Association for Information Systems
AIS Electronic Library (AISeL)

ICIS 2009 Proceedings

International Conference on Information Systems (ICIS)

2009

Building Theory from Quantitative Studies, or, How to Fit SEM Models

Joerg Evermann
Memorial University of Newfoundland, jevermann@mun.ca

Mary Tate
Victoria University of Wellington, Mary.tate@vuw.ac.nz

BUILDING THEORY FROM QUANTITATIVE STUDIES, OR, HOW TO FIT SEM MODELS

Completed Research Paper

Joerg Evermann
Memorial University of Newfoundland
St. John's, Newfoundland, Canada
jevermann@mun.ca

Mary Tate
Victoria University of Wellington
Wellington, New Zealand
mary.tate@vuw.ac.nz

Abstract

Covariance-based structural equation modeling (CB-SEM) is an increasingly popular technique for analyzing quantitative data in Information Systems research. As such, it is traditionally viewed as a method to test theory, rather than build it. However, many of the theoretical models tested with this technique show significant differences between the model and the data, which the IS community has been willing to overlook. This paper shows that as part of the pursuit of model fit, CB-SEM can provide deeper insights into a phenomenon, allowing us to build theories based on quantitative data.

Keywords: Structural equation modeling, Theory building, Quantitative analysis, data analysis, research methods

Introduction

Theories are sets of propositions that relate concepts or constructs, bounded by a specified context (Bacharach, 1989). Theories may exist for different purposes, among them explanation and prediction (Dubin, 1969; Glaser and Strauss, 1967; Gregor, 2006). Theories that explain and predict provide causal explanations of a phenomenon and testable hypotheses (Bacharach, 1989; Dubin, 1969; Gregor, 2006; Popper, 1968). A representation relationship relates the theoretical level of constructs and propositions, to the more specific level of variables and hypotheses: Variables represent constructs and hypotheses that relate variables are derived from the propositions that relate constructs (Bacharach, 1989).

Theory building has traditionally been based on qualitative data, using inductive case studies (Eisenhardt, 1989) or grounded theory approaches (Glaser and Strauss, 1967; Urquhart et al., 2009). An important aspect of both is the interplay between data and emerging theory. The researcher constantly compares data with theory, to refine theory based on data, to identify new themes in the data, and to ensure the emergent theory explains the data (Glaser and Strauss, 1967; Urquhart et al., 2009). However, Glaser and Strauss (1967) point out that "there is no fundamental clash between the purposes and capacities of qualitative and quantitative methods or data … each form of data is useful for both verification and generation of theory." (pg. 17f)

Covariance-based structural equation modeling (CB-SEM) is an increasingly popular technique for analyzing quantitative data. CB-SEM operates on Bacharach's (1989) specific level. In the statistical model, latent variables represent theoretical constructs, and hypothesized regression relationships between them represent hypothesized causal propositions between constructs. CB-SEM is typically used for theory testing as it allows simultaneous testing of multiple relationships and provides a test of how well the statistical model, which represents the theory, fits the observed data.

The test of model fit is a $\chi 2$ test, which tests whether a model, with its estimated parameters, explains the observed covariances (Bollen, 1989). For model-implied covariances that are not significantly different from observed ones, the $\chi 2$ test should be insignificant. Researchers, also in the information systems (IS) discipline (Gefen et al. 2000), and even textbook authors (Hair et al., 2005) suggest ignoring the $\chi 2$ test. They argue that the test statistic depends on sample size, and it is therefore impossible to specify a well-fitting model with realistic sample sizes. This problem of ignoring substantial misfit is widespread in the IS literature. Of the 54 studies published in MISQ, ISR, JMIS and JAIS between 2004 and 2008 that use CB-SEM, only three achieve a non-significant $\chi 2$ test for their model, and only one of the remaining 51 studies acknowledges the misfit and makes an effort at identifying or discussing the substantive reasons for the misfit. Typically, researchers provide a set of approximate fit indices to claim that, despite evidence to the contrary, their models fit the data.

Models that do not fit the observed covariances are still worthwhile publishing, because, given the extensive theory building and data collection effort that goes into a well-executed study, we can learn much from them. But rather than ignoring misfit, researchers need to identify the reasons for the misfit in order to improve the theory their models are based on: "We need to understand what is problematic if we are to do better next time around" (Hayduk et al. 2007, p. 845).

The goal of this paper is three-fold. First, we want to demonstrate that it is possible to construct well-fitting models even with realistic sample sizes. Second, we want to demonstrate that the process of fitting a model to data can lead to additional insights into the data, akin to theory building approaches in qualitative inquiry. Third, we wish to impress on the reader the importance of paying close attention to development of measurement indicators for constructs. The paper proceeds with a discussion of model fit and a description of the data used in this paper. This is followed by an extensive presentation of an example application. We then discuss the contributions and limitations, and revisit our three goals.

Model Fit

The only *statistical test* for CB-SEM models is the $\chi 2$ test (Barrett, 2007). Yet, as indicated, most published models in IS research fail this test (only 3 of 54 report non-significant $\chi 2$) and many ignore the test completely. Ignoring the $\chi 2$ test is tantamount to ignoring evidence that falsifies a theory, which is the basic method of scientific research (Popper, 1968). Hayduk et al. (2007, pg. 848) warn that "overlooking indications of potentially huge problems are the kinds of things lawyers will gladly describe as malfeasance, dereliction of responsibility, or absence of due

diligence." Researchers instead resort to approximate fit indices and threshold values to argue for well fitting models despite evidence to the contrary: "Where it has all gone badly wrong in this type of SEM model testing is that many investigators have actively avoided the statistical test of fit of their models, in favor of a collection of ad hoc indices which are forced to act like 'tests of fit'" (Barrett, 2007, pg. 819). This is also our impression of the IS literature, where 9 of 54 studies do not even present the $\chi 2$ statistic and 39 of 54 studies claim good model fit despite significant $\chi 2$ differences between model and data. Barrett (2007, pg. 819f) nicely summarizes our impression of the IS literature: "Indeed, one gets the feeling that social scientists cannot actually contemplate that most of their models do not fit their data, and so invent new ways of making sure that by referencing some kind of ad hoc index, that tired old phrase 'acceptable approximate fit' may be rolled out as the required rubber stamp of validity."

One of the main issues raised by critics of the $\chi 2$ test is sample size dependence. Critics argue that all models are misspecified, hence leading to a non-central $\chi 2$ statistic, which increases with sample size. For example, Bentler (2007, pg. 828) suggests that "a model is liable always to be misspecified, and hence to be rejected by any 'exact' test." Similarly, Goffin (2007, pg. 835) argues that "the models we develop in psychology should virtually never be presumed to contain the whole truth and therefore be subjected to a test of perfect fit". Steiger (2007, pg. 894) concludes with "Why test a hypothesis that is always false?" We are not this fatalistic and disagree with this sentiment, as do Hayduk et al. (2007), who state in a response to Barrett (2007) that "We cannot prevent Barrett from claiming that all his models are detectably wrong in general, but we can encourage everyone to strive for models that are properly specified, not wrong. Observing that at least some wrong models are more assuredly detected by larger samples (because of decreased sampling variability) is good methodological news to those seeking proper models!" (pg. 844) The sample size dependence of the $\chi 2$ test for misspecified models is a useful feature that provides the statistical power to detect these models.

Even if it were not possible in principle to construct a true model, it should be incumbent upon researchers to at least strive for exact fit. Anything less might be construed as "simple intellectual laziness on the part of the investigator." (Barrett, 2007, pg. 823). A non-significant $\chi 2$ test is a necessary (but not sufficient) condition for a true model and evidence to the contrary should not be dismissed.

In contrast, approximate fit indices are known to be problematic. Barrett (2007, pg. 817) summarizes recent studies by Beauducel and Wittmann (2005), Fan and Sivo (2005), Marsh et al. (2004) and Yuan (2005) and concludes that "single-valued indicative thresholds for approximate fit indices were impossible to set without some models being incorrectly identified as fitting 'acceptably' when in fact they were misspecified to some degree." In contrast to the $\chi 2$ test, which tests whether a model reproduces observed covariances, fit indices are developed to indicate goodness of approximation. However, as Barrett (2007, pg. 819) points out, "the problem is that no-one actually knows what 'approximation' means in terms of 'approximation to causality' in the many areas in which SEM modelling is used." Millsap (2007, pg. 877) echoes this: "In my view, the most cogent argument against setting particular thresholds for approximate fit is our lack of explicit knowledge about how values of various approximate fit indices can be translated into statements about particular types of model misspecifications." In fact, even Hu and Bentler (1999) in their much-cited paper warned researchers that "it is difficult to designate a specific cutoff value for each fit index because it does not work equally well with various conditions." (Pg. 27) Despite these warnings and mounting empirical evidence, researchers continue to treat fit indices and index thresholds as if they were statistical tests with known type I and II error rates.

