Burnout and job engagement in dentistry

te Brake, J.H.M.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
CHAPTER 3

On the phases of burnout: A longitudinal study using multiple imputation

Summary – This study uses a longitudinal design to examine the sequence of the three subscales of the Maslach Burnout Inventory (MBI): emotional exhaustion (EE), depersonalisation (D), and personal accomplishment (PA). Special precaution was taken to address the issue of wave nonresponse, which is an ubiquitous problem within longitudinal research. Multiple imputation (MI) was used in completing the original data, thereby offering a more advanced approach than the often used case deletion (CD). Using Structural Equation Modelling (SEM), the fit of several chronological models proposed in earlier research was compared among a representative sample of Dutch dentists. Results indicate that the original model suggested by Maslach and Jackson (1981) (EE→D→PA) showed an adequate fit, although an alternative model (PA→EE→D) showed an even better fit. However, an exploratively constructed, and empirically based ‘best fitting’ model indicated that EE should not be discarded as an early sign of burnout. Also, PA varies in the position it takes in relation to EE.

The Maslach Burnout Inventory (MBI - Maslach et al., 1996) is the most widely used instrument to measure burnout in current scientific research. There are important practical reasons to examine the chronological sequence of the three MBI subscales (emotional exhaustion – EE, depersonalisation – D and personal accomplishment – PA). First, knowledge of the sequence of development of burnout implies knowledge on those factors that are probable to surface in the early stages of the onset of burnout. Such an early recognition of burnout is considered essential for possible intervention. Second, from a preventive point of view, efforts to avoid a possible burnout can be

4 This chapter is submitted for publication.
concentrated on those factors that are known to develop first. By knowing the sequence of burnout early recognition and prevention can be facilitated (Brouwers, 2000; Van Dierendonck, 1997). Finally, knowledge on this sequence can help develop and empirically improve process models that can add to a further understanding of the antecedents and consequences of burnout (Lee & Ashforth, 1993, In: Van Dierendonck, 1997). In earlier research, structural equation modelling (SEM) was used to investigate the chronological order of the subscales. Although SEM is convenient for this purpose, a longitudinal design is necessary to rule out the reverse causation possibility (Cordes et al., 1997; Zapf, Dormann & Frese, 1996) and, overall, to make an empirically more viable point of chronology and causality.

Several models of sequence are proposed. First, the original EE→D→PA sequence proposed by Maslach and Jackson (1981) found some support in earlier research (e.g., Cordes et al., 1997; De Vries, 2001). Second, the so-called progressive phase model (Golembiewski & Munzenrider, 1988) proposes a D→PA→EE sequence. Although the progressive phase model also found some evidence (e.g., Gryskiewicz & Buttner, 1992), it was severely criticized on methodological grounds by Leiter (1989). Third, in a study among physiotherapists, evidence was recently found for a sequence of PA→EE→D (De Vries, 2001). In addition to these three models, the current study will focus on an additional sequence. In a longitudinal study, Van Dierendonck, Schaufeli and Buunk (2001) report evidence for the sequence PA→D→EE among general physicians. Although conceptually this model constitutes a reversal of the Maslach et al. (1981) model, in a mathematical sense the two models are identical. As a result, structural equation modelling based on cross-sectional data would yield identical results; only in a longitudinal design these two models can be distinguished from each other. In short, the present study seeks not only to examine several models of burnout in a longitudinal setting, but it also seeks to extend the models that are compared.

Within the social sciences, and especially within longitudinal research, missing data are a ubiquitous problem. Modern missing data procedures are available that are considered to be superior to Case Deletion (CD, also known as listwise deletion and pairwise deletion). One such advanced procedure is Multiple Imputation (MI) (e.g.,
Schafer, 1997). MI deals with the problem by regressing the missing data on the observed data. Thus, in longitudinal research MI can be used to predict missing data in one wave on the basis of available data from earlier or later waves. MI also requires a weaker assumption regarding the missing data mechanism (i.e., it results in unbiased estimates when the data are at least MAR\(^5\)) than naïve methods like CD, that assume MCAR (see, e.g., Smits, 2003). Moreover, contrary to ad hoc procedures such as CD, MI tends to reduce bias, even when the assumption of MAR is unrealistic (Schafer & Olsen, 1998). Finally, MI accounts for the uncertainty that is a result of the occurrence of missing values (Schafer & Graham, 2002).

In this study, the sequence of the three MBI subscales is examined using longitudinal data gathered among a representative group of Dutch dentists. Using SEM, several proposed models of sequence are compared. On the basis of results found earlier (e.g., Brouwers, 2000; De Vries, 2001), the original model proposed by Maslach and Jackson (1981) is expected to be the best fitting. Furthermore, in incorporating a technique of handling missing values, MI, this study aims to illustrate the preference that should be given to such methods over the widely used CD methods.

