Extrapolation of Experimental Results through Analogical Reasoning from Latent Classes

Gerdien G. van Eersel, Gabriela V. Koppenol-Gonzalez, and Julian Reiss*†

In the human sciences, experimental research is used to establish causal relationships. However, the extrapolation of these results to the target population can be problematic. To facilitate extrapolation, we propose to use the statistical technique Latent Class Regression Analysis in combination with the analogical reasoning theory for extrapolation. This statistical technique can identify latent classes that differ in the effect of X on Y. In order to extrapolate by means of analogical reasoning, one can characterize the latent classes by a combination of features and then compare these features to features of the target.

Introduction. Experimental research is often regarded as the gold standard in the human sciences (e.g., Angrist and Pischke 2010; for a review in medicine, see Bothwell et al. 2016). In this kind of research, randomization aims to control for confounders by balancing the different conditions with respect to all possible causal factors (for limitations, see, e.g., Deaton and Cartwright 2018). If randomization is successful, experimental research warrants causal conclusions (e.g., Holland 1986). These conclusions are typically formulated in terms of group averages (for historical accounts, see Danziger 1987; La-

Received September 2017; revised July 2018.

*To contact the authors, please write to: Gerdien G. van Eersel, Department of Psychology, Education, and Child Studies, Erasmus University Rotterdam, PO Box 1738, 3000 DR Rotterdam, The Netherlands; e-mail: gerdienveersel@gmail.com.

†We acknowledge support by the Erasmus Open Access Fund of Erasmus University Rotterdam.

Philosophy of Science, 86 (April 2019) pp. 219–235. 0031-8248/2019/8602-0002$10.00

miell 2009). However, what applies in aggregate is not necessarily informative about what is true for every individual in the sample population (for papers addressing this problem of group averages, see, e.g., Rothwell 1995; Molenaar 2004; Navarro et al. 2006; Hamaker 2012; Speelman and McGann 2013; Grice 2015; Rousselet, Pernet, and Wilcox 2017). That is, if in an ideal experiment X raises the probability of Y, the only thing one actually knows is that X raises the probability of Y in some subgroup of the sample (Cartwright 2007).[1] A subgroup is a maximally homogeneous group of individuals with a fixed arrangement of causal factors. Because within a subgroup, all causes of Y other than X are fixed at a certain level, the effect of X on Y is also fixed. The effect of X on Y in other subgroups may vary greatly, depending on the type and degree of interaction between the other causal factors and X.

This limitation makes it difficult to extrapolate an average experimental outcome from the sample to the target population—the group of individuals to which the outcome is to be generalized. If the sample and the target population were each homogeneous and identical with respect to the settings of all other causal factors, the average result would represent the full sample, and the result could be generalized to all individuals in the target (Steel 2008, 3). However, populations in the human sciences are heterogeneous, which implies that it is not clear to which individuals in the target population the effect can be extrapolated. This has been called the problem of extrapolation (e.g., Steel 2008). According to Cartwright (2007), to justify the claim that X causes Y in some members of the target population, one needs to assume that at least one of the subgroups of the sample in which X causes Y is identical in causal structure to a subgroup of the target population. For this assumption to be met, one first needs to determine what the target population is and then draw a large random sample. As a result, the distribution of causal factors in the sample will be similar to the distribution of causal factors in the target population, and the larger the sample, the more likely that the distributions are indeed similar. The problem, though, is that in practice it is complicated to define a target population and then draw a random sample from this population (e.g., Sears 1986; for a proposal on how to do this, see Simons, Shoda, and Lindsay 2017). Furthermore, because we lack information about the target population and we do not know all other causal factors, it is impossible to test whether the assumption has been met.

Suppose that the assumption is not satisfied: the subgroups in which X causes Y, with their particular fixed arrangements of other causal factors,

1. While positive results in the ideal experiment demonstrate the existence of causation in a subpopulation, neutral results do not demonstrate its absence. If there is no difference between the experimental (X) and the control (~X) group concerning the probability of Y it can still be true that X causes Y in a subgroup of the sample. This is the case when there is another subgroup of the sample in which X has the opposite effect, and the two mutually cancel out (Cartwright 2007).

do not share their causal structure with any subgroup of the target population. Is it then still possible to extrapolate the result from the sample to the target population? Furthermore, even if the assumption is met, the question remains what the effect of X on Y is in the rest of the target population, which is composed of other (unidentified) subgroups. The answers to these questions depend on the distribution of the causal factors in the target and on the degree and type of interaction between those factors and X. This issue is an instance of what Steel (2008, 4) calls the extrapolator's circle: "The challenge . . . is to explain how the suitability of the model as a basis for extrapolation can be established given only limited, partial information about the target." So, to substantiate the extrapolation of a causal relation one needs to have detailed information about the target, but in that case extrapolation would be unnecessary (4).

In sum, experimental studies are very useful to infer causal conclusions, but extrapolation of the results to the target population can be problematic. In what follows, two theories will be discussed that address the problem of extrapolation: one by Steel (2008) and one by Guala (2003, 2005). Steel's method for extrapolation relies on mechanisms, but we will claim that this method is not helpful when extrapolating from experimental systems in the human sciences. We will then argue that the combination of Guala's theory and a certain statistical technique can facilitate the extrapolation of experimental results in the human sciences.