While we do believe that fit indices are useful, they should not be used to hide the fact that a model does not fit the observed data. Fit indices are useful when a model fits the data, i.e. satisfies the necessary condition of not failing the $\chi 2$ test. They can then be used to compare the parsimony of models because many incorporate the model degrees of freedom. Fit indices can also be used to compare non-nested model, using information-theoretic indices such as the AIC or BIC. As Hayduk et al. (2007, pg. 844f) state, "we do not object to multi-faceted model assessments … but none of these replace or displace model testing" (pg 844f).

Thus while the $\chi 2$ test provides a necessary but not sufficient condition for identifying a true model, fit indices provide neither. It is for these reasons that we focus on the $\chi 2$ test in this paper and do not present or discuss fit indices. We believe the IS community should apply more rigorous standards in model *testing*, rather than mere model evaluation using indices with unknown properties, uncertain meaning, and problematic threshold values. As McIntosh (2007, pg. 861) points out, "Merely settling for close fit could hinder the advancement of knowledge in a given substantive field, since there is little impetus to seek out and resolve the reasons why exact fit was not attained." Hence, one of our goals is to demonstrate that it is possible to construct models that satisfy the $\chi 2$ test.

As pointed out, many published models in the IS area do not fit. However, this does not mean they are not valuable. On the contrary, given the careful theory development and extensive data collection effort, we agree with Hayduk et al. (2007, pg. 845) who argue that "attentively constructed and theoretically meaningful models that fail ought to be carefully discussed and published… Any area that is unable to openly acknowledge and examine the deficiencies in its current theories is hampered from proceeding toward better theories… If a model fails, the authors should not proceed to discuss the model as if it were 'OK anyway'. They should publish a discussion of 'how the world looks from this theory/model perspective', and their diagnostic investigations of 'how and why this theory/model perspective on the world fails'. We need to understand what is problematic if we are to do better next time around." Barrett (2007) makes sensible suggestions on how to deal with this further, suggesting that "if the [distributional] assumptions appear reasonable, … and an author is curious to explore further, then begin examining the residual matrix for clues as to where misfit is occurring, adjust the model accordingly, refit, and proceed in this way to explore minor-adjustment alternative models until either fit is achieved, or where it becomes obvious that something is very wrong with the a priori theory".

At this point we take up the literature on inductive theory development. As we highlighted in the introduction, a key characteristic is the interplay between the developing theory and the data, the constant comparison of the two. Thus, we view the process of diagnosing misspecification of CB-SEM models and of improving models by comparing them to the data as very much related to inductive theory building. However, we acknowledge some differences. First, most CB-SEM based research starts out as a theory testing endeavor, thus providing an initial model which may subsequently be modified. In contrast, inductive theory development in qualitative research typically begins without an a-priori model (Urquhart et al., 2009). This however, is only a difference in degrees: The quantitative CB-SEM research could begin with a null model, in which no observed variable is related to any other. Model diagnostics, i.e. theory to data comparison, could begin with such an a-theoretic (and maximally misspecified) model. Second, theory construction from qualitative data has few restrictions on the available data. For example, depending on whether one follows Glaser's or Strauss & Corbin's approach to grounded theory, the researcher need only select a general area of inquiry or perhaps a specific phenomenon, and can gather data ad-hoc, as necessary. In contrast, the CB-SEM researcher is restricted to an a-priori defined phenomenon and has a set of pre-defined observations. Depending on the population that is sampled, it may or may not be practical to collect additional data. However, we believe that these differences do not change the fact that a constant comparison of theory and data, and a building of theory, is possible in CB-SEM, as we demonstrate in the following sections.

While Glaser and Strauss' (1967) work on grounded theory preceded the advent of CB-SEM, they write: "*The freedom and flexibility that we claim for generating theory from quantitative data will lead to new strategies and styles of quantitative analysis, with their own rules yet to be discovered*" (pg. 186, emphasis in original). This paper may be seen as an illustration of such a new analysis strategy.

Data

For illustrating how quantitative data can be used to generate rich theory, we use data from a previous study by Chin et al. (2008) on the Technology Acceptance Model (TAM). We use this example for a number of reasons. First, TAM is one of the most frequently used theories in IS research, well accepted, and familiar to most researchers. The measurement items have been virtually unchanged over many studies. Second, Chin et al. (2008) present one of the few commendable studies in IS research that publish covariance or correlation matrices to allow readers to independently verify their conclusions and extend their research, as we do here. Third, Chin et al. (2008) claim that their model fits the data well, despite strong evidence to the contrary. The authors overlook the results of the $\chi 2$ test that shows significant discrepancies, while at the same time trusting in the $\chi 2$ test for multi-group invariance testing. If the researchers do not trust the $\chi 2$ test for model fit, what makes them trust the same test for model differences? Furthermore, when the models do not fit the data, the $\chi 2$ statistic has a non-central distribution (Bollen, 1989), and it is therefore inappropriate to use a central-distribution $\chi 2$ difference test: "The practice of ignoring the global chi-square tests while at the same time conducting and interpreting chi-square difference tests between nested models should be prohibited as nonsensical." (Millsap, 2007, pg. 878)

While Chin et al. (2008) provide the covariance matrix, their reporting is incomplete in other aspects (Boomsma, 2000; McDonald and Ho, 2002). One important omission is the fit function. While the maximum-likelihood (ML) fit function is the most commonly used fit function, it assumes multivariate normality of the data. Chin et al. (2008) do not appear to have verified multivariate normality. Second, non-normal data can be addressed using Satorra-Bentler (SB) corrections to the $\chi 2$ statistic (Satorra and Bentler, 1994), but Chin et al. (2008) provide no information on

whether theirs is a corrected statistic. Third, Chin et al. (2008) do not provide the complete statistical model that would show whether items are reflective or formative, and whether errors are correlated. Nor do they later provide the estimated error variances and covariances. Finally, the provided covariance matrix is incomplete. Because Chin et al. (2008) collected information on two different TAM instruments from the same participants, covariances between the items of the two instruments should be available, but are not published.

To verify our assumptions about the estimation and fit methods, we first reproduce the results obtained by Chin et al. (2008). Using the reported sample size of 283, the ML method without SB corrections on a model with reflective indicators and uncorrelated errors reproduces exactly the reported results. We used multi-group comparisons to verify their claim in their footnote 6 that the factor structure is measurement invariant between the two instruments. We now turn to a more detailed analysis of the first TAM instrument, which employs the items from Davis (1989). The results presented by Chin et al. (2008) show a $\chi 2=181$ on df=101 for a highly significant p value, indicating that the model does not fit the observed covariance data. The complete model is shown in Figure 1 and the measurement items are shown in Table 2 (all items measured on 7-point Likert scales). The covariance matrix can be found in Appendix C of (Chin et al., 2008) and is therefore not reproduced here.

