \int v.\frac{du}{dx}.dx$. We set u = x, v = tanh(x), since – as before - $\int \text{sech}^2(x)\, dx = \tanh(x) + C$.

Hence the integral becomes $\int x.\text{sech}^2(x)\, dx = x.\tanh(x) - \int \tanh(x)\, dx$ and it is known that $\int \tanh(x)\, dx = \ln(\cosh(x) + C$.

Again we can change the variable of integration putting z= ½ x, dx = 2 dz , so that

$$¼\int x.\text{sech}^2(½x)\, dx = ¼\int 2z.\text{sech}^2(z).2dz = \int z.\text{sech}^2(z).dz$$

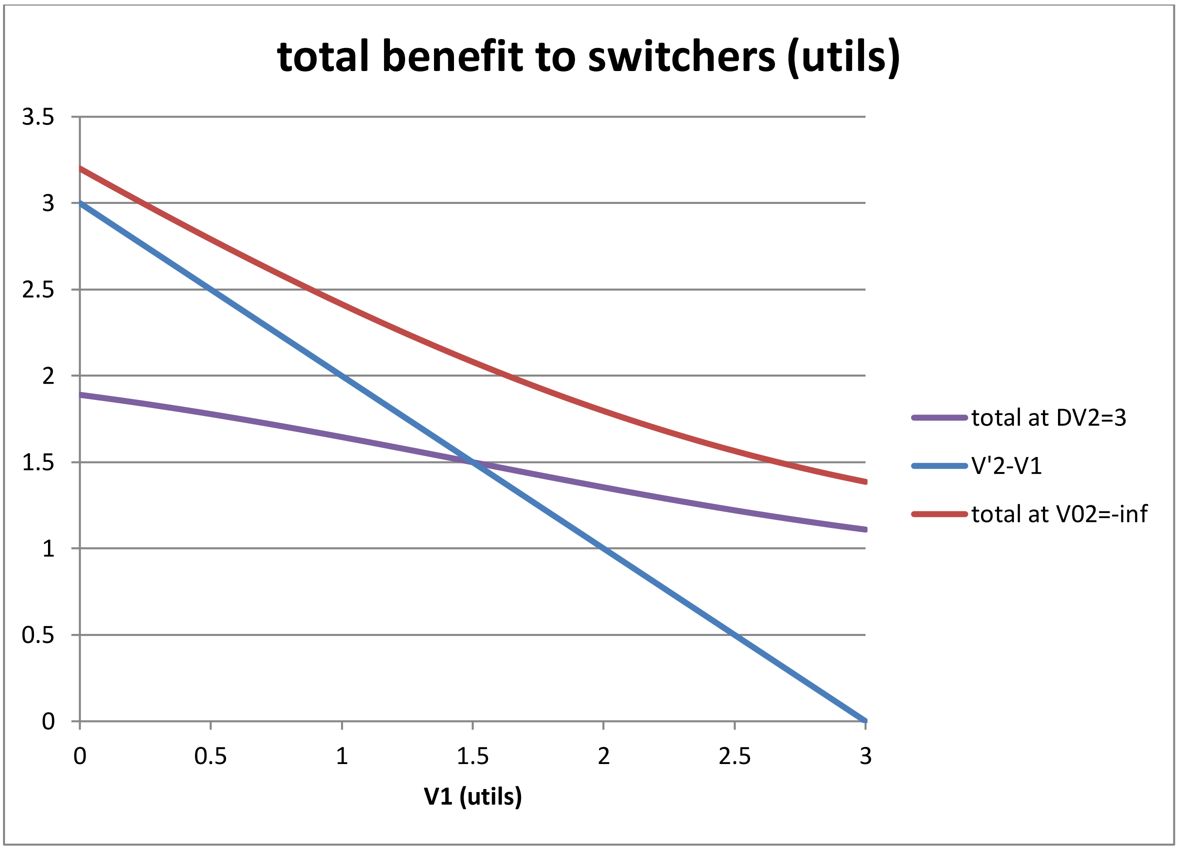
$$= z.\tanh(z) - \ln(\cosh(z) + C = ½x.\tanh(½x) - \ln(\cosh(½x) + C \qquad \text{(A7)}$$

Putting all this together, $\bar{U}_{12} = V_2' - V_1 + \dfrac{\left[½x.\tanh(½x) - \ln(\cosh(½x))\right]_{V_1 - V_2'}^{V_1 - V^0_2}}{\left[½x.\tanh(½x)\right]_{V_1 - V_2'}^{V_1 - V^0_2}}$ (A8)

When $V'_2 - V^0_2$ is small, $\bar{U}_{12}$ is approximately equal to ½ $(V'_2 - V^0_2)$ and is independent of V_1 , but as it increases, V_1 plays more of a role in the overall benefit. The benefit is not, however, equal to $V'_2 - V_1$ as long as there remains any random term.

