Original Article

Spatial maps in voting advice applications: The case for dynamic scale validation

Micha Germann[a,b,*], Fernando Mendez[a], Jonathan Wheatley[a] and Uwe Serdült[a]

[a]Centre for Research on Direct Democracy (c2d), Centre for Democracy Studies Aarau (ZDA), University of Zurich, Switzerland

[b]Centre for Comparative and International Studies, ETH Zurich, Switzerland

*Corresponding author.

Abstract Low-dimensional spatial representations of political preferences are a widespread feature of voting advice applications (VAAs). Currently, VAA spatial maps tend to be defined on the basis of *a priori* reasoning. This article argues that VAA spatial maps should be empirically validated to safeguard fundamental psychometric properties – in particular, unidimensionality and reliability. We suggest dynamic scale validation as a pragmatic method for improving measurement quality in VAA spatial maps. The basic logic of dynamic scale validation is to exploit early user data as a benchmark against which *ex-ante* defined maps can be evaluated. We draw on data from one of the most institutionalised VAA settings, Switzerland, to illustrate this dynamic approach to scale validation.

Acta Politica (2015) **50,** 214–238. doi:10.1057/ap.2014.3;
published online 28 February 2014

Keywords: voting advice applications; ideological space; dynamic scale validation; unidimensionality; reliability

Introduction

Voters in search of political information before elections increasingly turn to so-called voting advice applications (VAAs). Essentially, VAAs are internet tools that provide ‘voting advice’ by establishing ideological congruence with parties or candidates. The basic mode of functioning of a VAA is relatively straightforward. Visitors to the website indicate their preferences on a series of policy items, and upon completion, the VAA calculates the match with previously coded preferences of political parties or individual candidates. However, matching voters to political actors is simple only in the abstract. Recent work highlights that various aspects related to VAA design affect the quality of the ideological congruence estimates, including item wording (Gemenis, 2013), item selection (Walgrave *et al*, 2009), the matching algorithm (Mendez, 2012) and the rating scale format (Baka *et al*, 2012).

We contribute to the emerging literature on VAA design by extending the focus to a so far neglected aspect: the low-dimensional spatial visualization of policy preferences. Typically consisting of two dimensions, but in some cases more, the basic function of VAA spatial maps is to give users an indication of both their own and the political elites' standing in the political space. By this, they provide an implicit form of voting advice in the form of a relational cue. Most VAAs draw on spatial visualization techniques. Usually, low-dimensional representations accompany alternative, high-dimensional techniques. This is the case, among others, for the Swiss *smartvote*, the *Preference Matcher* family and the European-wide VAA *EU Profiler*. However, some prominent examples, such as the Dutch *Kieskompas*, even rely exclusively on low-dimensional spatial representations (Louwerse and Rosema, 2011).

Recent evaluations of construct validity suggest that the political dimension underlying VAA spatial maps did not always meet basic measurement criteria (Louwerse and Otjes, 2012; Gemenis, 2013). VAA designers should not take this finding lightly. Psychometric viability is essential for high-quality matching in a spatial framework. In this article, we are concerned with two of the arguably most fundamental properties needed for meaningful measurement: unidimensionality and reliability. We argue that the fundamental reason for the existing deficiencies is the current practice of basing VAA spatial maps on pure *a priori* reasoning. To maintain high-quality spatial matching, the political value dimensions need to be empirically validated. We suggest dynamic scale validation as a pragmatic method for safeguarding psychometric utility in VAA spatial maps. The basic logic of this dynamic approach to scale validation is to exploit data generated by actual VAA users who access the site soon after its launch in order to review and if necessary amend *ex-ante* defined scales. The potential of dynamic scale validation is underlined by means of a real-world example: the Swiss *smartvote* deployed before the federal elections of 2007.

The article proceeds as follows. In the first section we outline why the concepts of unidimensionality and reliability are crucial for VAA spatial maps. After this, we contrast various alternatives for validating spatial maps and make the case for dynamic scale validation. Next we discuss the specifics of dynamic scale validation, in particular methods suited for an evaluation of unidimensionality and reliability. In the fourth section, we apply our argument empirically. Drawing on the same empirical example, the final section addresses practical implications of dynamic scale validation. The conclusion reviews the main arguments.

Why Should We Care? Unidimensionality and Reliability in VAA Spatial Maps

Notions of space lie at the heart of political discourse (Benoit and Laver, 2012). Statements like 'party A is more conservative than party B' or 'C is more left than D' are prevalent not only among political scientists, but also among ordinary citizens.

By this, we are – wittingly or unwittingly – employing the language of space. From this perspective, it makes a lot of sense that VAAs make use of spatial forms of representing political preferences.

However, even if deeply ingrained in our political thinking, spatial concepts like left versus right fundamentally remain spatial *metaphors* (Benoit and Laver, 2012). Unlike with real space, it is impossible to physically and objectively observe an individual's position in the ideological space. We can only estimate it. VAAs do this by aggregating user and party/candidate ratings to a series of policy statements. For instance, answers to statements like 'income taxes should be increased' or 'there should be universal health care' are summed in order to get an estimate of an individual's position on the economic left–right dichotomy. The implicit idea behind this is that an individual's position on income tax and health care is caused by an underlying construct – economic left–right – and can therefore be used for measuring this construct.

The adequacy of a summated rating scale hinges on several criteria (Carmines and Zeller, 1979; Clark and Watson, 1995). In this article, we are concerned with the concepts of unidimensionality and reliability. Unidimensionality, on the one hand, requires that all items (the policy statements in our case) in a scale measure a single latent trait (a property called internal consistency), and only this trait (a property called external consistency; Gerbing and Anderson, 1988). Unidimensionality is the most critical assumption of measurement theory (Hattie, 1985). A lack of unidimensionality implies ambiguity in the composite score. For instance, a measure of economic left–right becomes pretty useless if it also measures liberal–conservative values, given that an individual's score does not exclusively reflect her position in terms of economic left versus right, but is influenced by culture-related attitudes as well.

Reliability, on the other hand, refers to the precision of a measurement instrument, or, technically speaking, to the share of error-free true score variance to total observed variance (Lord and Novick, 1968). The concept of reliability is equally fundamental to meaningful measurement (Carmines and Zeller, 1979); a perfectly valid, but very unreliable measure is similarly futile as a plainly invalid (for example, not unidimensional) measure because even if we get it right on average, we are most of the times far off the target (or, technically speaking, the true score).

Evidently, both concepts are essential to spatial mapping in VAAs. If the scales underlying VAA spatial maps lack unidimensionality and/or reliability, the positionings of both users and the political elite cannot be unambiguously interpreted and/or are too imprecise, giving rise to wrong conceptions on the side of the VAA user about her standing in the ideological space, about the position of the political elite and, by implication, the relative distances to the different political parties or candidates. In short, a lack of unidimensionality or reliability hampers the usefulness of VAA spatial maps, and in particular the quality of the implicit voting advice proffered to the user.

The Case for Dynamic Scale Validation

Recent evaluations suggest that VAA spatial maps did not always meet basic measurement criteria (Louwerse and Otjes, 2012; Gemenis, 2013). Scrutinizing the *EU Profiler*'s (the first pan-European VAA designed for the 2009 European elections) two-dimensional map, both Gemenis and Louwerse and Otjes concluded that its left–right scale was deficient in terms of unidimensionality. Taking a step back, the fundamental reason for the lack of unidimensionality is that the designers of the *EU Profiler* determined the nature of the spatial map exclusively on *a priori* grounds. That is, individual items were selected as indicators of the political dimensions underlying the map on the basis of a theoretical conception of the ideological space without subjecting the resulting scales to further scrutiny. The *EU Profiler* is by no means alone in this practice; to the best of our knowledge, almost every VAA draws purely on *ex-ante* reasoning for their spatial maps.[1] However, armchair theorizing has its limits; the extent to which required psychometric properties – such as unidimensionality and reliability – are fulfilled should always be empirically established (Cronbach and Meehl, 1955; Clark and Watson, 1995).

