Synthese (2019) 196:4329–4350
https://doi.org/10.1007/s11229-017-1660-0

Accommodation, prediction and replication: model selection in scale construction

Clayton Peterson[1]

Received: 12 April 2017 / Accepted: 18 December 2017 / Published online: 6 January 2018

Abstract In psychology, measurement instruments are constructed from scales, which are obtained on the grounds of exploratory and confirmatory factor analysis. Looking at the literature, one can find various recommendations regarding how these techniques should be used during the scale construction process. Some authors suggest to use exploratory factor analysis on the entire data set while others advice to perform an internal cross-validation by randomly splitting the data set in two and then either perform exploratory factor analysis on both parts or exploratory factor analysis on the first part and confirmatory factor analysis on the other. In spite of all these divergent recommendations, there is no consensus on which method yields the best result. In this paper, we analyze this issue in light of the prediction versus accommodation debate and argue that the answer to this question depends on one's conception of the criteria that should be used to achieve the goals of the scientific enterprise.

Keywords Exploratory factor analysis · Confirmatory factor analysis · Latent variables · Goodness-of-fit · Akaike information criterion (AIC) · Measurement model

1 Conundrum at the office

This is the story of three social scientists, Annie, Penny and Rosie, who worked together in a psychology department. They wanted to capture empirically some hypothesized latent constructs. Seeing the importance of their enterprise, they decided to get together

✉ Clayton Peterson
clayton.peterson@outlook.com

[1] Munich Center for Mathematical Philosophy, Ludwig-Maximilians-Universität München, Munich, Germany

and collaborate. The objective of their task was to obtain a measurement model. They began by formulating the items meant to be indicators of the hypothesized latent constructs. When all the items were properly developed, they turned to the data gathering process. However, when they reconvened afterwards, they got into an argument. None could agree as to how they should analyze the data to provide an adequate measurement model. Seeing their disagreement, they went their separate way, each one of them following a different guideline proposed in the literature. While Annie applied exploratory factor analysis to the entire data set, Penny randomly divided the data set in two parts and applied exploratory factor analysis to the first part but confirmatory factor analysis to the second. Rosie, who liked the idea of randomly splitting the data set to perform two parallel analyses, had read somewhere that exploratory factor analysis should rather be applied to both parts. When their analyses were done, they met again in order to examine what they had. To their surprise, they all obtained a different measurement model, with comparable but different factorial structures. But which one should they use for the paper they were supposed to write? They had to collaborate: It was *their* data. So they went to see Sara, the supervisor of their section, hoping she would be able to resolve the situation. After explaining their situation to Sara, they asked: *But which measurement model should we use? Which one is the best?* Sara did not know what to answer. She knew that each strategy was a plausible alternative discussed within the literature, but she did not know which one, if any, was best. So she asked for some time to think about the situation. She went home that night and discussed that problem with her friend, Piper the philosopher. Here is what Piper had to say.

2 Measurement models

One aim of psychological research is to identify latent constructs that can be used to explain psychological phenomena (cf. Velicer and Jackson 1990b). These latent constructs are identified through various measurement instruments. Measurement instruments in psychology are constructed from scales, which in turn are built from items. Latent constructs are identified indirectly via subjects' responses on these items. Items can take many forms (cf. Mellenbergh 1994). For instance, items can be dichotomous (e.g. yes or no), Likert scales (e.g. 1 = strongly disagree, 2 = disagree, 3 = neither agree nor disagree, 4 = agree and 5 = strongly agree) or points on a continuous line. There are four types of scales used in psychological measurement, which can be divided in two categories: Scales that produce categorical data and scales that produce continuous data (cf. Stevens 1946). Among the scales that yield categorical data are nominal scales (i.e. assigning numbers to individuals or concepts) and ordinal scales (which preserve order, e.g. a Likert scale). Continuous data is produced by interval (where 0 is a convention) and ratio scales.

The rationale behind scale construction is the conception that latent constructs can be used to causally explain the between-subjects variation on the items.[1] For

[1] This paper is primarily concerned with the issue of model selection during scale construction in psychology. With that being said, our analysis should not be considered as restricted only to that field. Our analysis

simplicity, consider an example with only one factor. Assume that a researcher thinks that there is some latent construct, say, *emotional exhaustion*, that can be used to explain some psychological phenomenon (e.g. Maslach and Jackson 1981). How would this researcher capture such latent construct? The researcher might postulate that if one is experiencing emotional exhaustion, then one is likely to score high on items such as "I feel emotionally drained from my work" and "I feel used up at the end of the workday". The researcher will thus develop many items that are meant to be indicators of the latent construct and then gather some data using these items. Using exploratory and/or confirmatory factor analysis techniques, the researcher will identify a subset of items that are good indicators of the factor. The scale used to identify emotional exhaustion will consist of the resulting subset of items. From this, the researcher obtains a measurement model. In this case, exploratory and/or confirmatory factor analysis is used to determine whether or not the between-subjects variation on the items can be modeled by a single factor, which is interpreted as emotional exhaustion. As a result, the researcher obtains a measurement model where the latent factor is taken as a causal explanation of the between-subjects variation on the items (see also Spirtes et al. 1991, 2000; Glymour 2001). In the aforementioned example, it is hypothesized that if one is experiencing (resp. not experiencing) emotional exhaustion, then one will score high (resp. low) on the items.[2] Hence, *emotional exhaustion* causally explains why some will score high while others will score low on the items. In this measurement model, one considers that there is a between-subjects variation on the items because some are experiencing emotional exhaustion while others are not.[3] This type of measurement model is known as a *reflective* measurement model (cf. Edwards and Bagozzi 2000).

It is important to emphasize that these measurement models are not meant to causally explain the *within*-subject variation on the items (cf. Borsboom et al. 2003; Borsboom 2008). They are not meant to indicate where an individual stands on a latent variable. Returning to the aforementioned example, the scale is not meant to measure to which extent one is experiencing emotional exhaustion. Measuring the variation on a latent variable would require a longitudinal design in which one observes the correlation between 'emotional exhaustion' at time 1 and at time 2, for example. Longitudinal designs are necessary to examine the variation on the latent variable. However, the reflective measurement model is obtained on the grounds of a cross-sectional study (with data gathered at only one point in time). In this respect, despite the terminology, measurement models are meant to identify latent variables rather than *measure* them per se. Instead of measuring to which extent one is emotionally exhausted (which would require a longitudinal research design), researchers identify this construct through its plausible empirical effects (i.e. the between-subjects variation on the items).[4]

Footnote 1 continued
can be extended to other disciplines that use factor analytic methods for the development of measurement instruments (e.g. sociology, education), although some idiosyncratic differences might occur.

[2] Some items might be reversed score items.

[3] See Hood (2013) for an analysis of the issue of realism with respect to measurement models.

[4] In such a measurement model, the items are not used to predict the latent factors. As we will see, a factorial structure is a space where the items are the points.

3 Scale construction

In the psychology literature, there are various recommendations regarding how one should proceed to construct a scale. To exemplify these recommendations, let us borrow and extend an example from Hitchcock and Sober (2004, p. 8). Consider three researchers, namely Annie, Penny and Rosie. Let D be a data set that can randomly be divided in two parts D_1 and D_2 (with $D = D_1 \cup D_2$). Annie uses an iteration of exploratory factor analysis (EFA) on D to identify the measurement model. This is the general procedure that is recommended for scale construction (e.g. Costello and Osborne 2005). In contrast, Penny uses an iteration of EFA on D_1 and then uses confirmatory factor analysis (CFA) on D_2 to confirm the measurement model (e.g. Comrey 1988; Hinkin 1998; Fabrigar et al. 1999). Finally, Rosie uses an iteration of EFA on D_1 and then tries to replicate the results on D_2 using EFA (e.g. Thompson 1994, 2004).[5] The iterative use of EFA is meant to identify a measurement model in which subsets of items are taken as good indicators of the factors. There is thus an issue of model selection. As a result of the analysis, one will choose a measurement model that may not use all the items that were initially present in the data set (cf. Sect. 5).