Steel's Theory of Comparative Process Tracing. Steel (2008) employs causal mechanisms as a starting point for extrapolation. According to Steel, a causal mechanism consists of "interacting components that generate a causal regularity between some specified beginning and end points" (40). In order to learn about these mechanisms, one uses a strategy called 'process tracing'. Starting from a beginning or an endpoint, one traces forward or backward to reconstruct the paths connecting the variables that constitute the mechanism. To then extrapolate knowledge of this mechanism from the model to the target, one follows a procedure called 'comparative process tracing' (89): comparing stages of the mechanism in the model with those in the target in which the two are expected to differ. The greater the similarities in the key stages of the mechanism, the more promising the extrapolation. Process tracing is most appropriate in cases where the causal relationships among components are more accessible than those among macrofeatures of the system (175). In order to infer causal relations among a set of variables that represent macrofeatures of a certain system, process tracing studies the relations among the features of the component parts of the system (187).

Steel's theory can be very useful if mechanistic knowledge is available. However, Steel (2008) acknowledges that there are many cases in the human sciences where there is knowledge available about causal relations,

but not about the mechanisms underlying them. This is particularly true in the case of experimental results, according to Steel, because mechanistic knowledge is hard to obtain from experiments (163). There is just knowledge of the probabilistic relation between cause and effect, but knowledge of how the two are connected is often lacking. Moreover, in many cases there are several plausible mechanisms (e.g., Gerring 2010). It is hard to discover which ones are actually operating, under what conditions, and how they interact. Furthermore, mechanisms themselves do not always behave in a regular manner (e.g., Howick, Glasziou, and Aronson 2013), and the same mechanism can have opposite effects in different contexts (Elster 1998). In addition, variables in the human sciences are mostly *latent*. They are not directly observed; there is no one-to-one relationship between a construct and its operationalization. This makes the construction of a mechanism even more difficult, as we have knowledge of mechanisms mostly at the level of operationalizations. Another problem of using Steel's mechanistic approach is that it is most appropriate for extrapolating *qualitative* causal claims (Steel 2008, 94). Many questions in the human sciences, however, are of a quantitative nature (Reiss 2008, 122). All in all, although the mechanistic account can be very useful, it is not meant to solve the problem of extrapolation from experimental systems in the human sciences.

Guala's Theory of Analogical Reasoning. Guala's (2003, 2005) theory is based on a form of analogical reasoning: if one has (*a*) some controlled initial conditions in the laboratory and (*b*) some observed experimental result on the one hand and given (*c*) some observed properties of the target system and (*d*) some observational data on the other, then (by analogy) *c* stands in the same (causal) relation to *d* as *a* stands to *b* (Guala 1998, 72). To extrapolate an experimental outcome, one has to ensure that the test system (i.e., the setting and the sample) and the target system (i.e., the setting and the population) are as similar as possible; the two systems should match in all respects that are relevant according to our current background knowledge (Guala 2005, 2010). One should make sure that differences between the test and the target system do not affect the relationship between X and Y. For example, it is only possible to extrapolate from rats to humans if there is evidence that the differences between the two species do not alter the causal relation being extrapolated. Moreover, an experiment cannot demonstrate much about a target system on its own, because different causal processes may generate similar patterns of data (Guala 1998). One therefore needs to rely on observational data. If the experimental and observational data are similar, by means of analogical reasoning, the causal relationship in question can be extrapolated to the target system.

To illustrate the analogical approach, Guala (2005, 2010) takes an example from experimental economics. Outer Continental Shelf (OCS) auctions

are common value auctions in which the auctioned good has the same value for all participants without anyone knowing what the value is. These auctions turned out to be systematically overpriced, possibly because of the specific auction rules. To test this explanation, Kagel and Levin (1986) performed an experimental study in which bidders did not know an item's value and were provided with a private or public information signal. According to the winner's curse hypothesis, public information reduces the uncertainty concerning the value of the item, thereby reducing overestimation. Kagel and Levin observed that participants' bids decreased when a public information signal was provided (917), confirming the winner's curse hypothesis. Now the question was whether the winner's curse, observed under experimental conditions, could explain the overpriced OCS auctions in the target system. To answer this question, Kagel and Levin discussed an observational study by Mead, Moseidjord, and Sorensen (1983) about the profits from the leases of two different kinds of tracts: wildcat and drainage tracts. More public information was available about the productivity of drainage tracts. The winner's curse hypothesis predicts that more public information reduces uncertainty about the item's value, thereby lowering bidders' overestimation of the value and raising profits (Kagel and Levin 1986, 915–16). Mead and colleagues (1983) indeed found higher profits from drainage leases compared to wildcat leases, which again confirmed the winner's curse hypothesis.

Guala illustrates the parallels between the experimental and real-world processes in figure 1 (2010, 1077). The arrows represent causal relationships, and the double lines represent analogical correspondences. The dashed lines depict the conclusion following from the extrapolation. Guala writes: "The analogy was backed up by looking for consequences or side-effects of the (relevantly similar) mechanisms that supposedly govern the laboratory and the target systems" (1077), and "by comparing analogous variations in

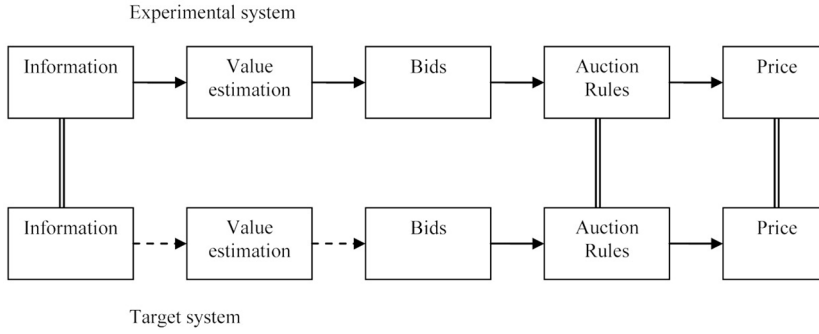

Figure 1. Analogical reasoning in the Outer Continental Shelf example. Source: Guala (2010, 1077).