Suppose $V'_2 - V^0_2 = 3$, which, given the scale of the error term, is a significant improvement. Set $V^0_2 = 0$, arbitrarily (since only differences in utility are relevant). The figure below shows how $\bar{U}_{12}$ varies with V_1 over the range [0,3] when $\Delta V = 3$.

In fact, for the binary case it can be shown that as $V^0_2 \to -\infty$ (equivalent to the "new modes" case), the numerator of the integral quotient in Eq (A5) tends to $\ln(2) - \left(½\left(V_1 - V_2'\right).\tanh(½\left(V_1 - V_2'\right)) - \ln(\cosh(½\left(V_1 - V_2'\right))\right)$, while the denominator tends to p'_2. In this case, therefore, it appears to be possible to estimate the benefit when there is a new mode, as a function of the change in GC between the car and the new rail option. It is not, however, equal to $V'_2 - V_1$, as is shown in the Figure.



Appendix B: Key observations from Bates (2015)[8]

Consumer Surplus continues to be the mainstay of transport appraisal, and while theoretically "correct" measures such as the "logsum" are often advocated (e.g. de Jong et al, 2005), they are rarely used in practice. A central reason for this appears to be the fact that, unlike the well-known "Rule of a half" [RoH] approximation, the logsum is not decomposable into constituent elements of "utility" (or "generalised cost"), and hence conveys less information than is considered desirable. A further reason for the prevailing preference for the RoH is that agencies may wish to allocate different weights to the elements of generalised cost from those that are being used in the demand model, even though this may give rise to inconsistencies (as pointed out, e.g., by Pearce & Nash, 1981).

The case of new options in the choice set is a perennial problem. Provided that the choice process is "reasonably random", the addition of a new option – even if no better than the worst existing option – is likely to convey considerable benefit as measured by the "logsum", and this is sometimes referred to as "diversity benefit". However, while this term is in principle well motivated, it turns out to be difficult to quantify it reliably. In addition, there are (bad) examples from practice where the concept has been seriously misused, leading to potentially erroneous conclusions.

…….

It is important to recognise that there is **no theoretically justified** way in which the composite cost C_* can be represented as the sum of ("appropriately composited") components:

$$C_* = \sum_k \beta_k . C_*^k \qquad (12) \qquad \text{****invalid****}$$

This is because the composite cost depends only on the (weighted) sum of the individual components for each alternative, and not the individual components independently. Failure to recognise this has led to significant errors in practical applications. A common assumption has been to carry out the following (wrong!) calculations for the components:

$$\tilde{C}_*^k = \sum_r p_r . C_r^k$$

where r ranges over the choices being modelled, and p_r denotes the predicted share. Since, as is well known, weighting these (share-based average) values for the components by the component weights β_k and adding over k will **not** deliver C_* (except in the case where $\lambda = \infty$, i.e. a deterministic choice model), a residual term is calculated as

$$C_* - \sum_k \beta_k . \tilde{C}_*^k$$

and variously referred to as an additional benefit component associated with "choice" or "diversity". On this basis, it might seem that the calculations for the change in the individual components can be made.

It is easy to show, however, that these calculations are not valid ….. The reason is that, while the decomposition of the absolute composite cost is illegitimate, there is a legitimate approximation, under certain conditions, for the **change** in composite cost, along the following lines.

[8] Equation numbers in this Appendix are as in Bates (2015)

By a fundamental principle of the differential calculus, where y = f(x), the change in y, Δy, resulting from a change of Δx in x can be approximated by:

$$\Delta y \approx \frac{dy}{dx}.\Delta x = f'(x).\Delta x$$

This extends straightforwardly to the case where y is a function of multiple x-arguments. Hence, we can write

$$\Delta C^* \approx \nabla C^*.\Delta \mathbf{C} \qquad (13)$$

where ∇C^* is the vector of partial derivatives of C* with respect to the costs of the elemental alternatives r, and $\Delta \mathbf{C}$ is the vector of cost changes of the elemental alternatives r. In other words,

$$\Delta C^* \approx \sum_r \left(\frac{\partial C^*}{\partial C_r}.\Delta C_r \right) \qquad (14)$$

Now it can be shown that the vector ∇C^* is precisely equal to the vector of choice probabilities **p**: this applies not merely to a single level logit model but a hierarchical (nested) model as well. Hence

$$\Delta C^* \approx \mathbf{p}.\Delta \mathbf{C} \qquad (15)$$

Since we expect the changes in cost to have some effect on the choice probabilities, a better version of the approximation can be obtained by evaluating ∇C^* at both the before and after situations and taking the average. In this way, we obtain another version of the RoH, whereby

$$\Delta C^* \approx ½\sum_r \left(p'_r + p_r\right).\Delta C_r \qquad (16)$$

and since the elemental costs **can** be decomposed into their components k, we have:

$$\Delta C^* \approx ½\sum_r \left[\left(p'_r + p_r\right).\left(\sum_k \beta_k.\Delta C_r^k \right) \right] \quad (17)$$

where ΔC_r^k is the change in component k between the base and policy cases for alternative r. Note crucially that this derivation requires the choice set {r} to be the same in both the before and after situations.