4 Prediction, accommodation and overfitting

The prediction versus accommodation debate is an old one, which has received the attention of many philosophers (see Maher 1988; White 2003 for an overview and references). In its initial formulation, the predictivist thesis amounted to the position that novel (predicted) data provide better support in favor of a hypothesis than previously existing (thus accommodated) data (cf. Maher 1988, p. 273). Hitchcock and Sober (2004) provided an interesting analysis of the prediction versus accommodation debate.[6] Starting from the premise that science should aim at predictive accuracy (on this point, see also Forster and Sober 1994; Forster 2002; Sober 2004), they argue that accommodation risks overfitting, which undermines predictive accuracy. A model overfits the data when it incorporates idiosyncratic properties of the sample. Hence, if science aims (among other things) at predictive accuracy, then researchers should insure that overfitting does not happen.

The elegance of Hitchcock and Sober's proposal and its relevance to model selection during scale construction consist in the fact that it relies on clear conceptual clarifications regarding the key concepts involved. Among the concepts central to the debate between prediction and accommodation is the notion of *novel* data. Following Musgrave (1974), Hitchcock and Sober (2004, pp. 4–5) distinguish between

[5] Note that Penny's and Rosie's partitions of D can be different.

[6] Hitchcock and Sober's analysis takes place within the context of an instrumentalist perspective and, as such, they consider that science should aim at predictive accuracy (see Hitchcock and Sober 2004, pp. 2–3). It should be noted, however, that our analysis is independent of the realism/instrumentalism debate. As we will see, our analysis only relies on the assumption that empirical adequacy is, among others, one goal of the scientific enterprise, and that predictive accuracy is, among others, an indicator of the overall empirical adequacy of a theory.

temporal, *heuristic* and *theoretical* novelty. Temporal novelty refers to data that was not available at the time the theory (model, hypothesis) was formulated; theoretical novelty concerns data that was not predicted by the theory's (model's, hypothesis's) rivals and heuristic novelty simply regards data not used for the construction of the theory (model, hypothesis).[7] Hitchcock and Sober adopt heuristic novelty, also known as use-novelty (cf. Worrall 2002). Heuristic novelty is particularly relevant to model selection given that it allows considering data that was not used in previous analyses (cf. Schurz 2014). On these grounds, the predictivist thesis can be formulated more broadly without reference to the time at which the data was obtained. It thus amounts to the assumption that data provide stronger support in favor of a hypothesis if it was predicted rather than accommodated. From this conception, predicted data can be already available data, as long as it is not used for the formulation of the hypothesis (model, theory).

Heuristic novelty is relevant for factor analysis given that it allows one to consider mathematical regressions as predictions in the context of the prediction versus accommodation debate. Considering that time is taken out of the equation, one can legitimately call the independent variable a *predictor* in a regression analysis insofar as the predictor does not necessarily need to be measured at a time prior than the criterion (dependent variable).

In addition, Hitchcock and Sober (2004, p. 3) distinguish between *global* and *local* predictivism. While global predictivism implies that prediction is always better than accommodation, local predictivism only stipulates that in some contexts it is. Further, they distinguish between *strong* and *weak* predictivism. Advocates of strong predictivism defend that prediction is better than accommodation per se, whereas weak predictivism asserts that the superiority of one concept over the other depends on external epistemically relevant characteristics (Hitchcock and Sober 2004, p. 4). The same conceptualization applies to global/local and strong/weak accommodationism. Hitchcock and Sober's position can therefore be described as local weak predictivism and local weak accommodationism. In some contexts prediction is better, in other contexts accommodation is better, and what makes one better than the other is external: It depends on the extent to which it safeguards against overfitting.

The key concept central to their analysis is *overfitting*. Prediction can be better than accommodation insofar as it protects against overfitting and it can be used to show that overfitting has not occurred. Nonetheless, accommodation can be better than prediction when one makes sure that overfitting does not happen (Hitchcock and Sober 2004, p. 3). In order to make explicit the relationships between prediction and accommodation, Hitchcock and Sober provide a fictional example to guide their analysis:

> Consider two scientists, Penny Predictor and Annie Accommodator. In the course of their investigations, they acquire identical sets of data. Let us call this data D. Both researchers advance theories on the basis of this data, but there is a difference. Penny formulates her theory after acquiring the initial data fragment D_1, and then uses her theory to make predictions about the remaining data. Her

[7] See also Gardner (1982) for a discussion of novelty.

> theory predicts the remaining data D_2 (where $D = D_1 \cup D_2$) to a high degree of accuracy. By contrast, Annie formulates her theory only after she has acquired all of the data in D, constructing her theory deliberately so as to accommodate this data (Hitchcock and Sober 2004, p. 8).

In light of this example, the question they ask is thus:

> Does Penny's theory, by virtue of accurately predicting data D_2, enjoy a privileged epistemic status that Annie's does not? [. . .] Suppose that Penny constructs a theory T_p that successfully predicts data D_2, whereas Annie constructs a (possibly different) theory T_a, by accommodating the entire data set D. Does the fact that T_p predicted D_2 whereas T_a was designed to accommodate this data give us reason to believe that T_p is the better theory (Hitchcock and Sober 2004, p. 8)?

To answer this question, they distinguish between four strategies that can be used by Penny and Annie (Hitchcock and Sober 2004, pp. 31–32):

I Accommodation of D by maximizing fit to the data;
II Accommodation of D_1 by maximizing fit to the data and prediction of D_2;
III Accommodation of D using AIC;
IV Accommodation of D_1 using AIC and prediction of D_2.

Even though Hitchcock and Sober's presentation uses the Akaike Information Criterion (AIC; Akaike 1987) as a criterion for model selection, it should be emphasized that their argument does not rely on AIC per se but rather relies on any criterion that enables to balance goodness-of-fit and simplicity in order to avoid overfitting (e.g. the Bayesian Information Criterion BIC; cf. Hitchcock and Sober 2004, p. 14). The relevance of a criterion such as AIC for model selection comes from the fact that it shows the relationship between goodness-of-fit and complexity (see Sect. 5 below). This relationship can be understood in light of the curve fitting problem. In a nutshell, the curve fitting problem consists in the fact that simplicity usually comes at the price of goodness-of-fit (Forster and Sober 1994, p. 2). More complex models will tend to have better fit given that there is always a model that can fit exactly all the points in a data set (Hitchcock and Sober 2004, p. 11). As such, the more complex the model is, the higher fit it will tend to have. However, the more complex the model is, the more likely it will overfit the data. Overfitting happens with complex models insofar as when there are more adjustable parameters, it is more likely that the model will fit to the noise in the data (Forster and Sober 1994, p. 8). Hence, a model that overfits the data will not be predictively accurate, though it will have a good fit. Basically, AIC insures that there is a balance between the complexity of the model and its goodness-of-fit, thus safeguarding against overfitting (Forster and Sober 1994, p. 11). The advantage of using AIC as a rule to select the best model is that it indicates whether or not it is worth selecting a more complex model (see also Forster 2002).