Table 1: Instrument items from (Chin et al., 2008), all measured on a 7-point Likert scale

Item	Wording
Perceived Ease of Use	
eou1	Learning to operate the (task-related) platform portions of (system) is easy for me
eou2	I find it easy to get the (task-related) portions of (system) to do what I want it to do
eou3	My interaction with the (task-related) portions of (system) has been clear and understandable.
eou4	I find the (task-related) portions of (system) to be flexible to interact with.
eou5	It is easy for me to become skillful at using the (task-related) portions of (system)
eou6	I find the (task-related) portions of (system) easy to use.
Perceived Usefulness	
use1	Using (system) as a (technology type) enables me to (accomplish tasks) more quickly
use2	Using (system) improves my (ability to accomplish task)
use3	Using (system) as a (technology type) increases my productivity
use4	Using (system) as a (technology type) increases my effectiveness in accomplishing (task)
use5	Using (system) makes it easier to do my (task)
use6	I find (system) useful in my (task completion)
Predicted Usage	
lu1	If the choice of a (technology type) platform were up to me, it would likely be (system).
lu2	If I need to (accomplish task) and the choice was up to me, I would expect to use (system) as a (task-related) platform.
lu3	If asked, I would likely recommend (system) as a (task-related) platform
lu4	For future (task-oriented) tasks that are totally within my control, I would probably use (system) as a (task-oriented) platform

Application

We began by estimating the complete model shown in Figure 1, which produced significant misfit ($\chi 2=182.25$, df=101, p=0.0000), as reported by Chin et al. (2008). Examining the residuals, i.e. the difference between model-implied and observed covariances, showed that the model explains the variances well but the covariances less so. However, there was no single variable that had unusually large residuals with other variables, nor were there any other patterns immediately identifiable in the residual matrix. Hence, we examined each construct individually, to assess whether Perceived Ease of Use, Perceived Usefulness, and Predicted Usage individually fit the factor models that were hypothesized for them.

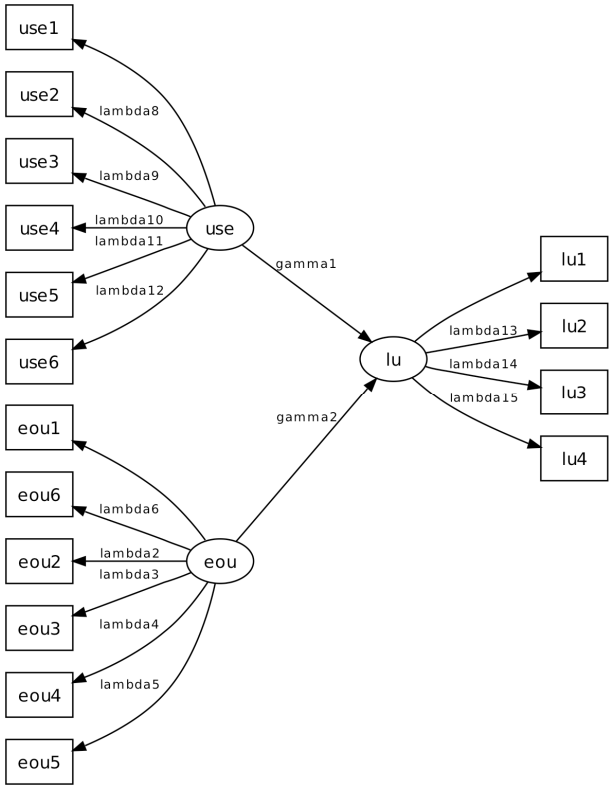

Figure 1: TAM Model

Perceived Ease of Use

With free error variances, uncorrelated errors, and one path coefficient constrained for scaling purposes, a six-indicator EOU factor model showed significant misfit ($\chi 2$=34.693, df=9, p=0.0001), showing that EOU is not a simple construct as hypothesized.

We therefore began to critically analyze the EOU item wording in detail. Table 1 shows that, rather than being homogeneous in meaning, i.e. being questions about the same underlying phenomenon, they are in fact different. Item EOU4 deals with flexibility, item EOU3 deals with clarity and understandability, items EOU1 and EOU5 are concerned with the ease of learning and mastering of a system and only EOU2 and EOU6 deal directly with ease of use. We therefore began modeling EOU2 and EOU6 as two items of a single EOU factor (Figure 2). To achieve identification, we chose to constrain error variances to 20% of the item variances. While this is in line with the estimates in the full TAM model, there is also a theoretical motivation. The error variance "quantifies your assessment of how similar or dissimilar your concept is to the best indicator." (Hayduk, 1996). We believe that such a small error variance is defensible because the item wording is strongly and directly related to the latent construct, Perceived Ease Of Use. This model showed good fit ($\chi 2$=0.1950, df=1, p=0.6588). Finally, we chose small error variances because, as Miles and Shevlin (2007) point out, the larger the error variances, the easier it is to fit the model.

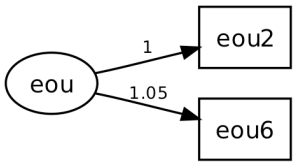

Figure 2: EOU items ($\chi 2$=0.1951, df=1, p=0.6587)

In the remainder of this article, we provide complete model specification and estimation results for all models to allow future critique of this analysis. We used the `sem` package in the open-source R system for our analysis. Models are specified by one line for each path. Regression paths are specified with `->` while covariance paths are specified by `<->`. The second entry on each line is the name of the parameter or NA if the parameter is fixed. The third entry is the value of a fixed parameter or the starting value for a free parameter. Covariance paths for observed variables represent error covariances; covariance paths for endogenous latents represent disturbance covariances. Comments begin with a `#`. For further details see (Fox, 2006). The model specification and estimation results for the model in Figure 2 follow.

```
# The EOU concept          | eou -> eou6, lambda1, 1     | eou2 <-> eou2, NA, 0.36
eou -> eou2, NA, 1         | eou <-> eou, phi1, NA       | eou6 <-> eou6, NA, 0.40
```

```
Parameter Estimates
        Estimate Std Error z value Pr(>|z|)
lambda1 1.0479   0.046374  22.5972 0        eou6 <--- eou
phi1    1.4793   0.154257   9.5896 0        eou <--> eou
```

Based on the item wording, EOU4 expresses Perceived Flexibility of the system, rather than Perceived Ease of Use. While a system that is perceived flexible may be perceived to be easy to use, the two are not the same. Hence, we model Perceived Flexibility as one cause (among other possible ones) of Perceived Ease of Use. Similarly, EOU3 expresses clarity and understandability, which may lead to a system that is perceived to be easy to use, but, again, the two concepts are not identical. The error variances of the indicator variables are again constrained at 20% of the item variances, for the reasons described above.

The model is shown in Figure 3 and the estimation showed good fit ($\chi2$=4.1967, df=3, p=0.241). Note that in the estimation results below, flexibility is not significantly related to ease of use. This is plausible, as a system that is flexible may offer more options to the user, making it more complex and hence harder to use, thus negating any inherent increase in ease of use by flexibility. A flexible system may also provide multiple ways of achieving the same result, possibly confusing the novice user and making it appear harder to use, further diminishing any positive effect that flexibility might have.

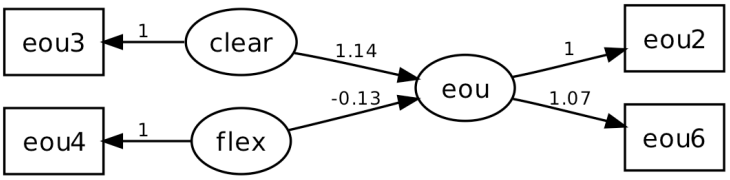

Figure 3: EOU, Flexibility, and Clarity ($\chi2$=4.1967, df=3, p=0.2410)

```
# The EOU part             | # Perceived Clarity         | flex -> eou4, NA, 1
eou -> eou2, NA, 1         | clear -> eou3, NA, 1        | eou4 <-> eou4, NA, 0.32
eou -> eou6, lambda1, 1.05 | eou3 <-> eou3, NA, 0.34     | flex <-> flex, phi3, NA
eou <-> eou, zeta1, NA     | clear <-> clear, phi2, NA   | flex -> eou, gamma2, NA
eou2 <-> eou2, NA, 0.36    | clear -> eou, gamma1, NA    | # Covariances
eou6 <-> eou6, NA, 0.40    | # Perceived Flexibility     | flex <-> clear, phi1, NA
```

```
Parameter Estimates
        Estimate  Std Error z value Pr(>|z|)
lambda1  1.068587 0.046081  23.1892 0.0000e+00 eou6 <--- eou
zeta1   -0.049852 0.048739  -1.0228 3.0638e-01 eou <--> eou
phi2     1.420000 0.148683   9.5505 0.0000e+00 clear <--> clear
gamma1   1.144443 0.205550   5.5677 2.5809e-08 eou <--- clear
phi3     1.328000 0.138944   9.5578 0.0000e+00 flex <--> flex
gamma2  -0.133897 0.207013  -0.6468 5.1776e-01 eou <--- flex
phi1     1.251000 0.125622   9.9585 0.0000e+00 clear <--> flex
```

Having established the ease of use latent as above, we must take care to prevent interpretational confounding. This is a situation in which, due to free parameters, a latent takes on new meaning by virtue of significantly different path coefficients to other variables. The path coefficients from EOU to EOU2 and EOU6 differ in the two estimations above (1.0479 versus 1.068587) and we need to test whether the differences are significant. Restricting the coefficient in the latter model to its value in the former allows a $\chi 2$ difference test with one degree of freedom, which, in this case, is non-significant ($\Delta\chi 2$=0.1657).