In short, the ubiquitous practice of defining political dimensions *ex-ante* does not guarantee that VAAs deliver meaningful voter-elite comparisons. Thus, there is a compelling case that VAA designers empirically validate the quality of the underlying political dimensions on which the spatial maps are based. The question is how they should go about scale validation.

Let us proceed from the ideal case scenario. Strictly speaking, matching voters to parties on the basis of summated rating scales presumes that the scales are equivalent across both voters and elites (Davidov, 2009). Most fundamentally, this means that both share the same understanding of ideology – that is, that the scales tap the same latent ideological dimension (and only this dimension) for both users and elites. This is not a trivial assumption; issues do not always relate the same way for voters and elites (Kriesi *et al*, 2006; van der Brug and van Spanje, 2009). If equivalence is not given, comparisons will be systematically biased and the estimates of ideological congruence will not reflect 'true' differences. Ideally we would thus analyse both voter- and elite-level data, and establish that ideological scales are sufficiently unidimensional and reliable in both samples (followed by further equivalence tests, see Horn and McArdle, 1992). However, the problem with the ideal case scenario is that most VAAs compare users to a mere 5–10 parties. The scarcity of data on the elite side in all but the most exceptional circumstances prevents us from testing the adequacy of the political dimensions on the supply side.[2]

In most contexts, the only viable alternative remaining is therefore voter-based validation. Obviously we have no *N*-problem on the voter side. The drawback is that voter-centred validation necessarily neglects the elite side, or in other words, it necessarily implies the superimposition of the voters' ideological space on the elite space. However, we believe voter-centred validation can be justified not only by necessity, but also on conceptual grounds. Superimposing the voter space implies that voters are matched to parties within a spatial framework as it is defined by voters. Arguably, this is very much in line with the main function of a VAA: giving *voters* an indication of which party or candidate best matches *their* preferences.[3]

Hence, demand side-based validation is not only a matter of necessity for the typical VAA, but is also conceptually viable. But what form should it take? Evidently, an ideal solution would involve the pre-administration of the VAA questionnaire to a representative test sample before the VAA launch. On the basis of this survey, we could conduct extensive psychometric testing, and thereby define the spatial map. However, such a survey would require very substantial financial resources, which we suspect renders this approach impracticable for most VAA designers.

Given the usually tight budgets, we suggest *dynamic scale validation* as a pragmatic alternative to conducting a test survey. The basic logic of dynamic scale validation is to exploit data generated by actual VAA users who access the site soon after its launch. If the analysis of early user data suggests that the *ex-ante* defined scales meet psychometric standards (in particular, are both unidimensional and reliable), all the better. However, if it turns out that measurement quality could be improved by changing the composition of the scales, the spatial map can be adjusted early on in the launch phase. This will improve measurement quality over the remaining time the VAA is online.

Critically, our suggestion of dynamic scale validation rests on two basic assumptions. First, contra Converse (1964), we must assume that voters exhibit a reasonable degree of ideological constraint. If too few voters have consistently left, right or centrist policy positions, it is evidently impossible to construct a meaningful left–right dimension. Note that the issue constraint argument is directed not only against dynamic validation. If Converse were indeed right, we should abandon the idea of spatial mapping altogether: Converse's argument would imply that it is altogether impossible to construct a (unidimensional scale based on voter data). However, against this contention, a significant body of research has shown that Converse's view is overly pessimistic (for example, Achen, 1975; Ansolabehere *et al*, 2008; van der Brug and van Spanje, 2009; Germann *et al*, 2012; Wheatley, 2012; Wheatley *et al*, 2012). We can thus be confident that citizens' preferences are sufficiently structured for dimensional analysis, and also that voters are able to make sense of spatial representations.

Second, we must assume that early bird data provides a reliable indication of patterns found over the full course of a VAA. A dynamic validation based on early user data is practical only if the resulting scales continue to work reasonably well for later users. A potential objection to this is that early users may differ from late users (for example, in terms of political interest), which may cause scales that perform well in the early user sample to perform worse in the late user sample. However, while this is true, these differences are unlikely to be fundamental; voters with higher political interest may be slightly more ideologically consistent, but do not see the world in a completely different way (Leimgruber *et al*, 2010). Thus, it should be fair to assume that early user data provides a reasonably robust benchmark against which *ex-ante* defined VAA scales can be evaluated. Having said this, there are strategies available to check whether or not this assumption holds and to suggest potential remedies (see below).

A further potential objection against our suggestion is that dynamic validation necessarily implies that measurement quality cannot be guaranteed over the full course of a VAA. In the immediate post-launch phase, spatial maps will either be based on (potentially deficient) *ex-ante* defined scales or remain deactivated. However, given typical user figures scale validation and refining is more a matter of hours than days or even weeks. Still, the relatively short interval wherein spatial maps remain unvalidated (or deactivated) is a small price to pay for more valid and reliable spatial matching, and requires VAA designers to be fully transparent about this.

Despite these minor caveats, we believe dynamic scale validation offers a pragmatic method by which the quality of spatial matching in VAAs can be greatly improved. Admittedly, there are methodologically superior alternatives – including conducting a representative survey before the launch of a VAA – but these are hardly feasible in most situations. Conversely, dynamic validation allows for an essentially cost-free, universally applicable, demand side-based empirical validation of VAA spatial maps.

The Method of Dynamic Scale Validation

Before turning to an empirical demonstration, in this section we discuss the specifics of dynamic validation in some more detail. After a few general comments, we go on to consider methods appropriate for evaluating unidimensionality and reliability in the VAA context.

For practical applications of dynamic validation, a crucial question concerns the cut-off point at which scales should be validated. From a purely technical perspective, a few hundred observations are easily sufficient for scale validation; however, in reality it may often make sense to await a slightly bigger sample. VAAs may for instance diffuse from within a university setting, rendering the very first entries rather unrepresentative of the average VAA user. From our own experience, drawing on the first 2 000–5 000 entries tends to lead to good results. This number may appear large, but for established VAAs, reaching such a sample size may take just a matter of hours.

As argued above, even when drawing on a broadened sample, early and late users may differ on relevant variables, such as political interest. This may cause scales that perform well in an early user sample to perform worse in the late user sample. Two strategies are available for tackling this problem. First, scales can be retested at later points in time, and if necessary re-adjusted. Second, assuming clear hypotheses about how the average early user differs from the average late user, it is possible to test for equivalence across these traits in the early user sample and if necessary adjust the scales on this basis (Horn and McArdle, 1992). However, we do not think that it is absolutely necessary to repeat the validation exercise and/or test for equivalence. In most cases, a single dynamic validation suffices to push measurement quality to a fair level.

A critical requirement for a user-data based scale validation is that the data is 'clean'. VAA users often experiment with the tool. For validation, multiple entries and random click-throughs need to be filtered out. The research community has taken up the issue and several techniques have been suggested to filter out rogue entries (Andreadis, 2012). We strongly recommend that VAA designers implementing dynamic validation make use of such cleaning techniques.

Unidimensionality

We now turn to methods suited to dynamic validation of VAA scales. Focusing first on unidimensionality, a crucial observation to make is that the nature of VAA data renders the use of standard techniques for unidimensionality testing – confirmatory or exploratory factor analysis – problematic. VAA items tend to be ordered in terms of difficulty, meaning that some items are more easily endorsed than others. The *smartvote* data we introduce below illustrates this point. While 79 per cent of users endorse an open-minded foreign policy (item 61), only 44.7 per cent endorse EU accession (item 60). Put differently, the EU accession item is less popular and thus more difficult in terms of social liberalism than the open foreign policy item. Varying item difficulties violate the fundamental assumption of factor analysis that items are parallel (same means and frequency distributions). The most critical consequence of this is that unidimensionality can erroneously be rejected because of the extraction of one or more additional 'difficulty factors' (van Schuur, 2003).