As a result of their analysis, Hitchcock and Sober argue that, in general, **III** is better than **IV**, which in turn is better than **II**, which is better than **I**. The ranking of **III** over **I** follows from the fact that, given the use of AIC, the model "will have the highest estimated predictive accuracy, relative to the entire data set D" and, further,

that it "balances simplicity against goodness-of-fit as to avoid overfitting (Hitchcock and Sober 2004, p. 31)." The same justification is used to rank **IV** over **II**. As to the ranking of **III** over **IV**, it follows from the principle of total evidence, which states that the best model should be obtained by "considering *all* the data available (Hitchcock and Sober 2004, p. 31)". The ranking of **II** over **I** is justified by the fact that prediction can be seen as a safeguard to overfitting. That is, an accurate prediction is an indicator that overfitting did not happen (cf. Hitchcock and Sober 2004, p. 18).

From these considerations, they conclude that prediction is better than accommodation when one cannot accommodate using a criterion that balances simplicity and goodness-of-fit (i.e. that safeguards against overfitting, such as AIC), whereas accommodation is better than prediction when one can use such a criterion, in which case this should be accomplished on as much data as possible.

5 Model selection

Although there are slight differences between the notions of accommodation and prediction as they are used in EFA/CFA and within Hitchcock and Sober's framework, their analysis provides us with a good starting point to determine which strategy is best during scale construction. Indeed, if Annie and Penny, as introduced in Sect. 3, can be compared to Annie and Penny in Hitchcock and Sober's case, then we might be able to apply their analysis to obtain a partial answer to Sara's problem. To determine if this is the case, however, we must first examine in what precisely consist EFA and CFA.

5.1 Exploratory factor analysis

EFA is comparable to (and is sometimes confused with) a similar but different statistical technique known as principal component analysis (see Jolliffe 2002). Principal component analysis is a data reduction technique that represents principal components as linear combinations of variables. As such, principal component analysis provides a *formative* measurement model (cf. Edwards and Bagozzi 2000). Principal component analysis, as a tool of descriptive statistics, is an accommodative technique that produces a unique solution. When all the principal components are retained, an analysis by principal components produces a model that perfectly fits the data. In this respect, principal component analysis is, by definition, prone to overfitting. One important aspect of using statistics in psychology, however, is to be able to take into consideration error and uncertainty (cf. Suppes 2007). Consequently, principal component analysis is not a suitable statistical tool to use during scale construction.

By comparison to principal component analysis, EFA models the *variables* as linear combinations of latent factors but includes error terms that account for possible error in measurement (cf. Jolliffe and Morgan 1992; Finch and West 1997; Park et al. 2002; Abdi and Williams 2010; Tabachnick and Fidell 2013; Johnson 2016). The terminology 'principal components' is used in the case of principal component analysis, whereas 'factor' is reserved for factor analysis. In factor analysis, factors refer to the axes of a n-dimensional space. A factorial structure with n factors is a n-dimensional space where

the items are the points. While the principal components describe *all* the variance in the data set, the factors only explain a portion of the variance, the rest of the variance being accounted for by the error terms.[8]

EFA, as a method of inferential statistics, is an accommodative procedure that is constrained by methodological decisions. The first step of an EFA is factor extraction: One has to determine the number of factors that will be used to model the data. Various criteria can be used to accomplish this (cf. Velicer and Jackson 1990a; Finch and West 1997), including parallel analysis. Parallel analysis, which was initially developed for principal component analysis (cf. Horn 1965), has been adapted for EFA. It retains factors that are statistically significant, the null hypothesis being that the factors come from a randomly generated data set (cf. O'Connor 2000). Parallel analysis is considered as one of the most objective criteria for factor selection. In addition to the choice of the number of factors to retain, the second step of an EFA is the choice of a factor extraction method. The most widely used extraction technique is perhaps maximum likelihood extraction (cf. Tabachnick and Fidell 2013), which relies on the assumption that the variables are continuous[9] and normally distributed[10] (Byrne 2012).[11] After extraction, the factorial structure is rotated to a simple structure (cf. Thurstone 1947, 1954). There are various rotation methods that can be used to accomplish this, including orthogonal and oblique rotations. Orthogonal rotations constraint factors to be uncorrelated while oblique rotations allow for correlation between factors.

During scale construction, EFA is used as an iterative procedure (cf. Mulaik 1991). The iterative use of EFA aims at obtaining a factorial structure where subsets of variables are highly correlated with only one factor (i.e. a simple structure). Factor loadings greater than .3 are considered to be significant (cf. Field 2009, p. 644). As it is usually recommended (e.g. Hinkin 1998), researchers start with a large pool of items and then use EFA to reduce these items to smaller subsets that are good indicators of the factors. After performing EFA on the data set, researchers eliminate items with double loadings (i.e., which between-subjects variation can be explained via more than one factor). This process is repeated until a proper factorial structure is obtained. Hence, there is an issue of model selection.

The choice of a measurement model in the context of scale construction can be justified through various criteria that are meant to achieve the goal(s) of the scientific enterprise (cf. Kuhn 1983). Though realists might argue that the ultimate goal of science is truth, everyone should agree with the weaker claim that, perhaps among

[8] Conceptually, part of the unexplained variance might be due to something different than error.

[9] Even though the hypothesis of continuity is often violated, it has been shown that, under the assumption of normality, violation of the assumption of continuity is likely negligible (cf. Byrne 2012).

[10] There are other extraction techniques that can be used in cases where the assumption of normality is violated (cf. Flora et al. 2012). Consequently, despite their importance, these two conditions are not necessary for the use of factor analysis during scale construction.

[11] It has been argued by Michell (1997) that, given the violation of the assumption of continuity, measurement of psychological attributes is not possible (see also Michell 2003, 2004). His argument revolves around the assumption that only continuous quantities can be measured and, since the hypothesis that psychological attributes are continuous is not tested, it follows that one is not justified to believe that it is possible to measure psychological attributes. Borsboom and Mellenbergh (2004) answered this objection and showed that this hypothesis is actually tested, though not in isolation.

others, one goal of science is empirical adequacy (e.g. van Fraassen 1980). The choice of a measurement model can thus be justified via specific criteria that are meant to insure that psychological research is empirically adequate.

Maximizing goodness-of-fit can be seen as a criterion meant to achieve empirical adequacy. By maximizing the goodness-of-fit, one is maximizing the extent to which a measurement model fits the points in the data set. There are limits, however, to the idea that maximizing goodness-of-fit can be used to reach empirical adequacy, for maximizing goodness-of-fit will tend to provide a model that overfits the data. Maximizing goodness-of-fit should therefore not be seen as a criterion that, alone, is sufficient to reach empirical adequacy.

In the context of factor analysis with maximum likelihood factor extraction, AIC serves as a badness-of-fit index (Akaike 1987, p. 318). It is given by $(-2)max\ log\ likelihood + 2k$, with k the number of parameters. If maximum likelihood extraction has not been used, then the log likelihood can be used instead of the maximum log likelihood (Akaike 1987, p. 322). The use of AIC for selecting models in EFA regards the problem of factor retention (Akaike 1987; Preacher et al. 2013). In an EFA model, the factors are the adjustable parameters. Accordingly, choosing the model with the lowest AIC score insures that goodness-of-fit does not come at the price of complexity and, therefore, it can be seen as a safeguard to overfitting.

For example, assume that a researcher starts with i items. Different EFA models with $1, \ldots, n$ factors can be used to model this data. Using AIC as a criterion for model selection in this context amounts to choose the EFA model with the lowest AIC score. The choice of the number of factors to retain, however, should not only be a matter of AIC. In addition to theoretical considerations, the researcher should also verify with parallel analysis whether or not the factors are statistically significant (cf. O'Connor 2000). After choosing the proper number of factors, the researcher will look at the covariation matrix to see whether some items are poor indicators of the factors (e.g. items with multiple loadings). The researcher will exclude these items and repeat the process until she obtains adequate subsets of indicators. To some degree, Annie, Penny and Rosie all use this strategy, the main difference being that Annie applies this strategy to the entire data set while Penny and Rosie only use it with respect to D_1.