The item wording suggests that items EOU1 and EOU5 are related to the ease of learning a system, rather than the ease of using a system. Again, the concepts may be related, but they are not identical. In fact, the direction of causality is not even clear: If a system is easy to use, it may well be perceived to also be easy to learn. On the other hand, it is also plausible that if a person perceives a system to be easy to learn, she will also believe the system to be easy to use. Adding the Learnability concept with the indicators EOU1 and EOU5 into the model in Figure 3 proved to be unsuccessful. Modeling it as an independent cause of EOU showed misfit ($\chi 2$=47.657, df=10, p=0.0000), as did modeling it as consequence only of EOU ($\chi 2$=184.71, df=12, p=0.0000). Realizing that Perceived Flexibility and Perceived Clarity might influence Learnability in the same way as they impact Perceived Ease of Use, we modeled it as a consequence of both Perceived Flexibility and Perceived Clarity (Figure 4). While showing improvement, this model also did not fit our data ($\chi 2$=37.309, df=10, p=0.0001).

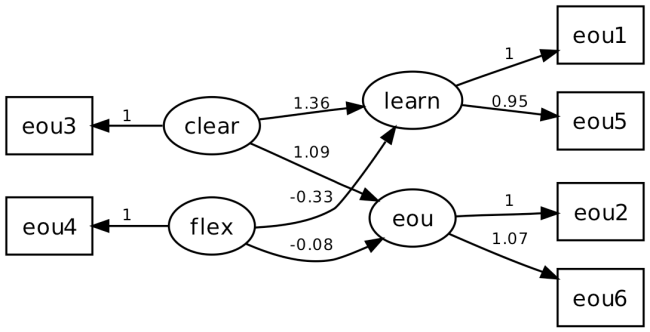

Figure 4: EOU, Learnability, Flexibility, and Clarity ($\chi 2$=37.309, df=10, p=0.0001)

```
# The EOU concept
eou -> eou2, NA, 1
eou -> eou6, lambda1, 1.05
eou <-> eou, zeta1, NA
eou2 <-> eou2, NA, 0.36
eou6 <-> eou6, NA, 0.40
# Perceived clarity
clear -> eou3, NA, 1
eou3 <-> eou3, NA, 0.34
clear <-> clear, phi2, NA
clear -> eou, gamma1, NA
# Perceived flexibility
flex -> eou4, NA, 1
eou4 <-> eou4, NA, 0.32
flex <-> flex, phi3, NA
flex -> eou, gamma2, NA
# Covariances
flex <-> clear, phi1, NA
# Learnability
learn <-> learn, zeta2, NA
learn -> eou1, NA, 1
learn -> eou5, lambda5, NA
eou1 <-> eou1, NA, 0.42
eou5 <-> eou5, NA, 0.4
# Causal relationships
flex -> learn, gamma3, NA
clear -> learn, gamma4, NA
```

```
Parameter Estimates
        Estimate   Std Error z value  Pr(>|z|)
lambda1  1.0709785 0.045459  23.55942 0.0000e+00 eou6 <--- eou
zeta1   -0.0354484 0.025560  -1.38687 1.6548e-01 eou <--> eou
phi2     1.4283542 0.147325   9.69527 0.0000e+00 clear <--> clear
gamma1   1.0916229 0.127883   8.53609 0.0000e+00 eou <--- clear
phi3     1.3281374 0.138984   9.55601 0.0000e+00 flex <--> flex
gamma2  -0.0838928 0.134661  -0.62299 5.3329e-01 eou <--- flex
phi1     1.2497466 0.125631   9.94774 0.0000e+00 clear <--> flex
zeta2    0.0097682 0.041329   0.23635 8.1316e-01 learn <--> learn
lambda5  0.9485977 0.041987  22.59262 0.0000e+00 eou5 <--- learn
gamma3  -0.3288684 0.185190  -1.77585 7.5758e-02 learn <--- flex
gamma4   1.3584442 0.175915   7.72216 1.1546e-14 learn <--- clear
```

Closely examining the residuals for these three models yielded no clues as to what caused the model misfit, as large residuals occurred between differing and unrelated variables. We finally examined whether items EOU1 and EOU5 (under the assumption of 20% error variance) are even caused by the same latent. The model showed significant misfit ($\chi2$=7.742, df=1, p=0.0054). This leads us to question the validity of items EOU1 and EOU5 (or the reliability of the items, if larger error variance would yield a fitting model). While their wording suggests they are related, and are not direct indicators of Perceived Ease of Use, none of the plausible models fit the data.

Perceived Usefulness

We applied the same approach to the Perceived Usefulness construct as we did for Perceived Ease of Use. An initial factor model with six indicators, free error variances, uncorrelated errors, and one path coefficient constrained for scaling yielded significant misfit ($\chi2$=32.607, df=9, p=0.0002).

While the items for Perceived Usefulness appear more homogeneous than the items for EOU, we can still make out distinct groups. Items USE1, USE2, and USE3 deal with efficiency, performance, and productivity, i.e. with amount of work per time unit. On the other hand, USE4 deals with effectiveness, which is quite distinct from efficiency. USE6 is the most immediate and direct measure of usefulness. The only problematic item is USE5, which deals with making the task easier. While this is clearly related to usefulness, effectiveness, and efficiency, the causal relationships are not immediately clear. A system can be effective without making the task any easier, or it can make the task easier, but not be effective. Similarly, efficiency and easing of the task are not necessarily related. Separating efficiency, effectiveness, and usefulness leads to the model in Figure 5 (free error variances where possible, uncorrelated errors, one path coefficient constrained for scaling each latent), which fits the observed data ($\chi2$=7.6693, df=4, p=0.1045).

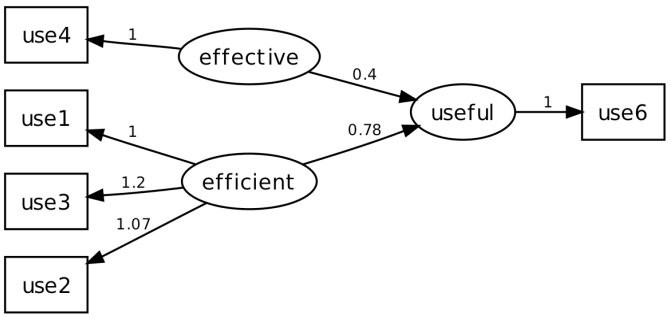

Figure 5: Usefulness model ($\chi2$=7.6693, df=4, p=0.1045)

```
# Perceived efficiency
efficient -> use1, NA, 1
efficient -> use2, lambda2, 1
efficient -> use3, lambda3, 1
efficient <-> efficient, phi1, NA
use1 <-> use1, delta1, NA
use2 <-> use2, delta2, NA
use3 <-> use3, delta3, NA
# Perceived effectiveness
effective -> use4, NA, 1
use4 <-> use4, NA, 0.24
effective <-> effective, phi2, NA
effective <-> efficient, phi3, 1
# Perceived usefulness
useful -> use6, NA, 1
useful <-> useful, zeta1, 1
use6 <-> use6, NA, 0.30
# Causal relationships
efficient -> useful, gamma1, NA
effective -> useful, gamma2, NA
```

```
Parameter Estimates
        Estimate Std Error z value Pr(>|z|)
lambda2 1.07423  0.071373  15.0511 0.0000e+00 use2 <--- efficient
lambda3 1.20188  0.077912  15.4261 0.0000e+00 use3 <--- efficient
phi1    0.63996  0.082885   7.7211 1.1546e-14 efficient <--> efficient
```

```
delta1  0.37504  0.039071   9.5988 0.0000e+00 use1 <--> use1
delta2  0.36750  0.039643   9.2701 0.0000e+00 use2 <--> use2
delta3  0.38656  0.043796   8.8264 0.0000e+00 use3 <--> use3
phi2    0.97200  0.102150   9.5155 0.0000e+00 effective <--> effective
phi3    0.72728  0.077438   9.3918 0.0000e+00 efficient <--> effective
zeta1   0.13732  0.045717   3.0037 2.6670e-03 useful <--> useful
gamma1  0.78095  0.273660   2.8537 4.3210e-03 useful <--- efficient
gamma2  0.39921  0.220848   1.8076 7.0667e-02 useful <--- effective
```