https://doi.org/10.1086/701956 Published online by Cambridge University Press

information and prices at a downstream stage of each mechanism, it is possible to extrapolate the winner's curse result from the laboratory to the field" (1078). In sum, the experimental system was analogous to the target system, and the observational data corresponded to the experimental results. As a result, the relationship between the amount and type of information and the bidders' profits could be extrapolated by means of analogical inference from the laboratory to the target system. Note, however, that it is always possible that the similar outcomes observed in the laboratory and the target were still caused by different factors, as inductive inferences like these are by nature fallible.

Mechanistic Knowledge. The OCS case is one of the few examples Guala gives to illustrate his theory. The extrapolation in this example is based on just two analogies: the amount and type of information on the item's value and the bidders' profits. Moreover, the example is *mechanistic* in nature. In the passage above, Guala invokes mechanistic knowledge, which he does in other writings as well (e.g., Guala 2003). Now if mechanistic knowledge is available, as in the OCS case, it might be used to extrapolate by means of analogical reasoning. However, for many experimental results in the human sciences, mechanistic knowledge is hard to obtain.

Although Guala (2010) stresses that the OCS example and the analogical approach in general do not necessarily rely on mechanistic explanation, he does not show how an extrapolation can be successful when not founded on mechanistic knowledge. Is it possible to extrapolate through analogical reasoning when there is no mechanistic knowledge on hand? We will demonstrate that it is. According to Guala, if there are sufficient analogies between the properties of the test system and the target system, and the experimental data and the observational data have similar outcomes, then the relationship at issue will presumably be valid in systems where the experimental conditions do not hold. We argue that it is not necessary that these properties are part of a mechanistic explanation. In what follows, a statistical technique will be discussed that can accommodate extrapolation according to Guala's approach but without the help of mechanisms.

Latent Class Regression Analysis. As discussed in the introduction, in a well-conducted experiment, if X raises the probability of Y, then X causes Y in some subgroup of the test population. However, the effect of X on Y in other subgroups can be very different. In this section, a statistical technique is discussed that identifies *latent classes* varying in the effect of X on Y, which can be used to extrapolate by means of analogical reasoning.

Latent Class Regression Analysis (LCRA; Wedel and DeSarbo 1994; Hagenaars and McCutcheon 2002; Magidson and Vermunt 2004) is a technique that can be used to predict the scores on a response variable (Y) from

https://doi.org/10.1086/701956 Published online by Cambridge University Press

a set of predictors (X). The difference between standard regression analysis and LCRA is that in the latter, a categorical latent variable is postulated. A latent variable is not directly measured but rather inferred from a set of observed variables. This latent variable can divide the sample into possible latent classes that differ with regard to the relationship between the predictor(s) and the response variable. The underlying idea is that the population consists of a number of subpopulations that differ with respect to the estimated parameters of the statistical model of interest. Scores of participants belonging to the same latent class are assumed to come from the same probability distribution. A latent class consists of individuals for whom the predicted effect of X on Y is similar but whose predicted effect is substantially different from that of individuals from other latent classes, and a regression model is estimated for each of the classes separately.[2]

Note that this approach differs from subgroup analysis (e.g., Szczech, Berlin, and Feldman 1998; Kent et al. 2002) in the sense that one does not need to know in advance what the subgroups are (for methodological problems of subgroup analysis, see Lanza and Rhoades 2013). Instead, the subgroups are empirically established using the data at hand. As an illustration of the difference between LCRA and subgroup analysis, consider the following study by Koppenol-Gonzalez, Bouwmeester, and Vermunt (2014). A sample of 210 children age 5–12 years was presented with memory tasks designed to differentiate between verbal and visual short-term memory (STM) processes. The tasks were serial order reconstruction tasks requiring participants to place a set of presented pictures in their correct serial order. One sequence had visually similar pictures and represented words that were phonologically dissimilar, and the other sequence had visually dissimilar pictures and represented words that were phonologically similar. Similarity was manipulated in order to distinguish between the use of verbal and visual STM processes. That is, interference of phonological similarity implies the use of verbal processes, whereas the interference of visual similarity implies the use of visual processes. The dependent variable was a vector of correct/incorrect responses of placing the pictures on the correct serial position. The performance across the positions was expressed in a serial position curve, which is the pattern of responses on each of the positions. The serial position curve of the visually similar task was compared with the serial position curve of the phonologically similar task. It was expected that when the serial position curve of the phonologically similar task showed significantly worse performance than that of the visually similar task, this indicated the use of specific verbal processing. Conversely, when the serial position curve of

2. Latent class analysis (LCA) does not require a regression model; there are other types of LCA as well, like latent profile analysis. For our purposes, the focus will be on LCRA.

the visually similar task showed significantly worse performance than that of the phonologically similar task, this indicated the use of specific visual processing.