As a result of an EFA, a researcher gets a measurement model indicating which factors can be seen as potential causal explanations of the between-subjects variation on subsets of items.[12] Formally, this measurement model takes the form of a covariation matrix, representing items as linear combinations of factors together with an error term. As such, an EFA produces a covariation matrix that specifies which items are indicators of which factors. Under the assumption of (i) a specific number of factors to extract, (ii) a factor extraction technique and (iii) a rotation method, EFA is an accommodative procedure that yields a unique solution. Put differently, given a specific number of factors to retain, a factor extraction method and the choice of an

[12] It should be mentioned that some authors argued that an algorithm that searches for pure measurement models should be used rather than factor analysis techniques to obtain a measurement model (e.g. Silva et al. 2006; Kummerfeld et al. 2014; see also Murray-Watters and Glymour 2015). The analysis of these alternative techniques is beyond the scope of this paper. We will concentrate on factor analysis given its place with regard to scale construction in the psychology literature.

appropriate rotation, EFA tells researchers which items load on which factor(s) and yields the best measurement model that fits the data. EFA can be performed without prior hypothesis regarding the factorial structure that will obtain (Huck 2012). Even if one can have an *a priori* hypothesis regarding the factorial structure, this hypothesis bears no consequences on the result of the analysis. EFA is an accommodative procedure that tells how the data behave, modulo the number of factors and some noise that might be in the data set. So Annie is an accommodator. She uses all the data at her disposal to formulate a measurement model. In contrast to Penny and Rosie, however, the validation of that measurement model is not included in the procedure.

5.2 Confirmatory factor analysis

In contrast, CFA is a statistical technique where researchers specify *a priori* the measurement model and then determine to which extent the covariation matrix of the proposed measurement model fits the covariation matrix of the data. Put differently, researchers specify beforehand which items are supposed to load on which factor(s) and then determine using CFA whether the hypothesized factorial structure can be used to explain the variance in the data. In contradistinction to EFA (where prior hypotheses regarding the factorial structure are unnecessary), the hypothesis guiding the CFA is a necessary condition for the analysis to be performed.

CFA is a special case of structural equation modeling (cf. Schreiber et al. 2006; Ullman 2013). Basically, CFA can be seen as a combination of factor analysis and multiple regression analysis that allows a complete and simultaneous examination of all the postulated relationships between the variables (latent or not) (Ullman 2013, pp. 681, 684). A model in structural equation modeling can be divided in two parts (cf. Ullman 2013, p. 683). While the measurement model specifies the relationships between the observed variables (the items) and the latent variables (the factors), the structural model specifies relationships between factors. During scale construction and scale validation, CFA is used to determine an appropriate measurement model only.[13]

When compared to EFA, CFA is generally presented as a hypothesis testing process. The rationale behind that conception is quite simple: CFA requires the specification of a measurement model prior to the analysis (cf. Hurley et al. 1997). If there is no hypothesis regarding the structure of the measurement model, then there is no point in trying to run a CFA. In this respect, CFA is not a data accommodation process. It does not tell what is happening in the data. Rather, researchers predict that the data will behave according to a specific factorial structure and test this hypothesis through CFA. The procedure generates an estimated population covariation matrix on the grounds of a specific hypothesized measurement model, and then evaluate how this estimated covariation matrix relates to the actual covariation matrix of the data (cf. Ullman 2013, p. 685). Formally, CFA determines using multiple regression analysis whether the hypothesized factorial structure can predict the between-subjects variation on the items

[13] Once a scale has been properly validated, researchers can use different measurement models and study the relationships between the latent factors through structural equation modeling. We will leave the analysis of model selection with respect to structural models for future research.

(cf. Brown 2015). That is, the specified factors are used to predict (mathematically, via regressions) the variance on items. Although understanding predictions through heuristic novelty allows us to disregard the time at which the data was obtained, it is noteworthy that, given that factors are taken as explanations of the variation on items, they can conceptually be considered temporally prior. Accordingly, CFA determines whether the hypothesized measurement model can be used to predict the variance in the data set. So Penny is a predictor.

The choice between different CFA models is justified via two sets of criteria. First, it is justified on the grounds of a chi-square difference test (Ullman 2013, p. 685). This allows to balance the complexity of the model against goodness-of-fit insofar as it determines whether the difference between chi-square values is sufficient (i.e. statistically significant) to justify a reduction of degrees of freedom (i.e. a more complex model). Second, the evaluation of a measurement model is accomplished through various goodness-of-fit criteria (e.g. Browne and Cudeck 1992; Hu and Bentler 1998; Bentler 2007; see also Kline 2005; Byrne 2012; Ullman 2013). Depending on the fit of the model that is tested during a CFA, the data will be taken to either confirm or refute the measurement model.

From these considerations, one can see why CFA is not used on the entire data set during the first step of scale construction (though, as we will see, it may be used on an entire data set during a cross-validation). When constructing a scale, researchers start with many items that are taken to be plausible indicators of specific latent constructs. The development of these items may (or may not) be theory driven. However, there is no guarantee that these items are indeed indicators of said constructs and, if they are, there is no guarantee that they are good indicators. As such, EFA is first used as an iterative procedure to clean up the (large) pool of items. The first step of scale construction cannot be accomplished through CFA insofar as this method does not allow to accomplish this. Although CFA can tell that there is something problematic regarding an item, it cannot tell exactly what is wrong. For example, it may be that the item does not load on any factor or that it has multiple loadings. Moreover, the issue of model selection would not make sense within this context. Indeed, model comparison must be accomplished with respect to the same data set. Otherwise, one is comparing apples and oranges. During scale construction, researchers use EFA to reduce the complexity of the initial data set and exclude some items. The issue of model selection regards the choice of the number of factors to retain with respect to the same data set. Once a model is selected, some items are discarded and the process is repeated. Incidentally, one can see that D_2 will actually be a fragment of D_2 composed only of the items retained from the iterative use of EFA on D_1. If CFA were used to accomplish this, however, researchers would need to compare measurement models applied to different data sets (seeing that different hypothesized measurement models that do not use the same items are models for different data sets).

6 Cross-validation

Various guidelines exist to orient researchers during scale construction. In general, EFA and CFA are presented as techniques that should be applied to the entire data

set (e.g. Costello and Osborne 2005; Tabachnick and Fidell 2013). Annie's strategy fits this case. Nonetheless, when the data set is large enough, researchers are often encouraged to randomly split the data set in two parts in order to perform a cross-validation. Penny's and Rosie's strategies fit this profile.

Looking at the psychology literature, the expression *cross-validation* can be understood in three different ways: (i) as a criterion for model selection, (ii) when the model obtained on the grounds of a data set is used to predict the behavior of another data set, and (iii) replicability across samples.

First, in a narrower sense, cross-validation is a technique that can be used as a criterion for model selection (cf. Stone 1974, 1977). There are different ways to cross-validate a model, depending on the proportion used to divide the data set D. Usually, researchers will randomly split the data set D in two equal parts D_1 and D_2, though different proportions could be used. Basically, the idea behind cross-validation is to randomly split a data set in two and then determine whether or not the model obtained via the accommodation of D_1 can predict D_2. This yields a correlation coefficient. The procedure is accomplished many times using different random partitions and a cross-validation score is computed from the average of the correlation coefficients. This is the type of cross-validation that is discussed by Hitchcock and Sober (2004, p. 13). The relevance of Stone's work with respect to Hitchcock and Sober's argument is the asymptotic equivalence between a cross-validation score and AIC. Though there are differences with cross-validation and Penny's strategy, they use the relationship between her strategy and cross-validation, on the one hand, and cross-validation and AIC, on the other, to argue in favor of **II** over **I**: Penny's strategy is comparable to cross-validation, cross-validation is asymptotically equivalent with AIC and AIC safeguards against overfitting. Thus, to some extent, Penny's strategy safeguards against overfitting. The relevance of cross-validation with respect to model selection in the cases of EFA and CFA has been analyzed by Cudeck and Browne (1983), who also noted the parallel with AIC and BIC.