If items are hierarchically ordered, item response theory (IRT) provides a viable alternative to factor analysis. IRT elegantly incorporates the concept of item difficulty by assuming that the probability of a particular response depends on both the characteristics of the person and the item. For the present purposes, we draw on the monotone homogeneity model (MHM), originally proposed in Mokken (1971) and extended to ordered polytomous items in Molenaar (1991). The MHM is a non-parametric form of IRT, and therefore often yields a better fit with empirical data compared with its parametric competitors, such as Rasch modelling (Hemker *et al*, 1995).

The empirical implications of the MHM are assessed with two tests. A scale can only be considered a unidimensional Mokken scale if both are passed (van Schuur, 2003). The first is the test of homogeneity, which draws heavily on Loevinger's *H*-coefficient. Two types of scalability coefficients play an important role. The overall *H*-score, on the one hand, indicates the overall precision of ordering individuals on the latent trait by means of the sum score (that is, average discrimination power). On the other hand, the item-specific H_i signifies discrimination power of individual items. Mokken (1971) suggested that for a scale to pass the test of homogeneity, both the overall *H* and all item-specific H_i need to exceed 0.3. According to a common rule of thumb, discrimination power is weak if $H \geqslant 0.3$, moderate if $H \geqslant 0.4$ and strong if $H \geqslant 0.5$.

The test of homogeneity can be applied in both a confirmatory and an exploratory mode. In its confirmatory mode, it is used for testing whether a given scale can be considered unidimensional. In its exploratory mode, it works as an automated search procedure for the identification of unidimensional scales similar to exploratory factor analysis. The exploratory search for unidimensional scales is *quasi*-inductive (and not fully inductive) in the sense that the results depend on the quantity and type of items in a test (Benoit and Laver, 2012). The exploratory mode works stepwise; items are consecutively added to scales based on the *H*-statistic until no item remains in the pool that fits the MHM (for more details, see Hemker *et al*, 1995; van Schuur, 2003).

The second test, the monotonicity test, builds on the fundamental implication of the MHM that items are monotone positively related to the latent trait. In short, it checks whether items are consistently non-decreasing functions of the latent trait (for a more detailed description, see Molenaar, 1991; van Schuur, 2003). The monotonicity test can only be employed for confirmatory purposes. Interpretation of the monotonicity test is facilitated by the diagnostic *crit* value devised by Sijtsma and Molenaar (2002). The *crit* value takes into account a number of aspects of model violation, whereby values above 80 are considered serious violations of monotonicity.[4]

Reliability

To this day, Cronbach's α remains the most-often reported reliability estimator. However, several attributes of VAA-type data render the α coefficient a bad choice. In particular, hierarchical item ordering by implication violates essential τ-equivalence and normality, leading to biased reliability estimates (Cortina, 1993; Sijtsma, 2009). While still reporting Cronbach's α, we propose to draw on an alternative estimator: the latent class reliability coefficient (LCRC) recently introduced by van der Ark *et al* (2011). Contrary to the α coefficient (as well as other reliability estimators, such as the Ω coefficient), the LCRC is well-suited for VAA-type data because it does not make rigid distributional assumptions.[5] The LCRC can be interpreted analogously to Cronbach's α: it ranges from 0–1, whereby higher values indicate better measurement precision (that is, a higher share of true score variance). Given that VAAs should provide individual users with reliable placements, the highest standards for measurement precision should apply. Generally speaking, the reliability estimate should push 1 for individual-level diagnosis. Values of 0.9 are often considered the lower bound of acceptance (Sijtsma, 2009).

Empirical Example

In the remainder of this article, we demonstrate the usefulness of dynamic scale validation by way of a real-world example: the 2007 version of the Swiss *smartvote* deployed before the federal elections. The 2007 version of *smartvote* featured multiple matching techniques. The primary system is high-dimensional, that is, it establishes issue-based ideological congruence. In addition, it featured two forms of spatial matching, one based on a two-dimensional framework (*smartmap*) and one based on an eight-dimensional framework (*smartspider*). The scales underlying these spatial maps were invariably determined *ex-ante*. Below, we emulate a dynamic validation of the two-dimensional *smartmap*.

Note that the *smartvote* setting is among the few that would allow us to go beyond user-based validation: given that it surveys hundreds of individual candidates for matching purposes, it would be possible (and recommendable) to establish a policy space common to both voters and elites. However, our aim is to illustrate user-based dynamic scale validation; we will therefore not exploit this unique avenue. More broadly, case selection was driven by the fact that s*martvote* passes easily as one of the most institutionalised VAAs (Fivaz and Nadig, 2010), and secondarily because the *smartvote* team was generous enough to share the data. Explicitly, it is not our aim to critique *smartvote*.

Two further comments are in order before we delve into the analysis. First, the *smartvote* questionnaire contains a total of 73 items; 63 items take the form of general policy statements and the remaining 10 relate to government spending.[6] We will not further consider the spending items, mainly because they employ a different answer format, which would complicate matters to a degree we deem unnecessary for the present purpose of illustration.[7] The remaining 63 policy items (see Table A1 in the Appendix) invariably employ the same four-point answer format ('yes', 'rather yes', 'rather no' and 'no'). Second, given that *smartvote* did not make use of data-cleaning mechanisms, we rely on data stemming from an additional opt-in survey. The self-selection mechanism associated with opting for an additional survey should guarantee that experimenting users are excluded from the scaling analysis. Moreover, the nature of the opt-in survey ensures that each user enters the analysis only once. The use of opt-in data reduces our *N* and may introduce some bias, but this is the only strategy available to access clean user data.

Examining the *ex-ante* scales

The 2007 version of the *smartmap* consisted of a left–right and a liberal–conservative axis. The underlying theoretical conceptualization was guided by earlier empirical work by Swiss researchers (Hermann and Leuthold, 2003). Because of this, the composition of the resulting dimensions appears rather unusual compared with internationally more established notions (for example, Kriesi *et al*, 2006; Marks *et al*, 2006). On the one hand, the left–right scale pertains not only to socio-economic issues, but also to some aspects of law and order (items 51, 54, 56 and 58) as well as military defence (items 52, 53, 55 and 57). On the other hand, the liberal–conservative scale contains a series of items referring to economic liberalism (items 5, 24, 26, 33, 34, 35, 36 and 37), including issues such as the introduction of a minimum wage and privatization of the state-owned phone company.[8]

Can *smartvote*'s *ex-ante* scales survive dynamic validation? Mimicking the situation of VAA designers, we draw on a sample of early users for scale validation. We selected as a cut-off point all users that had completed s*martvote* one calendar month before the federal elections of 21 October 2007. 3 872 out of 20 954 users in our data set accessed the site before our cut-off. Table 1 summarizes the results of the scaling analysis. It turns out that both *ex-ante* defined scales cannot be considered unidimensional. With their overall *H*-scores of 0.25 (for left–right) and 0.12 (for liberal–conservative), neither scale fits Mokken's MHM according to the guidelines stated above (the *H*-scores do not exceed 0.3). The test of monotonicity confirms the bad fit of the *ex-ante* defined scales. Several *crit* values come to lie above 80, a further indication of serious model violations. Meanwhile, the *ex-ante* scales perform much better in terms of reliability with estimates of 0.9 for left–right and estimates ranging from 0.79 to 0.82 for liberal–conservative. These figures are somewhat below desired levels (0.9 or more), but do not appear dramatic. However, marginally acceptable reliability cannot offset the lack of unidimensionality.