The data were analyzed using two approaches; subgroup analysis assuming age is an appropriate grouping variable, and LCRA assuming a latent grouping variable reflecting the use of verbal/visual STM processes (Koppenol-Gonzalez et al. 2014). The results showed that the two approaches led to different conclusions. Using age as a grouping variable, the children did not seem to use a specific type of STM (see fig. 2*a*) because the scores on the visually similar pictures did not differ significantly from the scores on the phonologically similar pictures. However, using LCRA five latent classes were distinguished that did differ in the use of verbal/visual STM (see fig. 2*b*). Moreover, age did not play an important role in the formation of the latent classes. Specifically, the age ranges of the five classes indicated that younger

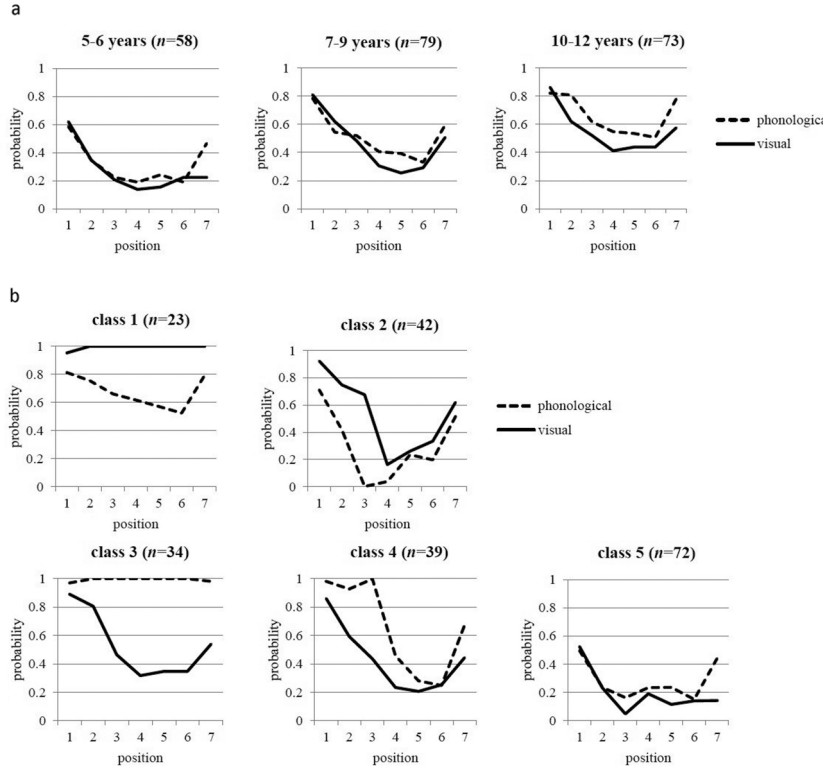

Figure 2. Serial position curves with the probability of correct responses to phonologically and visually similar items for each age group (*a*) and serial position curves for the latent classes (*b*). Source: Koppenol-Gonzalez et al. (2014).

https://doi.org/10.1086/701956 Published online by Cambridge University Press

and older children could be classified into any of the latent classes. Below we will explain the different steps of LCRA on the basis of this study. The LCRA starts by fitting the model with one latent class and subsequently with an increasing number of latent classes. The best balance between the fit of the model to the data and the parsimony of the model in terms of estimated parameters can be found using information criteria like the Bayesian information criterion (BIC), the Akaike information criterion (AIC), or the likelihood ratio chi-squared statistic (Magidson and Vermunt 2004). If the data are best described by multiple latent classes, then parameter estimations will differ across these classes. Participants' class posterior membership probabilities are estimated for each latent class. The class membership probabilities add up to 1 for each participant, and participants are assigned to the class for which their membership probability is the highest (Magidson and Vermunt 2004). The degree to which the highest membership probability differs from 1 can be regarded as classification error. In the study from the example above, the fit statistics showed that according to the BIC the three-class model showed the best balance between fit and parsimony, but according to the AIC3 the five-class model showed the best balance between fit and parsimony. Two classes were very similar in terms of interpretation in the three- and five-class models. The three additional classes of the five-class model made a further distinction in one of the classes of the three-class model. Using a combination of these fit statistics and substantive interpretation of the models, the model with five latent classes was chosen for interpretation (Koppenol-Gonzalez et al. 2014).

The next step is to identify the factors on the basis of which the latent classes can be differentiated. In order to do so, one needs to have developed hypotheses about potential factors on the basis of background knowledge. These hypotheses can be tested in two different ways. The first is to include the suggested factors as covariates in the LCRA directly. In this way class description is improved, as well as the classification of new cases (Magidson and Vermunt 2005). This procedure has several weaknesses, however; for example, it introduces new model-building problems and becomes impractical with a large number of covariates (see Vermunt 2010, 451). The alternative is to follow a three-step procedure. First, the latent class regression model is estimated. Second, cases are assigned to the class with the highest class posterior membership probability (Magidson and Vermunt 2004). Finally, one tries to identify the factors predicting class membership with the help of the appropriate statistical test. In the study of Koppenol-Gonzalez and colleagues (2014), the latter approach was applied. Each child was assigned to one of the five latent classes leading to a certain class size, or put differently, an overall probability of belonging to each of the latent classes. Each class was then characterized by a performance score. This score, together with the indicators of the STM processes (i.e., similar-

ity and position of the pictures), enabled the researchers to interpret each of the latent classes as follows: verbal STM with high performance (class 1), verbal STM with low performance (class 2), visual STM with high performance (class 3), visual STM with low performance (class 4), and overall low performance without a clear distinction between verbal and visual STM use (class 5). Note, however, that this three-step procedure underestimates the associations between the factors and the latent classes as a result of the classification errors in the second step (Bolck, Croon, and Hagenaars 2004), and therefore Vermunt (2010) has proposed two improved three-step procedures (see also Bakk, Tekle, and Vermunt 2013).