There are differences between Penny's strategy and cross-validation in Stone's sense. The main difference noted by Hitchcock and Sober (2004, p. 13) regards the fact that the cross-validation procedure is only accomplished one time in Penny's case, in contrast with cross-validation as proposed by Stone (1974, 1977) or Cudeck and Browne (1983). Nonetheless, Penny's strategy can still be understood as a form of cross-validation, though in this case the expression *cross-validation* gets another meaning.

A second way to understand cross-validation in psychological research is that results can be cross-validated when the model obtained on the grounds of D_1 is used to predict D_2 (cf. Howell 2010, pp. 549–550). This type of cross-validation is similar to cross-validation in the aforementioned sense but differs in that it is accomplished only one time. When performing CFA on D_2, Penny cross-validates her results in this sense, although there is a slight difference. Penny uses the model obtained on the grounds of accommodation of D_1 to predict the variance on the items of D_2 (rather than to predict D_2 per se).

Rosie, however, does not cross-validate her results in this sense.[14] As we mentioned earlier, even though a researcher might have some prior hypothesis regarding the factorial structure, this hypothesis bears no consequences on the result of the EFA. The measurement model obtained via accommodation of D_1 is not used to predict the variance in D_2. Rather, EFA, constrained by assumptions regarding the number of factors to extract, the factor extraction technique and the rotation method, accommodates D_2. Although Rosie *predicts* that the factorial structure can be replicated, the test she performs to examine this hypothesis is not a (multiple) regression.

7 Replicability

There is third, broader sense in which cross-validation can be understood in psychology. From a methodological perspective, cross-validation can be understood as *replicability across samples* (cf. Comrey 1988; Hurley et al. 1997; Hinkin 1998; Huck 2012). In this context, there is a distinction between *internal* and *external* replicability (cf. Thompson 1994, 2004). External replicability amounts to replicate the model obtained by accommodating D on another data set D' that is assumed to come from the same population. Thus understood, external replicability is a form of constructive replication (by opposition to direct or exact replication), where there might be differences in the research designs or in the data (cf. Brandt et al. 2014; Makel et al. 2012). In that sense, (external) replication means providing further similar evidence in favor of the same model. In contrast, internal replicability consists in randomly dividing a data set D in two parts D_1 and D_2, to obtain a model by accommodating D_1 and then to try to replicate this model via the same technique on D_2.[15] Rosie's strategy fits this profile. So Rosie is a replicator.

Rosie, like Penny, uses EFA iteratively on D_1 to obtain an appropriate measurement model. Using this procedure, she eliminates items with double loadings to identify a well-defined factorial structure. When she performs EFA on D_2, however, she does not follow the same guidelines. Indeed, she starts with the measurement model obtained via accommodation of D_1. As such, she does not use all the items that were present at the beginning of her analysis on D_1 but only uses specific subsets of items, resulting from her iterative use of EFA on D_1, that are supposed to be good indicators of the factors. Recall that EFA tells what is happening within the data, modulo a specific number of factor to extract and some noise in the data set. Using EFA on D_2, her aim is to determine whether the same factorial structure can be found within the data. Accordingly, Rosie tries to (constructively) replicate the measurement model obtained via accommodation of D_1 on D_2.

In light of these considerations, let us revisit the accommodation strategies presented in Sect. 4. Let M stand for a measurement model (the subscripts $_E$ and $_C$ refer to *exploratory* and *confirmatory*). Given that Hitchcock and Sober already argued in

[14] To some extent, trying to replicate results from previous studies can be understood as a prediction that the results can be generalized.

[15] Thus understood, replication amounts to the reproduction of a result using the same technique.

favor of the weak dominance of using AIC in favor of maximizing goodness-of-fit for model selection, let us concentrate only on strategies **III** and **IV**.

- $\mathbf{III}_E$ M is obtained by accommodation from D through EFA using AIC.
- $\mathbf{IV}_{EE}$ M is obtained by accommodation from D_1 through EFA using AIC and M is replicated on D_2 via EFA.
- $\mathbf{IV}_{EC}$ M is obtained by accommodation from D_1 through EFA using AIC and M is used to predict the variance in D_2 via CFA.

8 A new debate

8.1 Achieving empirical adequacy

Assuming that one goal of science is empirical adequacy, predictive accuracy can be seen as an indicator that can be used to achieve that goal. However, many would be inclined to argue that, though science should aim at predictive accuracy to achieve empirical adequacy, another important indicator of empirical adequacy is replicability. Replication of experimental results can be seen as a gold standard of the scientific enterprise (cf. Norton 2015). This conception also applies to psychology (cf. Francis 2012; Asendorpf et al. 2013). Hitchcock and Sober's analysis provides an answer to the debate between prediction and accommodation. However, the analysis of the different guidelines that are suggested for scale construction brings a third player to the table: Replication. We thus obtain a new debate, namely prediction versus accommodation versus replication.

Both prediction and replication can be used to safeguard against overfitting. Indeed, an accurate prediction is an indicator that overfitting did not happen (cf. Hitchcock and Sober 2004, p. 18). In strategy $\mathbf{IV}_{EC}$, if the model obtained via accommodation of D_1 is predictively accurate with regard to D_2, then it is an indicator that overfitting did not happen. Similarly, in strategy $\mathbf{IV}_{EE}$, we would not expect a replication to happen on D_2 if the model obtained through EFA via accommodation of D_1 were overfitted. As it stands, replicability is also a criterion that can be seen as an indicator that can be used to achieve empirical adequacy. After all, empirically adequate theories should have consequences that can be replicated. Put differently, given an empirically adequate theory, one should be able to reproduce similar results using the same techniques and procedures.

The idea that replication of a measurement model via EFA is more rigorous than trying to confirm it via CFA might seem self-evident for psychologists and perhaps some researchers in social sciences. After all, quantitative human and social sciences capitalize on chance. In the context of scale construction, the model obtained through EFA inevitably includes idiosyncratic characteristics of the data set (i.e., the model will include chance relationships found in the data; cf. MacCallum et al. 1992). As such, one might believe that it is *prima facie* improbable that the results obtained via EFA might be replicable on other data sets, so replication of a measurement model would be a worthy achievement.

To say that human and social science research capitalize on chance relationships that can be found in samples is another way to look at the problem of overfitting. Indeed,

when overfitting, one is taking into account characteristics that are idiosyncratic to the data set. It is noteworthy, however, that there are safeguards that are used to prevent overfitting during scale construction.[16] EFA, for instance, includes error terms that are meant to capture (at least part of) the noise in the data. In addition, using AIC as a criterion for model selection insures that one is not sacrificing simplicity for goodness-of-fit. The threshold of .3 to consider factor loadings as significant is also another way to avoid considering chance relationships that are specific to a given sample.

In light of these considerations, it is not improbable that one might be able to replicate a measurement model via EFA on further data sets. As such, it is worth attending to the role played by replication during scale construction and studying the epistemic significance of that methodology.

8.2 Predictive accuracy and replicability

In the context of scale construction, predictive accuracy, as it is achieved through CFA, and replicability, as achieved via EFA, are two related criteria. To see this, consider how Penny and Rosie would fare if the results of their analysis were positive.