Identifying unidimensional scales

Having found the *ex-ante* scales wanting in terms of unidimensionality, the question is whether it would have been possible to correct the scales. For this, we turn to the exploratory mode of Mokken scaling (Hemker *et al*, 1995; van Schuur, 2003). This *quasi*-inductive technique implicitly tests for the adequacy of the two-dimensional structure. We include the whole item bank, thus also items not attributed to either of the ex-ante scales.[9] Since ordering in terms of difficulty implies that all items in a scale must point in the same direction (for example, towards social liberalism), we run the search with items in both original and reversed orders. Each scale is therefore outputted twice (in reversed orders); however, only one of the duplicates is reported. The lower bound for inclusion in a scale was set at 0.3 (see van Schuur, 2003).[10]

Table 1: Examining the *ex-ante* dimensions in early user sample

Left–right			*Liberal–conservative*		
Item	H_i	*crit*	*Item*	H_i	*crit*
1*	0.18	49	5*	0.05	121
2	0.33	10	6*	0.13	100
3	0.22	29	9	0.09	104
6	0.31	14	11	0.07	128
7*	0.28	12	12*	0.04	165
8*	0.24	47	13*	0.18	24
9*	0.22	48	14*	0.16	35
10*	0.02	132	16*	0.2	57
15	0.37	16	17*	0.13	91
23	0.32	17	18*	0.16	64
27	0.29	28	19	0.23	19
28	0.31	16	20*	0.16	62
31*	0.29	22	21*	0.12	61
32*	0.15	38	24*	0.07	100
33	0.29	39	25*	0.2	44
37	0.21	34	26*	0.05	137
38	0.29	17	29*	0.14	46
40	0.24	25	33	−0.0	281
41	0.23	27	34*	0.12	91
43*	0.34	7	35*	0.15	69
44*	0.34	0	36*	0.09	84
48*	0.11	115	37	0.06	109
51*	0.3	40	39	0.2	27
52*	0.19	85	44*	−0.08	497
53	0.28	41	46	0.03	171
54	0.16	70	53*	0.18	63
55	0.27	21	59*	0.15	26
56*	0.2	39	60*	0.17	30
57	0.21	41	61*	0.21	32
58*	0.32	18	62*	0.21	22
			63*	0.2	32
H	0.25		*H*	0.12	
α	0.9		*α*	0.79	
LCRC	0.9 (6)		*LCRC*	0.82 (7)	
N	1962		*N*	1938	

Note: Items with an asterisk are reversed; the number of latent classes used for estimation of the *LCRC* is given in brackets; for the test of manifest monotonicity, minimum rest score group size was consistently set to 100.

The *quasi*-inductive search for unidimensional scales confirms that a two-dimensional structure is adequate.[11] Table 2 presents the two resulting scales.[12] The composition of the scales supports our earlier scepticism regarding the conflation of

Table 2: Evaluation of the *quasi*-inductive dimensions in early user sample

Economic dimension			*Cultural dimension*		
Item	H_i	*crit*	*Item*	H_i	*crit*
2	0.4	13	9	0.33	16
7*	0.32	23	13*	0.51	0
23	0.41	53	14*	0.48	0
24*	0.33	72	16*	0.47	8
27	0.4	19	17*	0.43	34
28	0.36	40	18*	0.34	17
33	0.42	14	19	0.48	19
36*	0.33	54	20*	0.32	28
37	0.35	22	50*	0.38	35
38	0.39	16	51	0.44	16
			53*	0.44	6
			55*	0.41	13
			57*	0.33	15
			58	0.45	21
			60*	0.43	12
			61*	0.46	31
			62*	0.44	15
H	0.37		*H*	0.42	
α	0.83		*α*	0.9	
LCRC	0.83 (5)		*LCRC*	0.9 (5)	
N	2341		*N*	2299	

Note: Items with an asterisk are reversed; the number of latent classes used for estimation of the *LCRC* is given in brackets; for the test of manifest monotonicity, minimum rest score group size was consistently set to 100.

socio-economic and cultural issues: both reflect more established interpretations of political dimensionality (for example, Marks *et al*, 2006; Kriesi *et al*, 2006). The first scale consists of 10 items invariably related to the socio-economic cleavage, including issues related to the welfare state (items 2 and 7), taxation (items 27 and 28) and state interventions (items 23, 24, 33, 36, 37 and 38). The second scale contains 17 items, all referring to the cultural cleavage, broadly understood. It includes items pertaining to national sovereignty (items 60, 61 and 62), immigration (14, 16 and 17), cultural liberalism (items 18, 19 and 20), the army (items 53, 55 and 57), law and order (items 51 and 58) and institutional reform (items 13 and 50). We label the two scales the economic and the cultural dimension, respectively, to avoid confusion with the *ex-ante* defined scales. Both *quasi*-inductively derived scales can be considered Mokken scales: they pass the test of both homogeneity (all H_i and $H \geqslant 0.3$) and monotonicity (all *crit*-values come to lie below 80).[13] Notably, the fact that we do find methodologically viable and substantively meaningful political dimensions supports our earlier conjecture that demand side issue constraint is sufficient for the creation of political value scales.

Turning to measurement precision, the cultural dimension yields satisfactory reliability (0.9), while the economic dimension is somewhat below the desired level with its estimated reliability of 0.83. This underperformance does not appear dramatic, however. Moreover, additional analyses suggest that we cannot improve reliability by item removals.

Effectiveness of dynamic scale validation

Dynamic scale validation would imply that *smartvote*'s *ex-ante* defined scales are replaced with the *quasi*-inductive scales. As argued above, the adjustments will increase measurement quality because early user data provides a reasonably robust indication of patterns found over the full course of a VAA. To test this conjecture, we repeat the scaling analysis in the late user sample, that is, for all users that accessed the site after the previously selected cut-off.

Table 3 shows that the results are indeed virtually identical to the early user-based analysis. On the one hand, both *ex-ante* scales are found wanting in terms of unidimensionality, whereby the liberal–conservative dimension is again singled out as particularly problematic with its H-score of 0.12 (meanwhile, the left–right axis yields an H-score of 0.22). Also reliability estimates are similar, with estimates of 0.87–0.88 and 0.77–0.81, respectively. On the other hand, the test statistics indicate that the two *quasi*-inductive scales obtained from the subset of early users continue to work fairly well also in the late user sample, with test-scalability amounting to 0.32 and 0.38 for the economic and the cultural dimension, respectively. A smaller caveat may be that the H_i of a few items (7, 28 and 36 on the economic, and 20 and 57 on the cultural dimension) have fallen slightly below the minimal threshold of 0.3. However, the deviations are rather marginal and should not pose a fundamental problem. Moreover, all items (including the aforementioned ones) pass the test of monotonicity. Again, also reliability estimates are similar, with the cultural dimension yielding acceptable and the economic dimension marginally insufficient precision (0.89–0.9 versus 0.8–0.81). Overall, we may conclude that the dynamic, early user-based validation provided a reliable indication of patterns to be found in the late user sample. Maybe most importantly, scale adjustments based on dynamic scale validation would have led to significantly improved measurement quality.

So What?