We have now described the various features of latent classes regression analysis. In the next section, we use class features to extrapolate causal conclusions following from an experiment.

Approaching the Problem of Extrapolation with Latent Classes. To extrapolate experimental results by means of Guala's analogical reasoning approach, LCRA can be of help. First, one examines whether there are latent classes in the sample that differ with respect to the effect of the predictors on the response variable. Note that this is possible without invoking strong theoretical assumptions about the latent variable(s), although in order to decide on the number of latent classes, it is useful to have defined hypotheses in this regard. Second, one seeks to identify the factors responsible for partitioning the sample into latent classes, which are then characterized by a combination of features. Finally, in order to extrapolate an effect, features of the classes are described as detailed as possible and then compared with features of the target population.

For example, the acute respiratory distress syndrome is defined by the clinical criteria of pulmonary opacities on a chest radiograph, arterial hypoxemia, and exclusion of cardiac failure as a primary cause (Calfee et al. 2014). Calfee and colleagues (2014) used LCA to explore whether multiple classes of the syndrome could be identified, using both baseline clinical data and biomarker levels. In two separate samples ($n = 473$ and $n = 549$), LCA showed that a two-class model provided the best fit to the data. These two classes of patients differed in plasma levels of inflammatory biomarkers, heart rate and total minute ventilation, serum bicarbonate, vasopressor use, prevalence of certain risk factors (i.e., sepsis and trauma), and ventilator-free and organ-failure-free days. Furthermore, one of the two samples came from a randomized controlled trial in which the treatment of mechanical ventilation with higher versus lower positive-end expiratory pressure (PEEP) was investigated. It turned out that within class 1, the mortality rate was higher after high PEEP (24%) than after low PEEP (16%), while in class 2, the mortality rate was higher after low PEEP (51%) than after high PEEP (42%). Also, this interaction effect between PEEP strategy and latent class was even

stronger for the outcome variables ventilator-free days and organ-failure-free days.

Now in order to extrapolate these results according to the method developed here, latent class features are compared to features of the target population. The claims to be extrapolated are conditional sentences. For instance, if an individual in the target has high plasma levels of inflammatory biomarkers, frequent vasopressor use, low serum bicarbonate, a high prevalence of sepsis, high heart rate and high total minute ventilation, and relatively few ventilator-free and organ-failure-free days, it is likely that the individual belongs to class 2, and high PEEP is recommended. By contrast, if these baseline measures and biomarker levels indicate that the individual belongs to class 1, low PEEP is the best option. In sum, the higher the similarities between the predicates in the antecedent and the features of the target, the stronger the basis for extrapolation.

As another example of LCRA, Magidson and Vermunt (2005) discuss a case study in which 400 persons were asked to indicate their 'purchase intent' regarding eight products (such as shoes) that differed in the following attributes: fashion (traditional vs. modern), quality (low vs. high), and price (lower vs. higher). The variables sex and age were included as covariates in the LCRA. One- to four-class models were estimated with and without covariates, and the three-class model with the covariates fitted the data best. People in the first latent class had a preference for modern, lower-price products. The attribute quality was not relevant in this class. The second latent class favored modern, lower-price, high-quality products. The third latent class was inclined to buy lower-price, high-quality products and was not influenced by the attribute fashion. Furthermore, the first class included a higher proportion of females than the other two latent classes; in the second class males were overrepresented. The first class also had a higher proportion of young people than the other two classes, and older people more often belonged to the third class than to the other two classes. Now if an individual in the target is a young female, then chances are high that she is inclined to buy cheap, modern products, without considering their quality. If an individual in the target is an older person, he or she will probably have a preference for cheap but high-quality products, regardless of their trendiness.

Note, however, that latent classes are not always as easy to interpret as in the example above (for another example, see Liew, Howe, and Little 2016). If this were the case, then LCRA would not even be necessary because the variables sex and age could just as well be included as predictors in a standard regression analysis (see Rothwell 2005). To illustrate the additional advantages of LCRA, consider the following study by Bouwmeester and Verkoeijen (2011). They investigated the testing effect: the phenomenon that taking a test of previously studied material improves retrieval more

than restudying. In a within-subjects design with the factor study method (restudy vs. test), 131 children studied 12 lists of semantically related words either through restudying of through taking a test. Each list had a strong backward association with one semantically related word ("lure") that was not in the list. One week later they took a final recognition test consisting of previously studied words, unrelated 'distractors', and the semantically related words. The false recognition of semantically related words was used as an indication of *gist trace processing*: the ability to form a gist memory of the elements that the words had in common. An LCRA found three latent classes within the sample. In the first class a large testing effect emerged: tested words were recognized much more often than restudied words. In the second class a small testing effect was observed, and in the third class no testing effect occurred. Furthermore, the latent classes differed in gist trace processing. In the second class, the false recognition of lures in the testing condition was lower than in classes 1 and 3. By contrast, the third latent class showed stronger gist trace processing in the restudy condition than the other two classes. This example demonstrates the additional advantages of LCRA as compared to the predefined groups in a standard regression analysis: the classes are interpreted on the basis of specific experimental tasks, and they are then externally validated by a new set of scores (i.e., the lures).