In the best case scenario, Rosie succeeds in replicating M on D_2. On the grounds of the measurement model obtained via accommodation of D_1, Rosie discarded some items in D_2 and was able to reproduce the factorial structure using EFA. Hence, the measurement model M is replicable. In this eventuality, it follows that it is also predictively accurate. Indeed, given specific constraints, EFA provides the best measurement model that fits D_2. As such, if Rosie were to perform CFA on D_2 with M as a hypothesis, then she would obtain a good fit. Performing CFA on the same data set using the measurement model obtained via EFA is somewhat redundant (cf. Hurley et al. 1997, pp. 675–677). This should not come as a surprise seeing that Rosie would take a model that yields the actual covariation matrix of D_2 to estimate a population covariation matrix that would then be compared to the actual covariation matrix of D_2. The hypothesized factorial structure will accurately predict the variance on the items in D_2 given that it is (modulo error and the number of factors that are extracted) the factorial structure of D_2. With respect to scale construction, it can thus be argued that replicability (achieved through EFA) is a higher standard than predictive accuracy (achieved through CFA) given that the satisfaction of the former implies the satisfaction of the latter. Consequently, Rosie wins on both accounts: Replicability and predictive accuracy are both satisfied.

Penny has the second best case scenario. If M has a good fit on D_2, then the factorial structure is predictively accurate. Predictive accuracy is thus satisfied. However, Penny will not know whether M is replicable per se. She will know that there is a good fit, but she will not know if the model can be replicated using EFA. Indeed, a measurement model can be predictively accurate with respect to CFA without being replicable through EFA. The predictive accuracy of the model is determined via statistically significant regression coefficients (factor loadings). As we mentioned earlier, EFA accommodates the data and provides the actual covariation matrix of the data, modulo

[16] This would not be the case if one were to use principal component analysis instead of EFA.

the number of factors and some noise that might be in the data set. Nonetheless, one can test a model via CFA that might turn out to be good enough, with acceptable regression coefficients and an acceptable fit. This would not imply, though, that it is the model that would be provided by EFA (which would represent the actual factorial structure of the data). If Penny were to perform an EFA, there is nothing guaranteeing that the factorial structure would be replicated. It is possible to have different factorial structures with the same number of factors with good fits and acceptable regression coefficients. This is actually why model selection is an important aspect of CFA. Therefore, if she succeeds, Penny only satisfies the criterion of predictive accuracy. She would have to actually perform an EFA to determine whether M is replicable, and there is nothing in the procedure guaranteeing it is.

8.3 Prediction versus accommodation versus replication

Hitchcock and Sober use two external criteria to rank the results of the different strategies, namely the principle of total evidence and predictive accuracy. As a result of our analysis, we propose to consider a third external principle, namely replicability.

Hitchcock and Sober's understanding of the principle of total evidence is that *all* available data should be taken into account when assessing a model. To some extent, Annie, Penny and Rosie all violate this understanding of the principle insofar as they discard some items during the iterative use of EFA. This formulation of the principle of total evidence, however, can be criticized. As Autzen (2016) argued, total evidence does not dictate to consider *all* available data, but rather specifies that all *relevant* available data should be considered. Accordingly, given two data sets D and D' such that D' is a proper subset of D, a measurement model should be based on D if these data are relevant for the model in light of D' (cf. Autzen 2016, p. 286). Applying this conception of total evidence to our example, D represents the entire data set used by Annie, Penny and Rosie before the iterative use of EFA, whereas D' represents the data set obtained after some items have been discarded.

To say that Annie, Penny and Rosie all violate the principle of total evidence amounts to say that total evidence dictates that researchers should not be discarding items during the process. Put differently, this amounts to the claim that the data set before the iterative use of EFA is relevant for the measurement model given the data set obtained afterwards. The formulation of the principle of total evidence as proposed by Autzen relies on a relevance criterion, which can be conceived in different ways. Essentially, however, this relevance criterion relies on a criterion for evidential assessment, specifying the conditions under which data provide support for a measurement model (cf. Autzen 2016, pp. 285–286).

A measurement model obtained through EFA is considered to be empirically supported by the data when it satisfies the simple structure criterion (cf. Thurstone 1947, 1954). Although further criteria might be used when there are competing simple structures (e.g. AIC), it is the simple structure criterion that is used to evaluate whether a factorial structure is appropriate. As a criterion for evidential assessment, the simple structure criterion is also a relevance criterion. As such, given that an adequate measurement model has to be representable via a simple structure, a data set from which

one cannot obtain an adequate simple structure is irrelevant for the model. Therefore, discarding items from the data set in order to be able to obtain a model that satisfies the simple structure criterion does not violate the principle of total evidence.

In light of these considerations, it can be argued that all three researchers satisfy the principle of total evidence, although they do so in different ways. Indeed, all three researchers use the entire data set obtained after the iterative use of EFA. The differences lie in the analyses that are subsequently made. Penny and Rosie also use all the data but they do so differently. Although both Penny and Rosie accommodate D_1 using EFA, Rosie uses the remaining of the data to apply the same statistical analysis whereas Penny uses it to perform a CFA. Given that, to some extent, both Penny and Rosie satisfy the principle of total evidence, the aforementioned considerations provide us with an argument in favor of local weak replicationism. Granted that predictive accuracy and replicability are (among others) two indicators of empirical adequacy, then, in the eventuality that both strategies are successful, Rosie wins on both counts whereas Penny only achieves predictive accuracy. Hence the result of Rosie's strategy is superior to Penny's ($\mathbf{IV}_{EE} > \mathbf{IV}_{EC}$).

We have established that all three researchers do not violate the principle of total evidence by discarding items from the initial data set. Nonetheless, Annie might have an advantage over Penny and Rosie with respect to total evidence seeing that she accommodates the data while safeguarding against overfitting using *all* the relevant available data. To clearly understand why, we must see that all three research strategies share a common method, namely the iterative use of EFA. If all three strategies are compared solely on the grounds of this common method to examine how the items are discarded, then there is a case to be made in favor of Annie's satisfaction of the principle of total evidence.

Prior to the iterative use of EFA, Annie starts with the entire data set D whereas Penny and Rosie start with proper subsets D_1 of D. Given Autzen's formulation of total evidence, there is a case to be made in favor of Annie if D is relevant for the measurement model given D_1. As it stands, D is relevant for the measurement model at this stage of the research process, but its relevance is not assessed using the simple structure criterion. After all, prior to the iterative use of EFA, neither data set enable researchers to have a measurement model that satisfies the simple structure criterion.

To properly understand why D is relevant here but was not relevant earlier when we were considering discarding items, we must note a shift of meaning in the expression *all available data* from Hitchcock and Sober to Autzen's criticism of the principle of total evidence. When saying that all available data should be considered, Hitchcock and Sober are concerned with the number of *observations*, whereas Autzen is concerned with the number of *variables*.

To exemplify this point, let us represent the initial data set D via a 2-dimensional table. Each row in the table corresponds to a participant, whereas each column corresponds to a variable. An intersection row/column in that table corresponds the observation made for a participant on a given variable. When claiming that total evidence applies to relevant data, Autzen means that only irrelevant columns (variables) should be deleted from the data set. In contrast, when Hitchcock and Sober claim that all data should be considered, they mean that, given a fixed number of columns (variables), all rows (participants) should be considered.