Continuing with our *smartvote* example, we now turn to the 'so what' question and investigate implications of early scale adjustment for the message conveyed to its

Table 3: Examining the *ex-ante* and *quasi*-inductive dimensions in the late user sample

Left–right			*Liberal–conservative*			*Economic dimension*			*Cultural dimension*		
Item	H_i	*crit*	*Item*	H_i	*crit*	*Item*	H_i	*crit*	*Item*	H_i	*crit*
1*	0.15	83	5*	0.07	123	2	0.35	0	9	0.3	30
2	0.27	64	6*	0.12	113	7*	0.27	34	13*	0.46	7
3	0.2	71	9	0.08	172	23	0.36	5	14*	0.43	39
6	0.25	61	11	0.06	183	24*	0.3	16	16*	0.41	45
7*	0.23	71	12*	0.04	168	27	0.35	12	17*	0.39	11
8*	0.17	140	13*	0.17	73	28	0.29	49	18*	0.3	69
9*	0.19	99	14*	0.15	103	33	0.39	9	19	0.44	13
10*	0.0	275	16*	0.18	70	36*	0.26	36	20*	0.28	51
15	0.31	66	17*	0.14	102	37	0.3	43	50*	0.33	40
23	0.28	49	18*	0.14	87	38	0.36	13	51	0.4	44
27	0.24	51	19	0.21	37				53*	0.41	44
28	0.25	80	20*	0.16	61				55*	0.37	41
31*	0.23	65	21*	0.12	82				57*	0.29	51
32*	0.11	119	24*	0.06	191				58	0.41	61
33	0.26	61	25*	0.2	45				60*	0.39	50
37	0.17	76	26*	0.08	129				61*	0.41	19
38	0.25	52	29*	0.12	88				62*	0.41	44
40	0.21	75	33	0.0	449						
41	0.23	60	34*	0.13	99						
43*	0.3	53	35*	0.13	69						
44*	0.29	73	36*	0.1	116						
48*	0.08	182	37	0.06	186						
51*	0.26	72	39	0.17	73						
52*	0.15	148	44*	−0.07	806						
53	0.25	71	46	0.03	184						
54	0.15	87	53*	0.16	71						
55	0.25	67	59*	0.12	90						
56*	0.14	126	60*	0.16	135						
57	0.17	73	61*	0.2	61						
58*	0.27	69	62*	0.2	76						
			63*	0.18	54						
H	0.21		*H*	0.12		*H*	0.32		*H*	0.38	
α	0.87		*α*	0.77		*α*	0.8		*α*	0.89	
LCRC	0.88 (12)		*LCRC*	0.81 (12)		*LCRC*	0.81 (8)		*LCRC*	0.9 (11)	
N	7297		*N*	7328		*N*	9461		*N*	9299	

Note: Items with an asterisk are reversed; the number of latent classes used for estimation of the *LCRC* is given in brackets; for the test of manifest monotonicity, minimum rest score group size was consistently set to 100.

users. In addition, we provide further evidence that dynamic scale validation improves validity by drawing on external association with established ideology measures. Given that scale adjustments would have taken place only after the previously set cut-off, we consistently focus on late users for our assessment of practical implications of dynamic scale validation.

The effect on placements in the ideological space

As argued above, deficient measurement quality in VAA spatial maps will affect the positionings in the ideological space (for the effect on the implicit voting advice, see below). To gauge the extent of these deviations, we examine the 'agreement' between the *ex-ante* and the *quasi*-inductive placements. We will consider both VAA users and candidates (*smartvote* is among the rare cases that matches with individual candidates), because scale adjustments evidently affect the placements of both. The agreement is evaluated separately by dimension, meaning that we compare placements on the *ex-ante* defined left–right scale with placements on the *quasi*-inductively defined economic dimension, and analogously placements on the liberal–conservative scale with placements on the *quasi*-inductive cultural dimension. For the comparisons, we added up all items belonging to a given scale (in their respective direction) and normalized the resulting measures so that they always range from 0 to 1. Lin's (1989) concordance correlation coefficient, ρ_c, is used as measure of agreement. The logic of ρ_c is most easily understood by thinking of a square scatterplot of two measures. If the two measures are exactly the same (have perfect concordance), the scatterplot would look like a 45° line falling through the origin. The ρ_c evaluates the degree to which pairs of observations fall on this 45° line. In essence, it does this by multiplying a measure of dispersion (Pearson's r) by a measure of the deviations from the 45° line (denoted as C_b). Practically speaking, a low r represents random measurement error and a low C_b systematic measurement error. Lin (1989) suggested that $\rho_c > 0.9$ represents good concordance.

Table 4 gives the results. Concordance is generally low, indicating stark differences between *ex-ante* and *quasi*-inductive placements. This is true in particular for the cultural cleavage, which yields ρ_c-values of 0.64 and 0.56 for users and candidates, respectively. Moreover, the low C_b-values indicate that the differences

Table 4: Concordance of *ex-ante* and refined placements

Sample	*Scales*	*N*	ρ_c	*95 per cent CI*	*r*	C_b
Late users	Left–right	7045	0.79	[0.78, 0.8]	0.82	0.96
	Cultural	6983	0.64	[0.63, 0.65]	0.72	0.89
Candidates	Left–right	2694	0.91	[0.91, 0.92]	0.93	0.98
	Cultural	2694	0.56	[0.54, 0.58]	0.71	0.79

Note: ρ_c gives Lin's concordance correlation coefficient; 95 per cent CI gives the 95 per cent confidence interval; r gives Pearson's correlation coefficient; and C_b gives the bias-correction factor.

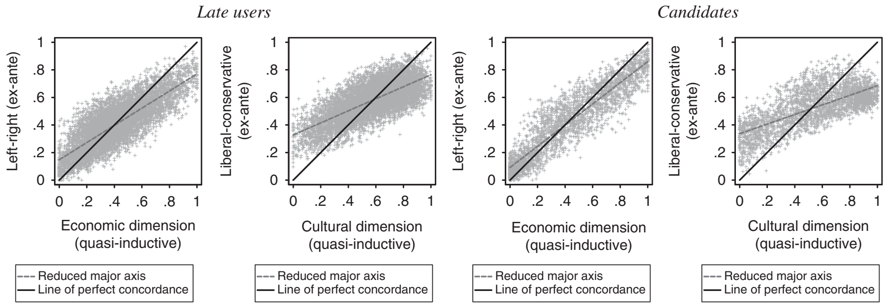

Figure 1: Concordance of *ex-ante* and refined placements.
Note: Points jittered 5 per cent.

are, at least in part, systematic. Figure 1 gives an indication of the distribution of these biases. The deviations (in grey) from the line of perfect concordance (in black) suggest that in comparison to the *quasi*-inductive version, *smartvote*'s *ex-ante* scale places conservatives systematically more towards the liberal end, and vice versa. Meanwhile, the differences are less pronounced for the left–right dimension. However, at least for the user side, concordance remains low with a ρ_c of 0.79, even though this is mostly because of random error.