In order to extrapolate these results (Bouwmeester and Verkoeijen 2011) according to the method expounded in this article, one compares features of the latent classes to features of the target population. For children in the target who draw on gist trace processing when restudying, restudy might be at least as effective as testing. For children in the target who do rely on gist trace processing when they are tested but not when they restudy, in contrast, testing will probably be more effective than restudying. This approach to extrapolation is in line with Guala's theory. His theory states that we need to rely on the correspondence between observed features of the target and observed features of the experimental system (Guala 2003). In the present proposal, the observed features of the experimental system are the features of the latent classes. The more analogies between them and the features of the individuals in the target population and the more the experimental results correspond to observational data (Guala 2003, 2005), the more reliable the extrapolation of the causal relationship at issue.

According to this approach, it is not necessary to reveal the mechanism connecting cause and effect. In the purchase intent example above, gender and age are not part of a causal mechanism, but they do form the relevant features by which the latent classes are described. It is possible, but not necessary, that these features come out as components of a mechanism connecting X and Y. For example, gist trace processing can be part of a mechanism explaining the effect of testing on retrieval. The factors determining class

membership can be derived from a mechanistic explanation for the effect of X on Y, which LCRA may help to uncover. However, it is not necessary that the features of the classes are part of a causal mechanism in order to extrapolate the results from latent classes to subpopulations.

Causally Relevant Differences. As discussed in the introduction, a subgroup as defined by Cartwright (2007) has a fixed arrangement of causal factors other than the experimental treatment X. It is a maximally homogeneous group of individuals, in which all causes of Y other than X are held fixed at a certain level. This entails that within a subgroup, the effect of X on Y is also fixed. A latent class, however, is not maximally homogeneous. That is, a latent class consists of individuals who are similar but not equal with respect to the effect of X on Y. To some extent a latent class is homogeneous, but it does allow for some variation in the effect and X on Y and, accordingly, in the arrangements of other causal factors. This implies that, because not all other causes of Y are necessarily held fixed within a class, a latent class may consist of several subgroups.

Now if an experimental study yields a positive effect, this means that X causes Y in some subgroup of the sample. The effects of X in other subgroups may be very different. According to Cartwright (2007), to extrapolate the conclusion that X causes Y in some members of the target population, one needs to assume that at least one of the subgroups of the sample in which X causes Y is causally identical to a subgroup of the target. However, since we lack information about the target population, we cannot find out whether the assumption has been met or what the effect is of X on Y in other subgroups of the target.

LCRA is able to discover latent *classes* for which the predicted effect of X on Y differs in the sample. We describe the classes on the basis of some features and extrapolate the class-specific effects to individuals in the target population who meet the description of one of the classes. Since not all causal factors are identified and classes cannot be described exhaustively, it is possible that by applying this method, an effect is extrapolated to a person in the target whose subgroup was not represented in the sample. To take the purchase intent example, the first class had a preference for modern, lower-price products, regardless of quality. In this class, young females were overrepresented. On the basis of these findings, a young female in the target is predicted to buy modern, lower-price products. However, it is possible that the young female in the target has a high socioeconomic status (SES), while the young females in the first class of the sample had low/average SES. If SES is a relevant causal factor for shoe purchase intent, then the subgroup in the target to which the young female with high SES belongs is not represented in the sample, and the extrapolation may not be justified.

By extrapolating the result from the first class in the sample to the young female in the target, a causal relation is inferred to be present in a member of a subgroup that was not part of the sample. However, because many causal factors have not been identified as such (like SES in the example), this extrapolation is acceptable according to the present proposal. That is, a subgroup has a fixed arrangement of other causal factors, of which some are unknown to us. We are therefore unable to identify all subgroups in the sample, and we do not know whether all subgroups in the target are represented in the sample. As a result, it might happen that a relation is extrapolated to an individual whose subgroup was not causally identical to that of the sample, but we are unable to check whether this is the case.

The only necessary condition for extrapolation according to this proposal is that the individuals in the target meet the description of the appropriate class. This implies that inferences from one population to another are allowed. As with any inductive method, these inferences are uncertain; the purchase intent of the young female with high SES in the target might be different from the purchase intent found in the first class. It is therefore essential to characterize the classes as accurately as possible and to try to check whether the known differences between the sample and the target can confound the extrapolation.

Whether extrapolation is a sensible option depends on the nature of the explanatory variable, the number of replications of the study, the differences between the effect sizes in the samples, and the degree to which the classes are defined. Moreover, an important concern is whether the classification that emerges from an LCRA is stable. It is possible that the number of classes varies between studies, and in that sense LCRA suffers from the same shortcomings as standard regression analysis. In order to determine whether the number and the specification of the classes are robust, one therefore needs to perform replication studies (Ioannidis 2005; Schmidt 2009; Cumming 2012; Open Science Collaboration 2012; Pashler and Wagenmakers 2012; Klein et al. 2014).

Conclusion. According to Guala, the more an experiment resembles the natural situation, the more likely is it that the extrapolation in question is reliable. If there are sufficient parallels between the properties of the test system and the target system, and the experimental data and the observational data are similar, then the causal relationship under study will probably obtain in the target system. It is not required that these properties are part of a causal mechanism. If a mechanism is identified, it can be valuable for extrapolation purposes. However, for many phenomena in the human sciences, there is no mechanistic explanation available.