The relevance criterion used to establish whether columns (variables) should be deleted from the data set (i.e., whether some items should be discarded) is not the same criterion used to determine whether rows (participants) should be deleted (i.e., whether the set should be partitioned). While the simple structure criterion can be used to judge the relevance of a variable in the former case, sample size is a relevance criterion in the latter. Whether researchers are justified to discard items from the data set concerns the elimination of columns (variables). However, whether Annie's strategy satisfies total evidence better than Penny and Rosie during the iterative use of EFA regards the number of rows that are considered in the analysis. The number of rows is a matter of sample size, and sample size always matters, for it is directly linked to statistical power (e.g. Aron et al. 2013). In this sense, D is relevant for the measurement model, hence Penny and Rosie violate the principle of total evidence.

In light of these considerations, it is noteworthy that Annie's ($\mathbf{III}_E$) and Penny's ($\mathbf{IV}_{EC}$) strategy fit Hitchcock and Sober's analysis. Therefore, we have a partial answer to Sara's problem since their arguments allow us to rank these strategies ($\mathbf{III}_E > \mathbf{IV}_{EC}$). Given that we have also established that Rosie's strategy is superior to Penny's ($\mathbf{IV}_{EE} > \mathbf{IV}_{EC}$), it remains to determine what is the relationship between Rosie's and Annie's strategies.

Granted that the principle of total evidence, in addition to predictive accuracy and replicability, is also a criterion that should be used to evaluate the results of the different strategies, Rosie's strategy only fully satisfies two criteria (predictive accuracy and replicability). Even though she uses all the data at her disposal, one could argue that the model obtained through accommodation of D_1 violates the principle of total evidence insofar as it does not consider *all* the data available (cf. Hitchcock and Sober 2004, p. 31). In contrast, Annie's model fully satisfies the principle of total evidence. Further, Annie's strategy satisfies the criterion of predictive accuracy given that she used AIC to determine the appropriate factorial structure. However, Annie's strategy does not satisfy the criterion of replicability. As far as she knows, she cannot determine whether her model is replicable (she would need to externally cross-validate her result).

As a result of our analysis, it appears that the ranking of one strategy over the other reflects a deeper conflict between the criterion of replicability and the principle of total evidence. There are two avenues one can take to argue in favor of the ranking of one strategy over the other: Either one gives primary importance to the principle of total evidence or one rather insists on reaching the criterion of replicability.

The former avenue might be argued along the lines of Hitchcock and Sober's analysis. One could argue that as along as one is not overfitting the data, replication has no advantage over accommodation and one should consider all the data at hand (cf. Hitchcock and Sober 2004, pp. 17–18). In this case, one would rank the principle of total evidence over the criterion of replicability, thus $\mathbf{III}_E$ would be better than $\mathbf{IV}_{EE}$ since both achieve predictive accuracy but only $\mathbf{III}_E$ fully takes into consideration the principle of total evidence. As such, one would obtain a transitive ordering: $\mathbf{III}_E > \mathbf{IV}_{EE} > \mathbf{IV}_{EC}$.

Nonetheless, one might be inclined to argue that replication, as a criterion meant to achieve empirical adequacy, has priority over the principle of total evidence. For instance, one might argue that data should be used to provide an accurate account of causal parameters (e.g. Myrvold and Harper 2002). Hence, safeguarding against

overfitting is not sufficient to argue in favor of the principle of total evidence. Rather, one should achieve replicability seeing that reproducing a factorial structure is a good indicator of its empirical adequacy. In this case, $\mathbf{IV}_{EE}$ would be ranked over $\mathbf{III}_{E}$ since, in addition to balancing simplicity and goodness-of-fit, it also satisfies the criterion of replicability. Again, one would obtain a transitive ordering: $\mathbf{IV}_{EE} > \mathbf{III}_{E} > \mathbf{IV}_{EC}$.

With that being said, all three criteria are important, and steps should be taken to satisfy each of them. In the context of scale construction, the ranking of one strategy over the other depends upon the research plan. In the eventuality that the research team does not intend to externally cross-validate their results, then Rosie's model should be preferred. Indeed, the violation of the principle of total evidence is very small and should be neglected. Rosie does use all the data available, although she does so differently. Further, her strategy allows to satisfy replicability, predictive accuracy and avoid overfitting. Hence, in the context of an internal cross-validation, local weak replicationism holds.

But still, one could insist that there should be no compromise and that the measurement model should be obtained by accommodation (using AIC) of *all* the data available. In this case, however, the research team would need to externally cross-validate their results, otherwise the criterion of replicability would remain unsatisfied.

Consequently, if the research team intends to perform an external cross-validation and replicate their results, then Annie's model should be preferred and local weak accommodationism holds. Given that the sample size from which Annie's measurement model was obtained (through an iterative use of EFA) was bigger, her model is less likely to capitalize on chance and is therefore more likely to replicate.

Otherwise, if the research team is not able (or do not intend) to perform an external cross-validation (e.g. due to lack of time, participants, or resources), then Rosie's model should be preferred. In both cases, the three criteria are met, though there is a slight violation of total evidence in the latter.

9 Problem solved

On the next day, Sara invited Annie, Penny and Rosie to meet in order to discuss the issue of model selection with respect to scale construction. After greeting her colleagues, Sara went straight to the point. Assuming that one goal of your research is empirical adequacy, then the choice of the measurement model should be justified through criteria that are meant to achieve that goal. Among these criteria we find predictive accuracy and replicability. In addition to these criteria, the principle of total evidence, as a matter of rationality, should be taken into account. All three criteria are important and steps should be taken to satisfy each of them. That being said, the choice of the best model to use also depends upon your research plan. If you do not intend to externally cross-validate your model, then Rosie's model should be used. Indeed, Penny's strategy (predictive accuracy and, although with a slight violation, total evidence) as well as Annie's strategy (total evidence and predictive accuracy) only satisfy two out of three of these criteria, whereas Rosie's strategy satisfies all three (with a slight violation of total evidence). In this context, Rosie's violation of total evidence is negligible and replicability should have priority insofar as it is

a good indicator of empirical adequacy, which is an important goal of psychological research. Otherwise, if you intend to externally cross-validate your model (on data that is assumed to come from the same population), then you should use Annie's model and try to reproduce it via EFA. Either way, each criterion is met. In the eventuality that replication fails during the external cross-validation, then use CFA to determine whether the measurement model is nonetheless acceptable. Annie, Penny and Rosie were quite satisfied by Sara's answer. They had more work to do, but they were happy to have a solution to their problem. They thus thanked Sara and began to see their way out of her office. It was a pleasure, Sara said, but one last thing: Next time, please just call Piper.

Acknowledgements I would like to thank Stephan Hartmann for valuable comments and suggestions made on a previous draft of this paper. I am also grateful to anonymous referees, whose comments and suggestions helped to improve this article, and to Sarah-Geneviève Trépanier, for enlightening discussions on the subject. This research was financially supported by the Social Sciences and Humanities Research Council of Canada.

References

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(4), 433–459.

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, *52*(3), 317–332.

Aron, A., Coups, E. J., & Aron, E. N. (2013). *Statistics for psychology* (6th ed.). London: Pearson.

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., et al. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*(2), 108–119.

Autzen, B. (2016). Significance testing, p-values and the principle of total evidence. *European Journal for Philosophy of Science*, *6*(2), 281–295.

Bentler, P. M. (2007). On tests and indices for evaluating structural models. *Personality and Individual Differences*, *42*(5), 825–829.

Borsboom, D. (2008). Latent variable theory. *Measurement*, *6*(1–2), 25–53.

Borsboom, D., & Mellenbergh, G. J. (2004). Why psychometrics is not pathological: A comment on Michell. *Theory & Psychology*, *14*(1), 105–120.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*(2), 203–219.

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., et al. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, *50*, 217–224.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York: Guilford Publications.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, *21*(2), 230–258.

Byrne, B. M. (2012). *Structural equation modeling with Mplus*. New York: Routledge.