In a nutshell, our assessment of concordance suggests that there are very substantial differences between the *ex-ante* and revised placements of both users and candidates. In principle, the fact that the latter by far outperform the former with regard to unidimensionality suffices to say that the *quasi*-inductively derived placements are more valid and should thus be preferred. Yet more intuitive evidence can be gained by investigating criterion-related validity (Carmines and Zeller, 1979). Criterion-related validation establishes the extent to which a measure compares with an alternative, established measure. The basic issue for us is the baseline of comparison; in particular, there are no established estimates of the ideological positions of individual VAA users. Instead, we compare aggregate positions of average party supporters and average party elites. For the demand side, our baseline is the average party supporter positions as estimated in Leimgruber *et al* (2010, p. 515), which are based on a post-election survey conducted shortly after the 2007 federal elections. For the supply side, the party positions as identified by the 2010 Chapel Hill expert survey (Bakker *et al*, 2012) serve as baseline.[14]

Our test of criterion-related validity is informal for two reasons. First, the limited number of parties pre-empts rigid statistical testing. We therefore make use of graphical representation for our comparison. Second, we cannot expect perfect concordance. VAA user data is plagued by self-selection, and cannot therefore be directly compared with a representative survey. On the other hand, the Chapel Hill expert survey was conducted 3 years after the 2007 federal election. The results of our comparison nonetheless bolster our confidence in the *quasi*-inductively derived

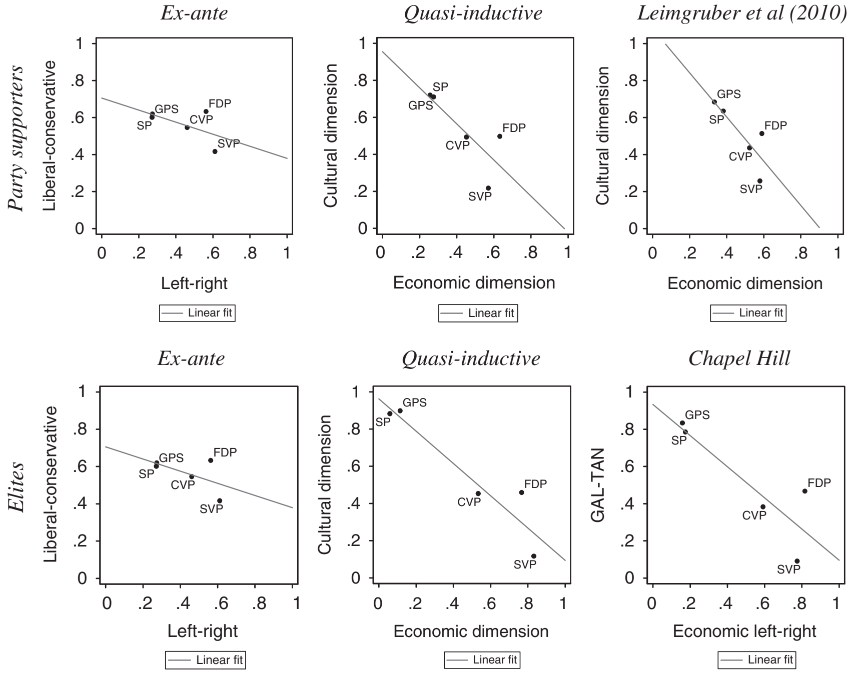

Figure 2: Assessing criterion-related validity.

scales (see Figure 2). Both a superficial glimpse at the slopes of the linear fit and a more detailed investigation of the positions of individual parties/party supporters suggest much stronger concordance between our *quasi*-inductive estimates and the more established estimates based on Leimgruber and his colleagues and the Chapel Hill expert survey, providing additional evidence that the *quasi*-inductively derived scales are superior.

The effect on the implicit voting advice

Arguably, the eye-catching element of spatial maps in VAAs is the implicit voting advice proffered. As argued above, a lack of unidimensionality has implications not only for placements in the ideological space, but also for the relational cue. The final question we ask is therefore: to what extent would scale adjustments based on dynamic validation have influenced the voting advice?

The scenario under which we assess such differences is hypothetical. S*martvote* matches users to individual candidates, but we will consider differences in matches to parties. Given the high number of candidates in most electoral districts, even minor

Table 5: Implications of scale adjustment on implicit voting advice

Party	*Ex-ante (in percentage)*	*Quasi-inductive (in percentage)*	*Difference (pp)*
CVP	16.19	16.18	–0.01
FDP	11.42	11.37	–0.05
SVP	6.32	5.26	–1.06
SP	4.68	5.19	0.51
GPS	16.35	20.04	3.69
GLP	22.85	13.16	–9.69
EVP	22.2	28.79	6.59

Note: $N = 6411$; *pp* denotes percentage points.

changes in the scales are likely to lead to fundamental changes in best matches. Therefore, matching to parties provides a more conservative estimate of the implications of dynamic scale validation; in addition, it renders our results more comparable with other VAA settings. We consider the seven largest parties. Party positions are calculated by averaging the positions of candidates from the seven respective main lists (for the distinction between main and subsidiary lists, see Bochsler, 2010). To do justice to the electoral setting in Switzerland where there is no single electoral district, we consider only candidates from the canton of Zurich. The Euclidean distance between the position of individual VAA users and parties serves as our measure of ideological congruence, whereby we follow the proximity model (Downs, 1957) and assume that the closest party constitutes the best match.[15]

It turns out that in our hypothetical scenario more than 4 out of 10 late users (42.21 per cent) would have received a different advice (that is, would have been placed closer to another party). Again, our estimate is conservative; had we based on individual candidates (as *smartvote* does) the numbers would have been much higher.

Table 5 shows how the shifts at the individual level translate to the aggregate level. The most striking result is that dynamic scale validation would have resulted in reducing the number of matches with the Green Liberals (GLP) by almost half. Meanwhile, upon scale correction an additional 6.59 per cent of late users would have been matched with the Protestant Party (EVP). By a much smaller margin, the Greens (GPS) would also have profited from scale adjustment, while the number of matches with the other parties would have remained relatively stable. Overall, this exercise has shown that dynamic scale validation can have significant implications for the implicit voting advice, both at the individual level and at the aggregate party level.

Conclusion

This article focused on a core property of many VAAs: the low-dimensional modelling of political preferences. Our principal concern was with the unidimensionality and reliability of the underlying latent ideology measures. Both concepts are essential for high-quality matching in a spatial framework. The current practice of assuming rather than empirically establishing their presence is therefore untenable.

Arguing that early user data provides a viable benchmark against which *ex-ante* defined maps can be evaluated, we made the case for dynamic validation of VAA spatial maps. Our recommendation is a pragmatic one. On the one hand, demand side-based validation implies that we superimpose the voter space on the elites. In the ideal case, we would rather want to match voters to parties in a framework that is common to both voters and elites. However, from a practical perspective this is in most cases simply not possible given the low number of observations on the political supply side. Moreover, we believe superimposing the voter space can be justified on conceptual grounds, given that the principal aim of a VAA is to provide voters with an indication of which party or candidate best matches their own preferences. On the other hand, dynamic validation necessarily implies that VAAs either launch with unvalidated scales or keep spatial features deactivated for some time. This obvious drawback could only be overcome by conducting a tailored survey before the launch. We suspect most VAA designers will lack the necessary resources to do this. Moreover, the price to pay is not dramatic given typical VAA user figures; often it will take a relatively short amount of time to validate (and potentially correct) spatial maps. Hence, in our view dynamic scale validation constitutes a viable compromise between methodological rigour and practicability, and offers a pragmatic means of maintaining fundamental properties like unidimensionality and reliability, at least in the (all too frequent) contexts where VAA designers are prevented from doing better.

To demonstrate the potential of dynamic scale validation, we evaluated an *ex-ante* defined spatial map from one of the most institutionalised VAA settings, Switzerland. While the map at hand consisted of two dimensions, our argument remains generalizable to frameworks involving more dimensions. Underlining the need for empirical validation, we found the *ex-ante* defined *smartmap* wanting in terms of unidimensionality. Critically, dynamic scale validation would have allowed spotting and correcting the deficiencies at a rather early stage of the VAA being online. Scale adjustments in line with the results obtained from early user-based validation would have paved the way for a significant improvement of the framework's psychometric utility. Perhaps most important of all, it was shown that dynamic scale validation is not a mere technicality. Early scale adjustments would have significantly affected the message carried by the spatial map, in particular in terms of the implicit voting advice.