In general, this approximation is extremely robust, and can in principle take into account any component deemed to form part of the cost C_r. The changes in the individual components calculated (correctly) by this method can be quite different from those using the fallacious values of $\tilde{C}_*^k$.

This makes the approximation for ΔC* rather convenient (provided, of course, the conditions on its validity are met).

…….

The implications are thus that the RoH should generally be applied at the level of the elemental alternatives, where the components breakdown is available. There is one exception to this – when the choice between elemental alternatives is deterministic. This is the normal case with highway route choice, and may also arise with alternative routes through the rail network between any given pair of stations (depending on the assumptions). In cases where multiple routes are predicted to be used under a deterministic choice allocation, it is legitimate to calculate the average components according to the route proportions, as noted earlier.

The reason why the RoH cannot be used when the choice between elemental alternatives is deterministic is that the costs of rejected routes are irrelevant to the benefit calculation. In a deterministic process, only routes with the minimum cost will be used, and the benefit is equivalent to the change in the minimum cost. The fact that multiple routes may have the same minimum cost is merely a minor complication from the point of view of calculating the components.

…….

We noted earlier that from a theoretical point of view, the consumer surplus can be directly calculated by integrating along the demand curve, and that for logit-based functions this gives the composite cost, **independently** of whether the choice sets change (although there are further issues relating to this, as we discuss below). From the point of view of the WebTAG guidance, however, this "locks in" the demand model weights, which may not be the same as the appraisal weights and also fails to deal with the disaggregation by components. In addition, the implied reduction in costs may be difficult to justify (at least to a non-specialist audience), as we discuss below.

Overall, therefore, from the point of view of UK guidance the position can be set out as in the following table:

	choice set constant between before and after scenarios	**choice set differs between before and after scenarios**
appraisal weights same as demand model weights	Composite cost correct but approximate components can also be calculated using RoH at level of elemental alternatives, or at next higher level in the case of a deterministic choice process	Composite cost correct but no reliable breakdown into components possible
appraisal weights different from demand model weights	Composite cost not correct but new weights can be applied to approximate components using RoH at level of elemental alternatives, or at next higher level in the case of a deterministic choice process	no reliable solution

In the worst case, the combination of changing choice sets and the need to use a different set of appraisal weights from those used for the demand model places us in the bottom right-hand box – "no reliable solution". In order to make progress, one of the conditions has to be relaxed. Keeping the choice set constant allows progress to be made in line with the WebTAG requirements, but as we will see later, this is not without its problems.

…….

A particular issue about the use of the composite cost in these circumstances, and the need to rely on "diversity benefit", is that the implied change in benefit may simply not be credible. For example, with a value of $\lambda = 0.01$ (which, on the basis of work for the UK Long Distance Model – see Rohr *et al* 2010 for a general description – is a reasonable value for the choice between "classic" rail and high-speed rail when C is measured in minutes for a one-way trip), adding a high speed rail service where the generalised cost was no better than the existing rail service would add a (diversity) benefit of 141 minutes. While this result is entirely consistent with the model, it tends to cast doubt on its appropriateness, particularly given the problems of how to introduce the new alternative.

Other more subtle issues may arise when the choice set changes: two examples associated with the rail mode may be noted. These examples have been recently encountered in practical applications reviewed by the author.

In the context of station choice, it is clear that the majority of station pairs will not be relevant for a particular zone-to-zone movement, as they are essentially "dominated" by other less expensive alternatives. Given the large number of potential station pairs, there is thus a natural desire to establish some "rules" for limiting the choice set, in order to reduce the level of computation. However, if these rules are formulated in a way which may be altered by policy (for example, all stations within 20 minutes driving time from the residence), then unexpected, and possibly unsuitable, changes in composite cost may result from changes in the choice set. It is also difficult to specify a set of rules which will be appropriate for all possible policies.
…….
For most transport problems, the RoH remains a highly accurate approximation to the true CS, and has the perceived advantages of a) being decomposable into the different components of generalized cost and, thereby, b) allowing for different weights to be used to achieve consistency of appraisal over projects

When large cost changes are involved, the accuracy can decrease markedly, but with some effort this can be rescued by a piece-wise linearization of the demand function between the before and after positions. The problem can be diagnosed by comparing the change in composite cost with the RoH, using the generalized cost components weights as in the demand model.

Much greater problems are encountered in circumstances where there are changes in the choice set. While the change in composite cost from a logit model remains a theoretically valid measure when the choice set changes, there are concerns about its appropriateness, though these are essentially related to the modelling difficulties associated with introducing new alternatives.

The alternative approach, advocated by the TUBA Guidance, is to create a "pseudo Do Minimum" case, in which the new alternatives are introduced but with high costs. Again, problems are encountered, and there appears to be insufficient practical experience of the approach.

It was noted that in addition to the issue of explicit new alternatives, there were other circumstances where the choice set might change as a result of introducing "rules", usually to keep the number of alternatives to manageable levels. Such cases warrant careful examination, to ensure that the implied benefits were legitimate.