The Exogenous Latents

At this stage we combine the two exogenous concepts, Perceived Ease of Use and Perceived Usefulness. While the initial TAM results suggest that Perceived Ease of Use influences Perceived Usefulness (Davis, 1989), Chin et al., (2008) model them as merely exogenously co-varying. As we have additional concepts available, i.e. Perceived Clarity and Perceived Flexibility, which are both exogenous, we hypothesize that increases in Perceived Flexibility might also cause increased Perceived Usefulness, irrespective of the Perceived Ease of Use. As we have seen above (Figure 4), a flexible system can plausibly be argued to be harder to use than an inflexible one. However, it is also plausible that a flexible system is more useful, as it is likely to be able to accomplish more tasks, or allow more efficient ways to perform the task. This could serve as an explanation for the effects of EOU on Usefulness suggested by Davis (1989). Hence, a first model to be tested is shown in Figure 6. Again, error variances are free where possible, errors are uncorrelated, and all four exogenous latents are allowed to co-vary. This model fits the data well ($\chi 2$=28.756, df=22, p=0.1520).

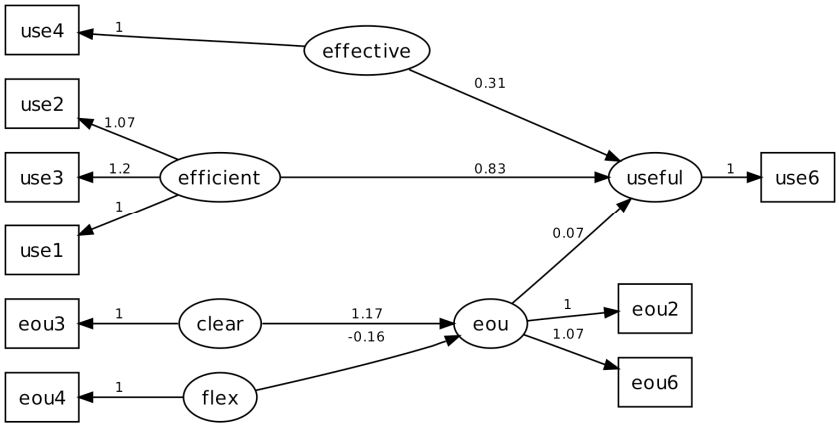

Figure 6: Exogenous latents ($\chi 2$=28.756, df=22, p=0.1520)

```
# The EOU concept
eou -> eou2, NA, 1
eou -> eou6, lambda1, 1
eou <-> eou, zeta1, NA
eou2 <-> eou2, NA, 0.36
eou6 <-> eou6, NA, 0.40
# Perceived clarity
clear -> eou3, NA, 1
eou3 <-> eou3, NA, 0.34
clear <-> clear, phi2, NA
clear -> eou, gamma1, NA
# Perceived flexibility
flex -> eou4, NA, 1
eou4 <-> eou4, NA, 0.32, NA
flex <-> flex, phi3, NA
flex -> eou, gamma2, NA
flex <-> clear, phi1, NA
# Perceived efficiency
efficient -> use1, NA, 1
efficient -> use2, lambda2, 1
efficient -> use3, lambda3, 1
efficient <-> efficient, phi4, NA
use1 <-> use1, delta1, NA
use2 <-> use2, delta2, NA
use3 <-> use3, delta3, NA
# Perceived effectiveness
effective -> use4, NA, 1
use4 <-> use4, NA, 0.24
effective <-> effective, phi5, NA
effective <-> efficient, phi6, 1
```

```
# Perceived usefulness
useful -> use6, NA, 1
useful <-> useful, zeta2, 1
use6 <-> use6, NA, 0.30
# Causal relationships
efficient -> useful, gamma3, NA
effective -> useful, gamma4, NA
eou -> useful, gamma5, NA
# Exogenous covariances
flex <-> efficient, phi7, NA
flex <-> effective, phi8, NA
clear <-> efficient, phi9, NA
clear <-> effective, phi10, NA
```

```
 Parameter Estimates
        Estimate  Std Error z value  Pr(>|z|)
lambda1  1.066844 0.045994  23.19538 0.0000e+00 eou6 <--- eou
zeta1   -0.047657 0.049361  -0.96548 3.3431e-01 eou <--> eou
phi2     1.420703 0.148752   9.55080 0.0000e+00 clear <--> clear
gamma1   1.166287 0.206443   5.64943 1.6098e-08 eou <--- clear
phi3     1.327058 0.138964   9.54962 0.0000e+00 flex <--> flex
gamma2  -0.158570 0.207667  -0.76358 4.4512e-01 eou <--- flex
phi1     1.254636 0.125566   9.99181 0.0000e+00 clear <--> flex
lambda2  1.073017 0.071159  15.07922 0.0000e+00 use2 <--- efficient
lambda3  1.198657 0.077625  15.44154 0.0000e+00 use3 <--- efficient
phi4     0.640068 0.082960   7.71541 1.1990e-14 efficient <--> efficient
delta1   0.374932 0.039052   9.60075 0.0000e+00 use1 <--> use1
delta2   0.369047 0.039516   9.33912 0.0000e+00 use2 <--> use2
delta3   0.391364 0.043708   8.95405 0.0000e+00 use3 <--> use3
phi5     0.974642 0.102297   9.52758 0.0000e+00 effective <--> effective
phi6     0.729410 0.077559   9.40463 0.0000e+00 efficient <--> effective
zeta2    0.133213 0.045175   2.94881 3.1900e-03 useful <--> useful
gamma3   0.834740 0.292963   2.84930 4.3816e-03 useful <--- efficient
gamma4   0.309456 0.227977   1.35740 1.7465e-01 useful <--- effective
gamma5   0.067931 0.045673   1.48735 1.3692e-01 useful <--- eou
phi7     0.511513 0.075680   6.75889 1.3905e-11 efficient <--> flex
phi8     0.637746 0.092410   6.90123 5.1552e-12 effective <--> flex
phi9     0.495305 0.074123   6.68225 2.3531e-11 efficient <--> clear
phi10    0.562385 0.089211   6.30402 2.9003e-10 effective <--> clear
```

To assess interpretational confounding, we conduct $\chi 2$ difference tests by restricting each path coefficient in Figure 6 to its value in Figure 4 or Figure 5. All $\chi 2$ difference tests are non-significant, showing that path coefficients, and hence the meaning of the latent concepts, remain unchanged.

Predicted Usage

The final concept, Predicted Usage, is represented by an endogenous latent variable with four hypothesized indicators. Examining the indicators closely shows that LU3 is slightly different as it deals with recommendations for others, rather than subject's own future intended usage. However, a test of the hypothesized four indicator factor model, with free errors, uncorrelated errors, and one path coefficient fixed for scaling showed good fit ($\chi 2$=1.6517, df=2, p=0.4379), so that we used this sub-model for the next step.

Full Model

We are now ready to assemble the full model, based on the model in Figure 6. The TAM theory suggests that both Perceived Usefulness and Perceived Ease of Use cause future intentions to use a system (Chin et al., 2008; Davis, 1989), so we introduce these two paths. We estimate the model again with errors free where possible, uncorrelated errors, and correlated exogenous latents. The model fits the data ($\chi 2$=67.338, df=55, p=0.1228). This model is shown in Figure 7.