In closing, a legitimate concern with our case selection may be that *smartvote* 2007 represents too easy a case for our argument. Admittedly, the conception of political dimensionality underlying *smartvote*'s two-dimensional map (see Hermann and Leuthold, 2003) deviates in many respects from standard understandings. It could therefore be argued that the deficiencies we found should have been obvious from the outset, and that VAA maps based on a more common conceptualization will not face the same problems. However, even if scales are based on established theoretical concepts, it is virtually impossible to attribute items to scales *a priori* without some margin of error. After all, this is precisely why the evaluation of scaling properties has become standard methodological practice in applied research (Clark and Watson, 1995). Therefore, the need for empirical validation extends beyond *smartvote*. In line with this, scaling analyses by Gemenis (2013) and Louwerse and Otjes (2012) have shown that the *EU Profiler*'s more standard scales were deficient (also see Germann and Mendez, 2013). Hence, given the growing number of VAA users worldwide, the upshot of scale validation, and dynamic scale validation in particular, is potentially very large indeed in terms of more valid and reliable spatial matches.

Acknowledgements

We would like to especially thank the *smartvote* team for their spirit of openness in encouraging us to critically assess their work and for making their data sets available to us. We would also like to thank Dominik Hangartner for sharing data with us as well as Andries van der Ark, Marco Steenbergen, Kostas Gemenis, Simon Otjes, Daniel Bochsler and two anonymous reviewers for their helpful comments. *R* (v3.0.1) was used for the scaling analysis, in particular the *mokken* (van der Ark, 2007, 2012) and the *poLCA* (Linzer and Lewis, 2011) packages. *Stata* (v11.2) was used for the assessment of the practical implications of dynamic scale validation. Data as well as replication codes will be made available upon publication on www.preferencematcher.org.

About the Authors

Micha Germann is a PhD researcher at the Centre for Research on Direct Democracy (c2d), University of Zurich and at the Centre for Comparative and International Studies, ETH Zurich. His research interests include direct democracy, political violence, e-voting and voting advice applications (VAAs).

Fernando Mendez is Director of the e-Democracy Centre and is also lecturer at the University of Zurich and the ETH Zurich. His research interests include federalism, European Union politics, direct democracy and e-democracy (especially e-voting and voting advice applications). He has published in all of these fields.

Jonathan Wheatley is a senior researcher at the Centre for Research on Direct Democracy (c2d). He is also lecturer at the University of Zurich and the ETH Zurich. His research interests include democratization, state-building, parties and party systems in developing democracies, and voting advice applications (VAAs) in both established and developing democracies.

Uwe Serdült is vice-director of the Centre for Research on Direct Democracy (c2d) and lecturer at the University of Zurich and the ETH Zurich. His research interests include direct and electronic democracy, decision making processes and structures, institutional change and comparative public policy.

Notes

[1] We know of a single exception to this rule: as of 2011, the Swiss VAA *smartvote* has started to validate the two-dimensional spatial map against elite-level data. However, for the reasons stated below, we would argue that *smartvote* should take into account the voter side as well.

[2] However, in the rare cases where there are sufficient elite-level observations, such as *smartvote*, VAA designers should in principle strive for a common space in order to ensure maximum-quality spatial matching, that is, establish equivalence across voters and elites.

[3] It follows that validation based exclusively on the political supply side appears less sensible. Where supply side validation is possible, VAA designers should always consider the voter side as well.

[4] For the test of monotonicity, we consistently set the minimum size of the rest score group to 100. The default algorithm for the determination of the rest score group size tends to deflate test results in case of a very large N (Sijtsma and Molenaar, 2002; van der Ark, 2007).

[5] Estimation of the LCRC requires prior determination of the optimal number of latent classes (van der Ark *et al*, 2011). We used the Bayesian information criterion to choose among different models.

[6] *Smartvote* users had the option of filling in a shorter version of the questionnaire.

[7] There are also substantive reasons against including spending items (Gemenis, 2013). Meanwhile, the differences between the original scales (with spending items) and our shorter scales (without spending items) are marginal ($\rho_c = 0.98$ for both).

[8] A further relatively odd feature is that some items are attributed to both *ex-ante* scales. This is bad practice. By definition, the assignment of items to multiple scales violates the requirement of external consistency, and thus unidimensionality.

[9] We included 58 out of the 63 items in the final quasi-inductive search we report, excluding items 6, 15, 43, 44 and 45. This is because additional analyses suggest that they correlate strongly with both latent traits. Cross-loading items can distort the search procedure, that is, they may lead to the creation of multidimensional scales (van Abswoude *et al*, 2004). Moreover, cross-loading items may violate external consistency. We decided to err on the side of caution, and excluded the strongly cross-loading items.

[10] In applications, it may be defendable to retain an item even if it falls slightly below the 0.3 level to safeguard reliability and/or content validity.

[11] The quasi-inductive search yields more than two scales, but the remaining scales consist of a maximum of three items, tend to be weak, similarly worded and tap very narrow constructs, which moreover are also represented in the two main scales. Hence, the remaining scales are best conceived of as method artifacts; we will not consider them further. Exploratory factor analysis (with polychoric correlations as input matrix) confirms the two-dimensional structure. The first five eigenvalues are 15.29, 4.89, 2.22, 1.67 and 1.09.

[12] The search procedure attributed an additional item (41) to the economic dimension, but its H_i fell below 0.3 by the end of the procedure. Following van Schuur (2003, p. 150), we removed the item from the scale.

13 Discrimination power of the two scales is sufficient but rather weak, at least for the economic dimension. Discrimination power could be increased by raising the threshold for inclusion in a scale. However, the resulting scales are significantly shorter and yield significantly lower reliability estimates. This seems too big a sacrifice to make.

14 To ensure comparability, the scores were normalized in both cases so that they range from 0 to 1. Note that we used −1 and 1 as the respective minimum and maximum for the figure reproduced from Leimgruber *et al* (2010, p. 515).

15 In reality, a user is free in her interpretation of a spatial map; she may in particular also employ a directional logic.

References

Achen, C.H. (1975) Mass political attitudes and the survey response. *American Political Science Review* 69(4): 1218–1231.

Andreadis, I. (2012) To clean or not to clean? Improving the quality of VAA data. Paper presented at the 2012 IPSA World Congress; 8–12 July, Madrid, Spain.

Ansolabehere, S., Rodden, J. and Snyder, Jr J.M. (2008) The strength of issues: Using multiple measures to gauge preference stability, ideological constraint, and issue voting. *American Political Science Review* 102(2): 215–232.

Baka, A., Figgou, L. and Triga, V. (2012) Neither agree, nor disagree: A critical analysis of the middle answer category in voting advice applications. *International Journal of Electronic Governance* 5(3/4): 244–263.

Bakker, R. *et al* (2012) Measuring party positions in Europe: The Chapel Hill expert survey trend file. *Party Politics,* 1999–2010, advance online publication 29 November, doi:10.1177/1354068812462931.

Benoit, K. and Laver, M. (2012) The dimensionality of political space: Epistemological and methodological considerations. *European Union Politics* 13(2): 194–218.

Bochsler, D. (2010) Who gains from apparentments under D'Hondt? *Electoral Studies* 29(4): 617–627.

Carmines, E.G. and Zeller, R.A. (1979) *Reliability and Validity Assessment*. Thousand Oaks, CA: Sage.

Clark, L.A. and Watson, D.B. (1995) Constructing validity: Basic issues in objective scale development. *Psychological Assessment* 7(3): 309–319.

Converse, P.E. (1964) The nature of belief systems in mass publics. In: D.E. Apter (ed.) *Ideology and Discontent*. New York: Free Press, pp. 206–261.

Cortina, J.M. (1993) What is coefficient alpha? An examination of theory and application. *Journal of Applied Psychology* 78(1): 98–104.

Cronbach, L.J. and Meehl, P.E. (1955) Construct validity in psychological tests. *Psychological Bulletin* 52(4): 281–302.

Davidov, E. (2009) Measurement equivalence of nationalism and constructive patriotism in the ISSP: 34 countries in a comparative perspective. *Political Analysis* 17(1): 64–82.