The method developed here is a refinement of Guala's theory. LCRA is able to discover classes in the sample for which the effect of the explanatory

variable varies. In order to extrapolate by means of analogy but without relying on mechanistic knowledge, the existence of latent classes should be checked and interpreted with the help of background knowledge. The latent classes are then described by a combination of features. The causal relations being extrapolated are formulated as a set of conditional sentences in which features of the classes are in the antecedent, and the direction and size of the effect are in the consequent. The more analogies between the features of the classes and features of individuals in the target, the stronger the case for extrapolation.

The only requirement for extrapolation according to this proposal is that individuals in the target meet the description of the appropriate class. Because we do not know all other causal factors and we cannot describe the classes exhaustively, it is possible that these individuals differ from the tested individuals in causally relevant ways. This means that, just like Steel's proposal (Howick et al. 2013), we cannot fully avoid the extrapolator's circle. Therefore, we should define the classes carefully and try to discover whether differences between the test and the target can confound the inference. Given the inductive framework, this is all that we are capable of doing.[3]

REFERENCES

Angrist, Joshua D., and Jörn-Steffen Pischke. 2010. "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics." *Journal of Economic Perspectives* 24:3–30.

Bakk, Zsuzsa, Fetene B. Tekle, and Jeroen K. Vermunt. 2013. "Estimating the Association between Latent Class Membership and External Variables using Bias-Adjusted Three-Step Approaches." *Sociological Methodology* 43:272–311.

Bolck, Annabel, Marcel Croon, and Jacques Hagenaars. 2004. "Estimating Latent Structure Models with Categorical Variables: One-Step versus Three-Step Estimators." *Political Analysis* 12:3–27.

Bothwell, Laura E., Jeremy A. Greene, Scott H. Podolsky, and David S. Jones. 2016. "Assessing the Gold Standard: Lessons from the History of RCTs." *New England Journal of Medicine* 374:2175–81.

Bouwmeester, Samantha, and Peter P. J. L. Verkoeijen. 2011. "Why Do Some Children Benefit More from Testing than Others? Gist Trace Processing to Explain the Testing Effect." *Journal of Memory and Language* 65:32–41.

3. One of the authors of the current article has expressed skepticism about the usefulness of the 'internal validity (X causes Y in the experimental population) versus external validity (X causes Y also in the target population)' distinction (Reiss 2018). In brief, describing reasoning about target systems of interests in this way suggests that in order to learn a causal relationship in a target, an experimental analogue of that system must first be identified and studied. But this is not the case, and Reiss (2018) discusses numerous aids for causal inference about a target system without the prior identification of an analogue system. In this article we address the reverse question: once we have data that draw on an experimental system, what can we infer from these data? LCRA is a powerful method for increasing our inferential abilities for these kinds of cases.

Calfee, Carolyn S., Kevin Delucchi, Polly E. Parsons, B. Taylor Thompson, Lorraine B. Ware, and Michael A. Matthay. 2014. "Subphenotypes in Acute Respiratory Distress Syndrome: Latent Class Analysis of Data from Two Randomised Controlled Trials." *Lancet Respiratory Medicine* 2:611–20.

Cartwright, Nancy. 2007. "Are RCTs the Gold Standard?" *BioSocieties* 2:11–20.

Cumming, Geoff. 2012. *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-analysis*. Melbourne: Routledge.

Danziger, Kurt. 1987. "Statistical Method and the Historical Development of Research Practice in American Psychology." In *The Probabilistic Revolution*, vol. 2, *Ideas in the Sciences*, ed. Lorenz Krüger, Gerd Gigerenzer, and Mary S. Morgan, 35–47. Cambridge, MA: MIT Press.

Deaton, Angus, and Nancy Cartwright. 2018. "Understanding and Misunderstanding Randomized Controlled Trials." *Social Science and Medicine* 210:2–21.

Elster, Jon. 1998. "A Plea for Mechanisms." In *Social Mechanisms: An Analytical Approach to Social Theory*, ed. Peter Hedstrom and Richard Swedberg, 45–73. Cambridge: Cambridge University Press.

Gerring, John. 2010. "Causal Mechanisms: Yes, But . . ." *Comparative Political Studies* 43:1499–526.

Grice, James W. 2015. "From Means and Variances to Persons and Patterns." *Frontiers in Psychology* 6:1007.

Guala, Francesco. 1998. "Experiments as Mediators in the Non-laboratory Sciences." *Philosophica* 62:57–75.

———. 2003. "Experimental Localism and External Validity." *Philosophy of Science* 70:1195–205.

———. 2005. *The Methodology of Experimental Economics*. New York: Cambridge University Press.

———. 2010. "Extrapolation, Analogy, and Comparative Process Tracing." *Philosophy of Science* 77:1070–82.

Hagenaars, Jacques A., and Allan L. McCutcheon. 2002. *Applied Latent Class Analysis*. Cambridge: Cambridge University Press.

Hamaker, Ellen L. 2012. "Why Researchers Should Think 'Within-Person': A Paradigmatic Rationale." In *Handbook of Research Methods for Studying Daily Life*, ed. Matthias R. Mehl and Tamlin S. Conner, 43–61. New York: Guilford.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81:945–60.

Howick, Jeremy, Paul Glasziou, and Jeffrey K. Aronson. 2013. "Problems with Using Mechanisms to Solve the Problem of Extrapolation." *Theoretical Medicine and Bioethics* 34:275–91.

Ioannidis, John P. A. 2005. "Why Most Published Research Findings Are False." *PLoS Medicine* 2:696–701.