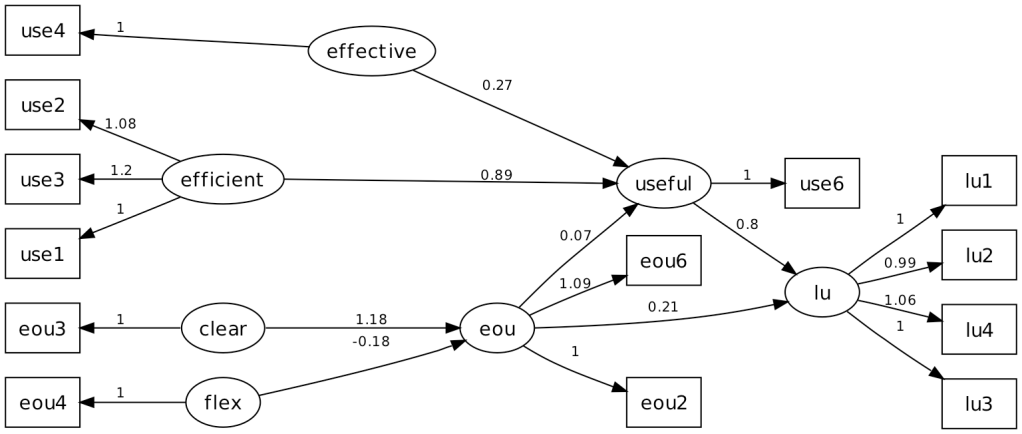

Figure 7: Full model (χ2=67.338, df=55, p=0.1228)

```
# The EOU concept
eou -> eou2, NA, 1
eou -> eou6, lambda1, 1
eou <-> eou, zeta1, NA
eou2 <-> eou2, deltaeou2, 0.36
eou6 <-> eou6, deltaeou6, 0.40
# Perceived clarity
clear -> eou3, NA, 1
eou3 <-> eou3, NA, 0.34
clear <-> clear, phi2, NA
clear -> eou, gamma1, NA
# Perceived flexibility
flex -> eou4, NA, 1
eou4 <-> eou4, NA, 0.32, NA
flex <-> flex, phi3, NA
flex -> eou, gamma2, NA
flex <-> clear, phi1, NA
# Perceived efficiency
efficient -> use1, NA, 1
efficient -> use2, lambda2, 1
efficient -> use3, lambda3, 1
efficient <-> efficient, phi4, NA
use1 <-> use1, delta1, NA
use2 <-> use2, delta2, NA
use3 <-> use3, delta3, NA
# Perceived effectiveness
effective -> use4, NA, 1
use4 <-> use4, NA, 0.24
effective <-> effective, phi5, NA
effective <-> efficient, phi6, 1
# Perceived usefulness
useful -> use6, NA, 1
useful <-> useful, zeta2, 1
use6 <-> use6, deltause6, 0.30
# Casual relationships
efficient -> useful, gamma3, NA
effective -> useful, gamma4, NA
eou -> useful, gamma5, NA
# Exogenous covariances
flex <-> efficient, phi7, NA
flex <-> effective, phi8, NA
clear <-> efficient, phi9, NA
clear <-> effective, phi10, NA
# Causal relationships
eou -> lu, gamma6, NA
useful -> lu, gamma7, NA
# Intention to use concept
lu -> lu1, NA, 1
lu -> lu2, lambda7, 1
lu -> lu4, lambda8, 1
lu -> lu3, lambda9, 1
lu <-> lu, phi50, NA
lu1 <-> lu1, deltalu1, 0.40
lu2 <-> lu2, deltalu2, 0.36
lu4 <-> lu4, deltalu3, 0.42
lu3 <-> lu3, deltalu4, 0.40
```

```
 Parameter Estimates
          Estimate  Std Error z value  Pr(>|z|)
lambda1    1.089881 0.049595  21.97582 0.0000e+00 eou6 <--- eou
zeta1     -0.062519 0.050980  -1.22635 2.2007e-01 eou <--> eou
deltaeou2  0.431831 0.047618   9.06867 0.0000e+00 eou2 <--> eou2
deltaeou6  0.352191 0.046959   7.49999 6.3727e-14 eou6 <--> eou6
phi2       1.419199 0.148732   9.54196 0.0000e+00 clear <--> clear
gamma1     1.179635 0.212421   5.55329 2.8034e-08 eou <--- clear
phi3       1.326353 0.138971   9.54409 0.0000e+00 flex <--> flex
gamma2    -0.183003 0.213274  -0.85806 3.9086e-01 eou <--- flex
phi1       1.256735 0.125539  10.01071 0.0000e+00 clear <--> flex
lambda2    1.080303 0.071361  15.13857 0.0000e+00 use2 <--- efficient
lambda3    1.204050 0.077735  15.48920 0.0000e+00 use3 <--- efficient
```

```
phi4        0.635372 0.082601   7.69210 1.4433e-14 efficient <--> efficient
delta1      0.379627 0.038907   9.75736 0.0000e+00 use1 <--> use1
delta2      0.364484 0.038589   9.44536 0.0000e+00 use2 <--> use2
delta3      0.389874 0.042826   9.10378 0.0000e+00 use3 <--> use3
phi5        0.973554 0.102237   9.52248 0.0000e+00 effective <--> effective
phi6        0.726194 0.077345   9.38902 0.0000e+00 efficient <--> effective
zeta2       0.016422 0.050580   0.32468 7.4543e-01 useful <--> useful
deltause6   0.415821 0.060199   6.90744 4.9349e-12 use6 <--> use6
gamma3      0.887276 0.259145   3.42386 6.1738e-04 useful <--- efficient
gamma4      0.271192 0.199119   1.36196 1.7321e-01 useful <--- effective
gamma5      0.065920 0.046114   1.42949 1.5286e-01 useful <--- eou
phi7        0.512962 0.075466   6.79721 1.0666e-11 efficient <--> flex
phi8        0.641682 0.092317   6.95086 3.6307e-12 effective <--> flex
phi9        0.489540 0.073684   6.64374 3.0582e-11 efficient <--> clear
phi10       0.557394 0.089074   6.25767 3.9076e-10 effective <--> clear
gamma6      0.205960 0.065753   3.13231 1.7344e-03 lu <--- eou
gamma7      0.795834 0.081594   9.75359 0.0000e+00 lu <--- useful
lambda7     0.986071 0.041914  23.52587 0.0000e+00 lu2 <--- lu
lambda8     1.064321 0.043661  24.37685 0.0000e+00 lu4 <--- lu
lambda9     1.000093 0.042114  23.74735 0.0000e+00 lu3 <--- lu
phi50       0.655979 0.078837   8.32066 0.0000e+00 lu <--> lu
deltalu1    0.402543 0.042545   9.46161 0.0000e+00 lu1 <--> lu1
deltalu2    0.336317 0.037043   9.07917 0.0000e+00 lu2 <--> lu2
deltalu3    0.318425 0.037697   8.44689 0.0000e+00 lu4 <--> lu4
deltalu4    0.320249 0.036215   8.84303 0.0000e+00 lu3 <--> lu3
```

We are now in a position to try to add the uncooperative items USE5, EOU1, and EOU5 into the model. At this stage, there are more possibilities for these items to fit into a theoretically plausible model. Additionally, because of the larger model, adding these items adds more degrees of freedom, which might permit a better fit statistically. We begin by adding USE5 as another indicator of Perceived Usefulness (Figure 8). The $\chi 2$ statistic shows that the model does not fit the data ($\chi 2$=89.837, df=67, p=0.0328). This confirms the earlier problems with USE5 and supports the conclusion that USE5 is not an indicator of Perceived Usefulness.

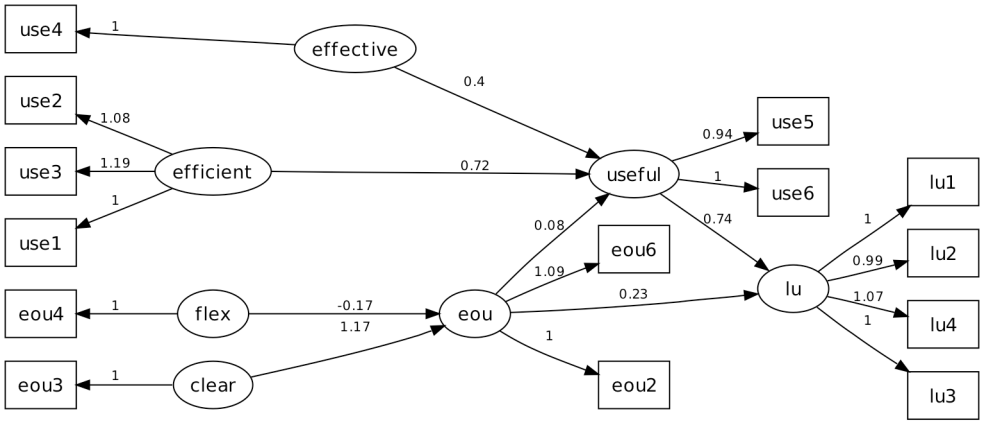

Figure 8: Full model with USE5 ($\chi 2$=89.837, df=67, p=0.0328)

Instead of trying to fit USE5, our attempt to add EOU5 as an indicator of Learnability, which in turn is caused by Perceived Ease of Use, shows more promise ($\chi 2$=80.204, df=67, p=0.1292) and is shown in Figure 9. Adding EOU1 as another indicator of Learnability reduces model fit to unacceptable levels ($\chi 2$=123.05, df=79, p=0.0011). Adding both EOU1 and USE5 (because of the combined degrees of freedom they contribute) also does not lead to improvement ($\chi 2$=151.85, df=94, p=0.0001). Hence, the most complex plausible model that can be fitted to the data is that in Figure 9.