Downs, A. (1957) *An Economic Theory of Democracy*. New York: Harper and Row.

Fivaz, J. and Nadig, G. (2010) Impact of voting advice applications (VAAs) on voter turnout and their potential use for civic education. *Policy & Internet* 2(4): 167–200.

Gemenis, K. (2013) Estimating parties' policy positions through voting advice applications: Some methodological considerations. *Acta Politica* 48(3): 268–295.

Gerbing, D.W. and Anderson, J.C. (1988) An updated paradigm for scale development incorporating unidimensionality and its assessment. *Journal of Marketing Research* 25(2): 186–192.

Germann, M., Mendez, F., Wheatley, J. and Serdült, U. (2012) Exploiting smartvote data for the ideological mapping of Swiss political parties. Paper presented at the 2012 Convegno SISP; 13–15 September, Rome, Italy.

Germann, M. and Mendez, F. (2013) Dynamic scale validation reloaded: Assessing the psychometric properties of latent measures of ideology in VAA spatial maps. Paper presented at the 2013 ECPR General Conference; 4–7 September, Bordeaux, France.

Hattie, J. (1985) Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement* 9(2): 139–164.

Hemker, B.T., Sijtsma, K. and Molenaar, I.W. (1995) Selection of unidimensional scales from a multidimensional item bank in the polytomous Mokken IRT model. *Applied Psychological Measurement* 19(4): 337–352.

Hermann, M. and Leuthold, H. (2003) *Atlas der politischen Landschaften. Ein weltanschauliches Porträt der Schweiz*. Zurich, Switzerland: vdf Hochschulverlag.

Horn, J.L. and McArdle, J.J. (1992) A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research* 18(3): 117–144.

Kriesi, H., Grande, E., Lachat, R., Dolezal, M., Bornschier, S. and Frey, T. (2006) Globalization and the transformation of the national political space: Six European countries compared. *European Journal of Political Research* 45(6): 921–956.

Leimgruber, P., Hangartner, D. and Leemann, L. (2010) Comparing candidates and citizens in the ideological space. *Swiss Political Science Review* 16(3): 499–531.

Lin, L.I. (1989) A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 45(1): 255–268.

Linzer, D.A. and Lewis, J.B. (2011) poLCA: An R package for polytomous variable latent class analysis. *Journal of Statistical Software* 42(10): 1–29.

Lord, F.M. and Novick, M.R. (1968) *Statistical Theories of Mental Test Scores*. Reading, UK: Addison-Wesley.

Louwerse, T. and Otjes, S. (2012) Design challenges in cross-national VAAs: The case of the EU profiler. *International Journal of Electronic Governance* 5(3/4): 279–297.

Louwerse, T. and Rosema, M. (2011) The design effects of voting advice applications: Comparing methods of calculating results. Paper presented at the ECPR General Conference; 25–27 August, Reykjavik, Iceland.

Marks, G., Hooghe, L., Nelson, M. and Edwards, E. (2006) Party competition and European integration in the East and West: Different structure, same causality. *Comparative Political Studies* 39(2): 155–175.

Mendez, F. (2012) Matching voters with political parties and candidates: An empirical test of four algorithms. *International Journal of Electronic Governance* 5(3/4): 264–278.

Mokken, R.J. (1971) *A Theory and Procedure of Scale Analysis Applications in Political Research*. New York: De Gruyter.

Molenaar, I.W. (1991) A weighted loevinger H-coefficient: Extending Mokken scaling to multicategory items. *Kwantitatieve Methoden* 12(37): 97–117.

Sijtsma, K. (2009) On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika* 74(1): 107–120.

Sijtsma, K. and Molenaar, I.W. (2002) *Introduction to Nonparametric Item Response Theory*. Thousand Oaks, CA: Sage.

Van Abswoude, A.A.H., van der Ark, A.L. and Sijtsma, K. (2004) A comparative study of test data dimensionality assessment procedures under nonparametric IRT models. *Applied Psychological Measurement* 28(1): 3–24.

Van der Ark, A.L. (2007) Mokken scale analysis in R. *Journal of Statistical Software* 20(11): 1–19.

Van der Ark, A.L. (2012) New developments in Mokken scale analysis in R. *Journal of Statistical Software* 48(5): 1–27.

Van der Ark, A.L., van der Palm, D.W. and Sijtsma, K. (2011) A latent class approach to estimating test-score reliability. *Applied Psychological Measurement* 35(5): 380–392.

Van der Brug, W. and van Spanje, J. (2009) Immigration, Europe and the 'new' cultural dimension. *European Journal of Political Research* 48(3): 309–334.

Van Schuur, W.H. (2003) Mokken scale analysis: Between the Guttman scale and parametric item response theory. *Political Analysis* 11(1): 139–163.

Walgrave, S., Nuytemans, M. and Pepermans, K. (2009) Voting aid applications and the effect of statement selection. *West European Politics* 32(6): 1161–1180.

Wheatley, J. (2012) Using VAAs to explore the dimensionality of the policy space: Experiments from Brazil, Peru, Scotland and Greece. *International Journal of Electronic Governance* 5(3/4): 318–348.

Wheatley, J., Carman, C., Mendez, F. and Mitchell, J. (2012) The dimensionality of the Scottish political space: Results from an experiment on the 2011 Holyrood elections. *Party Politics*, advance online publication 30 September, doi:10.1177/1354068812458614.

Appendix

Table A1: Item descriptions (shortened version)

Item	*Statement*	*Item*	*Statement*
1	Raise pension age	33	Minimum wage of 3 500 Swiss Francs
2	Income-based health insurance premiums	34	Total liberalization of shopping hours
3	Cover alternative medicine by basic health insurance	35	Permit parallel imports
		36	Privatize national telephone company
4	Limit free choice of doctor	37	Maintain network of post offices
5	Free choice of second pillar pension fund	38	Mandatory funds for apprenticeship places
6	Federal subsidies for day-care centres	39	Government should favour Swiss companies
7	Cut unemployment benefits	40	Introduce road pricing
8	Replace student grants by repayable loans	41	Raise environmental standards for new buildings
9	Special schools for troublesome children		
10	State subsidies for private schools	42	Relax protection provisions for wolves
11	Ban genetically modified food	43	New nuclear power stations
12	English as first foreign language in school	44	Limit association's right of appeal
13	Communal right to vote for foreigners	45	Reduce greenhouse gas emissions
14	Judicial appeal for asylum-seekers	46	Freeze construction zones
15	Collective residence permit for sans-papiers	47	Animal protection lawyer
16	Restrict use of the ballot box to decide on naturalization	48	Financial referendum at federal level
		49	Direct election of the Federal Council
17	More funds for integration of foreigners	50	Lower voting age to 16
18	Gay adoption	51	Tighten juvenile law
19	Ban of minarets	52	Supporting role of the army in internal security
20	Legalize cannabis	53	Free choice between military and civilian service
21	Legalize active euthanasia		
22	Federal ban on smoking in public buildings	54	Stricter controls of driving regulations
23	Check parity of pay between men and women	55	Storage of military weapons in the armoury
		56	Preventive monitoring of personal communication
24	Abolish fixed book prices		
25	Abortion	57	Abolition of military courts
26	VAT reform	58	Severer punishment for vandalism
27	Tax equality between cantons and communes	59	Foreign deployment of armed Swiss troops
28	Ban degressive tax rates	60	Start negotiations over EU membership
29	Individual taxation for married couples	61	Active and open foreign policy
30	Tax reform concerning commuters	62	Extend free movement of peoples to Romania/Bulgaria
31	Cut federal taxes		
32	Replace federal taxes with higher VAT rates	63	Facilitate agricultural imports from developing countries