Kagel, John, H., and Dan Levin. 1986. "The Winner's Curse and Public Information in Common Value Auctions." *American Economic Review* 76:894–920.

Kent, David M., Rodney A. Hayward, John L. Griffith, Sandeep Vijan, Joni R. Beshansky, Robert M. Califf, and Harry P. Selker. 2002. "An Independently Derived and Validated Predictive Model for Selecting Patients with Myocardial Infarction Who Are Likely to Benefit from Tissue Plasminogen Activator Compared with Streptokinase." *American Journal of Medicine* 113:104–11.

Klein, Richard A., et al. 2014. "Investigating Variation in Replicability: A 'Many Labs' Replication Project." *Social Psychology* 45:142–52.

Koppenol-Gonzalez, Gabriela V., Samantha Bouwmeester, and Jeroen K. Vermunt. 2014. "Short-Term Memory Development: Differences in Serial Position Curves between Age Groups and Latent Classes." *Journal of Experimental Child Psychology* 126:138–51.

Lamiell, James T. 2009. "Reviving Person-Centered Inquiry in Psychology: Why It's Erstwhile Dormancy." In *Dynamic Process Methodology in the Social and Developmental Sciences*, ed. Jaan Valsiner, Peter C. M. Molenaar, Maria C. D. P. Lyra, and Nandita Chaudhary, 31–43. New York: Springer.

Lanza, Stephanie T., and Brittany L. Rhoades. 2013. "Latent Class Analysis: An Alternative Perspective on Subgroup Analysis in Prevention and Treatment." *Prevention Science* 14:157–68.

Liew, Shi Xian, Piers D. L. Howe, and Daniel R. Little. 2016. "The Appropriacy of Averaging in the Study of Context Effects." *Psychonomic Bulletin and Review* 23:1639–46.

Magidson, Jay, and Jeroen K. Vermunt. 2004. "Latent Class Models." In *The Sage Handbook of Quantitative Methodology for the Social Sciences*, ed. David Kaplan, 175–98. Thousand Oaks, CA: Sage.

———. 2005. "A Nontechnical Introduction to Latent Class Models." *DMA Research Council Journal*, 2005, 1–15.

Mead, Walter J., Asbjorn Moseidjord, and Philip E. Sorensen. 1983. "The Rate of Return Earned by Lessees under Cash Bonus Bidding for OCS Oil and Gas Leases." *Energy Journal* 4:37–52.

Molenaar, Peter C. M. 2004. "A Manifesto on Psychology as Idiographic Science: Bringing the Person Back into Scientific Psychology, This Time Forever." *Measurement* 2:201–18.

Navarro, Daniel. J., Thomas L. Griffiths, Mark Steyvers, and Michael D. Lee. 2006. "Modeling Individual Differences Using Dirichlet Processes." *Journal of Mathematical Psychology* 50:101–22.

Open Science Collaboration. 2012. "An Open, Large-Scale, Collaborative Effort to Estimate the Reproducibility of Psychological Science." *Perspectives on Psychological Science* 7:657–60.

Pashler, Harold., and Eric-Jan Wagenmakers. 2012. "Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence?" *Perspectives on Psychological Science* 7:528–30.

Reiss, Julian. 2008. *Error in Economics: Towards a More Evidence-Based Methodology*. London: Routledge.

———. 2018. "Against External Validity." *Synthese*. doi:10.1007/s11229-018-1796-6.

Rothwell, Peter M. 1995. "Can Overall Results of Clinical Trials Be Applied to All Patients?" *Lancet* 345:1616–19.

———. 2005. "Treating Individuals." Pt. 2, "Subgroup Analysis in Randomized Controlled Trials: Importance, Indications, and Interpretation." *Lancet* 365:176–86.

Rousselet, Guillaume A., Cyril R. Pernet, and Rand R. Wilcox. 2017. "Beyond Differences in Means: Robust Graphical Methods to Compare Two Groups in Neuroscience." *European Journal of Neuroscience* 46:1738−48.

Schmidt, Stefan. 2009. "Shall We Really Do It Again? The Powerful Concept of Replication Is Neglected in the Social Sciences." *Review of General Psychology* 13:90−100.

Sears, David O. 1986. "College Sophomores in the Laboratory: Influences of a Narrow Data Base on Social Psychology's View of Human Nature." *Journal of Personality and Social Psychology* 51:515–30.

Simons, Daniel J., Yuichi Shoda, and D. Stephen Lindsay. 2017. "Constraints on Generality (COG): A Proposed Addition to All Empirical Papers." *Perspectives on Psychological Science* 12:1123–28.

Speelman, Craig P., and Marek McGann. 2013. "How Mean Is the Mean?" *Frontiers in Psychology* 4:451.

Steel, Daniel. P. 2008. *Across the Boundaries. Extrapolation in Biology and Social Science*. New York: Oxford University Press.

Szczech, Lynda A., Jesse A. Berlin, and Harold I. Feldman. 1998. "The Effect of Antilymphocyte Induction Therapy on Renal Allograft Survival. A Meta-analysis of Individual Patient-Level Data." *Annals of Internal Medicine* 128:817–26.

Vermunt, Jeroen K. 2010. "Latent Class Modeling with Covariates: Two Improved Three-Step Approaches." *Political Analysis* 18:450–69.

Wedel, Michel, and Wayne A. DeSarbo. 1994. "A Review of Recent Developments in Latent Class Regression Models." In *Advanced Methods of Marketing Research*, ed. Richard P. Bagozzi, 352–88. Cambridge, MA: Blackwell.