Because we have been unable to fit all observed variables, our model has fewer degrees of freedoms than the one presented in Chin et al. (2008). Hence, it may be argued that the good fit is due to low power. While power estimates in structural equation modeling are problematic due to the lack of an alternative hypothesis (Saris and

Satorra, 1993) and ideally require a simulation study (Muthén and Muthén, 2002), indicative power calculations using the method of MacCallum et al. (1996) show that a model with 67 df and a sample size of 283 has a power of .943 to detect deviations from exact fit.

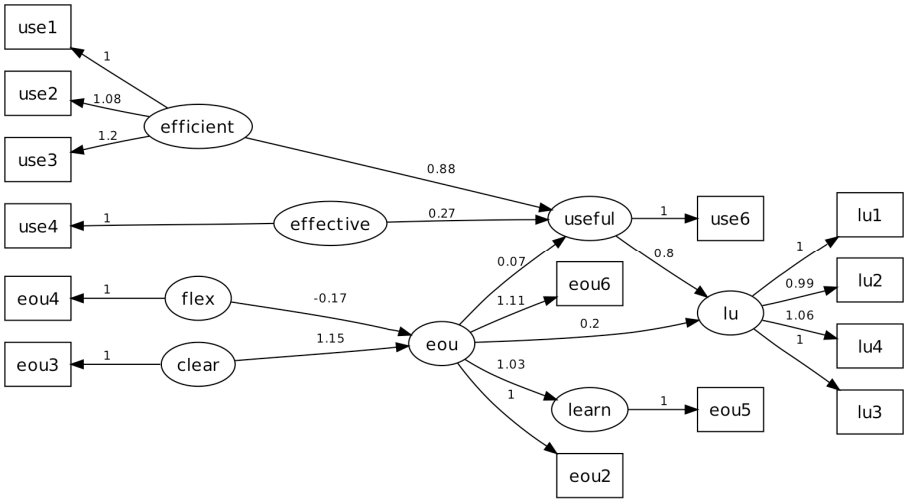

Figure 9: Full model with EOU5 (χ2=80.204, df=67, p=0.1292)

Discussion and Conclusion

We set out to achieve three goals in this paper. First, we wanted to demonstrate that it is possible to fit realistically sized CB-SEM models with realistic sample sizes. Our final model in Figure 9 has non-significant χ2 fit. We believe that the IS community needs to adopt stronger measures of model fit. As we have shown, it is unlikely that a, however well-conceived, a-priori model fits well. However, it is unhelpful when journal editors and reviewers insist on authors presenting "the" model and showing that it is correct. In the face of these inflexible requirements, it is perhaps not surprising that authors grasp at every straw to argue that their model fits well, despite significant χ2 fit. While we agree that an a-priori model should be tested without "peeking at the data", this is only sensible if, as a community, we are prepared to deal fairly with the outcome. As Hayduk et al. (2007, pg. 845) note, "we argue that attentively constructed and theoretically meaningful models that fail ought to be carefully discussed and published." Alternatively, if we are unwilling to acknowledge failures, we need at least be open to post-hoc model modifications of the type presented here. We believe that this process, in a briefer form, should be part of the discussion section of every paper. Our community is not well served if we have our collective heads in the sand and simply ignore the fact that most of our models fail to fit.

Second, we set out to demonstrate that, in trying to fit a CB-SEM model, we can generate theory. We believe that our final model in Figure 9 has achieved this. Through the constant comparison of theory and data we have identified five new theoretical constructs, "perceived effectiveness", "perceived efficiency", "perceived clarity of interaction", "perceived system flexibility", and "perceived learnability", that are linked in a nomological theory network to the existing TAM constructs. These constructs and their relationships deserve, and enable, further theoretical attention and empirical investigation. They can also provide links to related fields of research, such as learning and educational theories ("perceived learnability") or human-computer interaction ("perceived clarity", "perceived flexibility"). Additionally, in contrast to TAM, our model explains EOU and USE as endogenous variables. The additional constructs are not only theoretically interesting, they are also practically important, as they provide more specific guidance to interventions to improve the usage of information systems. In Dubin's (1969) terms, we have built theory by seeking intervening variables, a process that starts with the "admission that the starting theoretical model is inadequate and must be supplemented" (pg. 81). Finally, even the "discarded" items EOU1 and USE5 are theoretically useful, because "the theory builder is often well advised to inquire about the data that researchers collect but subsequently exclude from their research analysis. These data may be mined for important insights about new units" (Dubin, 1969, pg. 84).

As we introduced each new construct into the model, we have emphasized an at least plausible reason for the relationships in our models, because "any post-hoc model modification should be defensible on theoretical grounds,

as opposed to being purely data-driven.“ (McCallum, 2007, pg. 862). We do not propose or condone “fishing expeditions” in the data, e.g. by blindly chasing modification indices. Model changes should at least be plausible, if not supported by existing theories. The process demonstrated in this paper is clearly different from automatic model specification approaches based purely on modification indices (Marcoulides and Drezner, 2003; Schumacker, 2006).

Markland (2007, pg. 856) notes that “such model modifications based on observed discrepancies might be capitalizing on chance sampling fluctuations in the data, improving fit at the expense of theoretical meaningfulness. … Researchers who engage in this exercise … must justify any modifications they make and, preferably, any resulting respecified models should be replicated with independent data.” His caution mirrors that in qualitative inductive theory building. For example, Eisenhardt (1989, pg. 547) cautions us that “the risks are that the theory describes a very idiosyncratic phenomenon or that the theorist is unable to raise the level of generality of the theory.” We do not believe that quantitative theory building as shown here is in way different from that of qualitative inductive theory building and that researchers need to be explicit about the limitations in either case.

Another characteristic of the generated theory is that it lacks the parsimony of the original TAM model. The lack of parsimony is a potential issue that has long been recognized in the qualitative theory building literature. As Eisenhardt (1989, pg. 847) notes, “there is a temptation to build theory which tries to capture everything. The result can be theory which is very rich in detail, but lacks the simplicity of overall perspective.” However, a complex but well-fitting model, is preferable to a simple one that does not fit: “If the model $\chi 2$ test detects a causally mis-specified model, the biased estimates … become impotent and unconvincing.” (Hayduk et al, 2007, pg. 845). In general, the parameter estimates from a mis-specified model should not be substantively interpreted.

As a third goal, we set out to emphasize the importance of instrument construction and the lack of attention paid to it. Instrument construction is also related to the issue of parsimony due to the pervasive use of the factor model in IS research. The factor model offers great parsimony, but has limitations if the instrument is not very carefully constructed, as we have shown in this study where we examined a factor-analysis based instrument and highlighted the significant misfit to the data. We believe that careless construction of an instrument, an insistence on the factor model, and consequent neglect of interrelationships between subtle aspects of the items are a frequent cause of problems with model fit. One solution is to more carefully construct factor-based measures. However, researchers believed the measures in TAM are well specified reflectively, i.e. equivalent. Given our analysis here, we believe it is unlikely that, no matter how carefully constructed, as the TAM items have been, multiple items can be completely semantically identical, so that the factor model (and the notions of validity and reliability built on it) is no longer appropriate. Thus, we recommend IS researchers to embrace the additional flexibility that CB-SEM offers. CB-SEM is *not* just the combination of factor analysis with regressions between factors but can test models that may be more realistic than factor-analytic ones.