Comrey, A. L. (1988). Factor-analytic methods of scale development in personality and clinical psychology. *Journal of Consulting and Clinical Psychology*, *56*(5), 754–761.

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*, *10*(7), 1–9.

Cudeck, R., & Browne, M. W. (1983). Cross-validation of covariance structures. *Multivariate Behavioral Research*, *18*(2), 147–167.

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods*, *5*(2), 155–174.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272–299.
Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London: Sage.
Finch, J. F., & West, S. G. (1997). The investigation of personality structure: Statistical models. *Journal of Research in Personality*, *31*(4), 439–485.
Flora, D. B., LaBrish, C., & Chalmers, R. P. (2012). Old and new ideas for data screening and assumption testing for exploratory and confirmatory factor analysis. *Frontiers in Psychology*, *3*(55), 1–21.
Forster, M. R. (2002). Predictive accuracy as an achievable goal of science. *Philosophy of Science*, *69*(3), S124–S134.
Forster, M., & Sober, E. (1994). How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *British Journal for the Philosophy of Science*, *45*(1), 1–35.
Francis, G. (2012). Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin & Review*, *19*(6), 975–991.
Gardner, M. R. (1982). Predicting novel facts. *The British Journal for the Philosophy of Science*, *33*(1), 1–15.
Glymour, C. (2001). *The mind's arrows: Bayes nets and graphical causal models in psychology*. Cambridge: MIT Press.
Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods*, *1*(1), 104–121.
Hitchcock, C., & Sober, E. (2004). Prediction versus accommodation and the risk of overfitting. *British Journal for the Philosophy of Science*, *55*(1), 1–34.
Hood, S. B. (2013). Psychological measurement and methodological realism. *Erkenntnis*, *78*(4), 739–761.
Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185.
Howell, D. C. (2010). *Statistical methods for psychology* (7th ed.). Wadsworth: Cengage Learning.
Hu, L.-T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*(4), 424–453.
Huck, S. W. (2012). *Reading statistics and research* (6th ed.). London: Pearson.
Hurley, A. E., Scandura, T. A., Schriesheim, C. A., Brannick, M. T., Seers, A., Vandenberg, R. J., et al. (1997). Exploratory and confirmatory factor analysis: Guidelines, issues, and alternatives. *Journal of Organizational Behavior*, *18*(6), 667–683.
Johnson, K. (2016). Realism and uncertainty of unobservable common causes in factor analysis. *Noûs*, *50*(2), 329–355.
Jolliffe, I. T. (2002). *Principal component analysis* (2nd ed.). Berlin: Springer.
Jolliffe, I. T., & Morgan, B. J. T. (1992). Principal component analysis and exploratory factor analysis. *Statistical Methods in Medical Research*, *1*(1), 69–95.
Kline, R. B. (2005). *Structural equation modeling*. New York: The Guilford Press.
Kuhn, T. S. (1983). Rationality and theory choice. *The Journal of Philosophy*, *80*(10), 563–570.
Kummerfeld, E., Ramsey, J., Yang, R., Spirtes, P., & Scheines, R. (2014). Causal clustering for 2-factor measurement models. In T. Calders, F. Esposito, E. Hüllermeier, & R. Meo (Eds.), *Machine learning and knowledge discovery in databases, volume 8725 of lecture notes in computer science* (pp. 34–49). Berlin: Springer.
MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, *111*(3), 490–504.
Maher, P. (1988). Prediction, accommodation and the logic of discovery. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, *1*, 273–285.
Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research. *Perspectives on Psychological Science*, *7*(6), 537–542.
Maslach, C., & Jackson, S. E. (1981). The measurement of experienced burnout. *Journal of Occupational Behavior*, *2*(2), 99–113.
Mellenbergh, G. J. (1994). Generalized linear item response theory. *Psychological Bulletin*, *115*(2), 300–307.
Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, *88*(3), 355–383.
Michell, J. (2003). The quantitative imperative positivism, naive realism and the place of qualitative methods in psychology. *Theory & Psychology*, *13*(1), 5–31.

Michell, J. (2004). The place of qualitative research in psychology. *Qualitative Research in Psychology*, *1*(4), 307–319.
Mulaik, S. A. (1991). Factor analysis, information-transforming instruments, and objectivity: A reply and discussion. *The British Journal for the Philosophy of Science*, *42*(1), 87–100.
Murray-Watters, A., & Glymour, C. (2015). What is going on inside the arrows? Discovering the hidden springs in causal models. *Philosophy of Science*, *82*(4), 556–586.
Musgrave, A. (1974). Logical versus historical theories of confirmation. *British Journal for the Philosophy of Science*, *25*(1), 1–23.
Myrvold, W. C., & Harper, W. L. (2002). Model selection, simplicity, and scientific inference. *Philosophy of Science*, *69*(S3), S135–S149.
Norton, J. D. (2015). Replicability of experiment. *Theoria*, *30*(2), 229–248.
O'Connor, B. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavioral Research Methods, Intruments and Computers*, *32*(3), 396–402.
Park, H. S., Dailey, R., & Lemus, D. (2002). The use of exploratory factor analysis and principal component analysis in communication research. *Human Communication Research*, *28*(4), 562–577.
Preacher, K. J., Zhang, G., Kim, C., & Mels, G. (2013). Choosing the optimal number of factors in exploratory factor analysis: A model selection perspective. *Multivariate Behavioral Research*, *48*(1), 28–56.
Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. *The Journal of Educational Research*, *99*(6), 323–338.
Schurz, G. (2014). Bayesian pseudo-confirmation, use-novelty, and genuine confirmation. *Studies in History and Philosophy of Science*, *45*(1), 87–96.
Silva, R., Scheines, R., Glymour, C., & Spirtes, P. (2006). Learning the structure of linear latent variable models. *The Journal of Machine Learning Research*, *7*, 191–246.
Sober, E. (2004). Likelihood, model selection, and the Duhem-Quine problem. *Journal of Philosophy*, *101*(5), 221–241.
Spirtes, P., Glymour, C., & Scheines, R. (1991). From probability to causality. *Philosophical Studies*, *64*(1), 1–36.
Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, Prediction, and Search*. Cambridge: MIT Press.
Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677–680.
Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society B*, *36*(2), 111–147.
Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society B*, *39*(1), 44–47.
Suppes, P. (2007). Statistical concepts in philosophy of science. *Synthese*, *154*(3), 485–496.
Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). London: Pearson.
Thompson, B. (1994). The pivotal role of replication in psychological research: Empirically evaluating the replicability of sample results. *Journal of Personality*, *62*(2), 157–176.
Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington, DC: American Psychological Association.
Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago: University of Chicago Press.
Thurstone, L. L. (1954). An analytical method for simple structure. *Psychometrika*, *19*(3), 173–182.
Ullman, J. B. (2013). Structural equation modeling. In B. G. Tabachnick & L. S. Fidell (Eds.), *Using multivariate statistics* (6th ed.). London: Pearson.
van Fraassen, B. C. (1980). *The scientific image*. New York: Oxford University Press.
Velicer, W. F., & Jackson, D. N. (1990a). Component analysis versus common factor analysis: Some further observations. *Multivariate Behavioral Research*, *25*(1), 97–114.
Velicer, W. F., & Jackson, D. N. (1990b). Component analysis versus common factor analysis: Some issues in selecting an appropriate procedure. *Multivariate Behavioral Research*, *25*(1), 1–28.
White, R. (2003). The epistemic advantage of prediction over accommodation. *Mind*, *112*(448), 653–683.
Worrall, J. (2002). New evidence for old. In P. Gärdenfors, J. Wolenski, & K. Kijana-Placek (Eds.), *In the scope of logic, methodology and philosophy of science* (pp. 191–209). Dordrecht: Kluwer Academic Publishers.