Moving away from the factor model also means that the recent debate over formative versus reflective measurement (Hardin et al. 2008a;b; Marakas et al. 2007; 2008; Petter et al. 2007) becomes a non-issue, as it is only meaningful in the traditional factor model. The most important message to take away from Petter et al. (2007) is to pay increased attention to the causality between indicators and latent variables. We agree with this, and urge researchers not to get mired in the terminological debate of “formative” or “reflective”.

We note two important limitations. First, as we indicated earlier, model fit is necessary, but not sufficient, for identifying the correct model: “Finding a model that fits the covariance data does not say the model is the correct model, but merely that the model is one of the several potentially very causally different models that are consistent with the data.” (Hayduk et al., 2007, pg. 843) Consequently, even well fitting models should be carefully and critically examined. We believe this caution applies also to qualitative theory building approaches, where ruling out alternative explanations has not been as widely addressed as generating plausible explanations. Second, we add the caution that our final model is unsupported by evidence other than what it is generated from. It has thus the status of untested theory and must be tested against an independently collected sample. In this, its status is no different to a theory that is developed from qualitative data, which must be testable and subsequently be tested (Eisenhardt, 1989).

We close with some recommendations for researchers, reviewers, and editors. We believe that the IS research field needs to pay increased attention to model fit and not disregard important evidence. We also believe that there is value in publishing and discussing models that do not fit, if they are theoretically well motivated, with sensible indicators and data has been carefully collected. Clinging to misfitting factor models is not helpful in driving theory forward: The data has much more to offer, as we have shown in this example, and we believe it is incumbent upon us not to waste the enormous effort that typically goes into theorizing and data collection.

References

Bacharach, S.B. 1989. "Organizational Theories: Some Criteria for Evaluation," *The Academy of Management Review*, (14:4), pp. 496–516.

Barrett, P. 2007. "Structural equation modelling: Adjudging model fit," *Personality and Individual Differences*, (42), pp. 815–824.

Beauducel A. and Wittmann, W. 2005. "Simulation study on fit indices in confirmatory factor analysis based on data with slightly distorted simple structure," *Structural Equation Modeling*, (12:1), pp. 41–75.

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

Bentler, P. M. and Weeks, D. G. 1980. "Linear structural equations with latent variables," *Psychometrika*, (30:4), pp. 289–308.

Bollen, K. 1989. *Structural Equations with Latent Variables*. New York, NY: John Wiley and Sons.

Boomsma, A. 2000. "Reporting analyses of covariance structures," *Structural Equation Modeling*, (7:3), pp. 461–483.

Chin, W. W., Johnson, N., and Schwarz, A. 2008. "A fast form approach to measuring technology acceptance and other constructs," *MIS Quarterly*, (32:4), pp. 687–703.

Davis, F.D. 1989. "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS Quarterly*, (13:3), pp. 319–340.

Dubin, R. 1969. *Theory Building*. New York, NY: Free Press.

Eisenhardt, K. M. 1989. "Building theories from case study research," *The Academy of Management Review*, (14:4), pp. 532–550.

Fan, X. and Sivo, S.A. 2005. "Sensitivity of fit indices to misspecified structural or measurement model components: rationale of two-index strategy revisited," *Structural Equation Modeling*, (12:3), pp. 343–367.

Fox, J. 2006. "Structural equation modeling with the `sem` package in R," *Structural Equation Modeling*, (13:3), pp. 465–486.

Gefen, D. and Straub, D.W. and Boudreau, M.-C. 2000. "Structural Equation Modeling and Regression: Guidelines for research practice," *Communications of the AIS,* (4:7), pp 1–70.

Glaser, B.G. and Strauss, A.L. 1967. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Chicago, IL: Aldine Publishing Company.

Goffin, R. D. 2007. "Assessing the adequacy of structural equation models: Golden rules and editorial policies," *Personality and Individual Differences*, (42), pp. 831–839.

Gregor, S. 2006. "The nature of theory in information systems," *MIS Quarterly*, (30:3), 611–642.

Hair, J. E., Anderson, R. E., Tatham, R. L., and Black, W. C. 2005. *Multivariate Data Analysis, 6th ed.*, Upper Saddle River, NJ: Pearson/Prentice Hall.

Hardin, A. M., Chang, J. C.-J., and Fuller, M. A. 2008a. "Formative vs. reflective measurement: Comment on Marakas, Johnson, and Clay (2007)," *Journal of the Association for Information Systems*, (9:9), pp. 519–534.

Hardin, A. M., Chang, J. C.-J., and Fuller, M. A. 2008b. "Clarifying the use of formative measurement in the IS discipline: The case of the computer self-efficacy," *Journal of the Association for Information Systems*, (9:9), pp. 544–546.

Hayduk, L., Cummings, G., Boadu, K., Pazderka-Robinson, H., and Boulianne, S. 2007. "Testing! testing! One, two, three - testing the theory in structural equation models!" *Personality and Individual Differences*, (42), pp. 841–850.

Hayduk, L. A. 1996. *LISREL Issues, Debates, and Strategies*. Baltimore, MD: Johns Hopkins University Press.

MacCallum, R.C., Browne, M.W. and Sugawara, H.M. 1996. "Power analysis and determination of sample size for covariance structure modeling," *Psychological Methods, (1:2), pp. 130*–149.

Marakas, G. M., Johnson, R. D., and Clay, P. F. 2007. "The evolving nature of the computer self-efficacy construct: An empirical investigation of measurement construction, validity, reliability and stability over time," *Journal of the Association for Information Systems*, (8:1), pp. 16–46.

Marakas, G. M., Johnson, R. D., and Clay, P. F. 2008. "Formative vs. reflective measurement: A reply to Hardin, Chang, and Fuller," *Journal of the Association for Information Systems*, (9:9), pp. 535–543.

Marcoulides, G. A. and Drezner, Z. 2003. "Model specification searches using ant colony optimization algorithms," *Structural Equation Modeling*, (10:1), pp. 154–164.

Markland, D. 2007. "The golden rule is that there are no golden rules: A commentary on Paul Barett's recommendations for reporting model fit in structural equation modeling," *Personality and Individual Differences*, (42), pp. 851–858.

Marsh, H.W., Hau, K.-T., and Wen, Z. 2004. "In search of golden rules: comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings," *Structural Equation Modeling*, (11:3), pp. 320–341.

McDonald, R.P. and Ho, M.-H.R. 2002. "Principles and practice in reporting structural equation analyses," *Psychological Methods*, (7:1), pp. 64–82.

McIntosh, C. N. 2007. "Rethinking fit assessment in structural equation modelling: A commentary and elaboration on Barrett (2007)," *Personality and Individual Differences*, (42), pp. 859–867.

Miles, J. and Shevlin, M. 2007. "A time and a place for incremental fit indices," *Personality and Individual Differences*, (42), pp. 869–874.

Millsap, R. E. 2007. "Structural equation modeling made difficult," *Personality and Individual Differences*, (42), pp. 875–881.

Muthén, L.K. and Muthén, B.O. 2002 "How to use a Monte-Carlo study to decide on sample size and determine power," *Structural Equation Modeling,* (9:4), pp. 599–620.

Petter, S., Straub, D., and Rai, A. 2007. "Specifying formative constructs in information systems research," *MIS Quarterly*, (31:4), pp. 623–656.

Popper, K. 1968. *The Logic of Scientific Discovery*. New York, NY: Harper & Row.

Saris, W.E. and Sotorra, A. 1993. "Power evaluations in structural equation models", in *Testing Structural Equation Models* , Bollen, K.A. and Long, J.S. (Ed.), Sage Publications, Newbury Park, CA.

Satorra, A. and Bentler, P. 1994. "Corrections to test statistics and standard errors in covariance structure analysis.," in *Latent variable analysis: Applications to developmental research*, pp. 399–419.

Schumacker, R. E. 2006. "Conducting specification searches with Amos," *Structural Equation Modeling*, (13:1), pp. 118–129.

Steiger, J. H. 2007. "Understanding the limitations of global fit assessment in structural equation modeling," *Personality and Individual Differences*, (42), pp. 893–898.

Urquhart, C., Lehmann, H., and Myers, M. D. 2009 "Putting the 'theory' back in grounded theory: Guidelines for grounded theory studies in information systems," *Information Systems Journal*, forthcoming, pp. 1–25.

Yuan, K.H. 2005. "Fit indices versus test statistics," *Multivariate Behavioral Research*, (40:1), pp. 115–148.