**Material and methods**

**Participants and procedure**

Participants in this study were derived from the pool of dental practitioners registered in the files of Movir Insurance (which does include more than 77% of all active Dutch dentists). A random selection procedure was employed, using stratification for gender, region (twelve provinces) and age. At Wave 1, a total of 950 dentists was sent a questionnaire. At Wave 2 the same group of dentists was approached, excluding 65 dentists that had explicitly indicated to prefer not to participate any longer. Data-

---

\(^5\) Rubin (1976) introduced a classification of the mechanisms that cause missing values. When the observed data are a completely random subsample of the hypothetically complete data, the data are Missing Completely At Random (MCAR). The second type of missing data mechanism is called Missing At Random (MAR), in which the missingness does not depend on the missing values but possibly on observed data. In the worst case, the missing data are dependent upon the values of the data that are missing; these data are described as Missing Not At Random (MNAR).
collection took place in March and April 1997 (Wave 1) and between April and June 2000 (Wave 2). The procedure included an announcement, two reminders, and if necessary a complete re-sending of the questionnaire (following recommendations by Dillman, 1978).

**Material**

A Dutch version of the MBI was used to measureburnout (Schaufeli & Van Dierendonck, 2000). Like the MBI-human services survey (MBI-HSS - Maslach et al., 1996), this instrument consists of the three subscales: EE, D and PA. EE consists of eight items (e.g., 'I feel emotionally drained from my work'), D consists of five items (e.g., 'I don't really care what happens to some patients') and PA consists of seven items (e.g., 'I deal very effectively with the problems of my patients'). Each of the twenty items can be answered on a 7-point Likert scale, ranging from 0 ('never') to 6 ('every day'). With the publication of a new manual, this Dutch questionnaire was renamed Utrechtse Burnout Schaal (UBOS - Schaufeli & Van Dierendonck, 2000).

**Overview: The analyses of missing data**

In longitudinal studies participants may be present in some waves of data collection and missing in others. This kind of missingness may be called wave nonresponse (Schafer & Graham, 2002). To adequately represent the uncertainty of its values, missing data should be imputed (filled in) more than once. Therefore, before an imputed value can be used as observed value, this uncertainty should be accounted for. Using MI, a missing value is replaced by m>1 possible values drawn from a distribution of plausible values; the variability among the values provides a measure of the uncertainty with which the imputed values are derived from the observed ones (Schafer & Olsen, 1998). Application of MI requires three steps as illustrated in Figure 1: imputation, statistical analysis and pooling. The implementation of these three steps for the current study is discussed below. For a thorough introduction the reader is referred to Schafer and Graham (2002).
Imputation of missing values

A procedure called Data Augmentation (DA) was used to create 5 completed data sets. Using the software program NORM (Schafer, 1999; Schafer & Olsen, 1998), the imputation model included the (incomplete) item scores of the burnout scale from Wave 1 and Wave 2, and the completely observed variables gender and age. These last two variables were added to preserve the structure of the data.

Statistical analysis

Psychometric aspects of the MBI subscales are examined by calculating interscale correlations and internal consistencies (Cronbach’s alpha). To determine the sequence of the three burnout components, thereby answering the main research question, several two-wave longitudinal models were compared in terms of model fit using the program LISREL 8.54 (Jöreskog & Sörbom, 1996)⁶. Analyses consisted of three distinguishable steps. First, a stability model (M₀), shown in figure 2, was tested in which the three subscales were allowed to correlate with each other within each wave

⁶ Within SEM missing values can also be dealt with using Full Information Maximum Likelihood (FIML, see, e.g., Enders, 2001). In many situations, the performance of MI and FIML will be essentially identical (Schafer, 2001). Here, MI was preferred because of the straightforward use of auxiliary variables for the estimation of the missing values and its ability to study rates of missing information (Collins, Schafer & Kam, 2001; Schafer & Graham, 2002).
and in which the factors at Wave 2 were regressed on themselves at Wave 1. The purpose of the second step was to attain a best fitting model. To this end, a stepwise approach was followed, which started with a saturated model consisting of all possible longitudinal relations. In this saturated model Wald tests were used to identify the longitudinal relations that did not significantly differ from zero. In several consecutive steps, these non-significant regression weights were fixed to zero (thereby effectively removing this relation), until there remained only significant regression weights. This procedure results in a (parsimonious) model that, purely on an empirical basis, can be considered best fitting to the data ($M_1$). In a third and final step, four models described in literature were fitted. These models were: the model describing the sequence EE$\rightarrow$D$\rightarrow$PA ($M_2$) proposed by Maslach and Jackson (1981); the model suggesting the sequence D$\rightarrow$PA$\rightarrow$EE ($M_3$) proposed by Golembiewski and Munzenrider (1988); the model containing the sequence PA$\rightarrow$D$\rightarrow$EE ($M_4$) proposed by Van Dierendonck et al. (2001); and a model with the sequence PA$\rightarrow$EE$\rightarrow$D ($M_5$), proposed by De Vries (2001). Because the same items were administered twice, the residuals of the paired items are allowed to covariate over time in all models.

**Figure 2: Two-Wave Longitudinal Design, Stability Model.**

- EE = emotional exhaustion; D = depersonalisation; PA = personal accomplishment.
The goodness-of-fit of models was assessed by several widely used absolute and relative indices. The absolute fit indices were the chi-square goodness-of-fit index ($\chi^2$) with the accompanying degrees of freedom and the Root Mean Square Error of Approximation (RMSEA). A lower value of chi-square indicates a better fit of the model. The value of the RMSEA should approach zero, whereby values smaller than 0.06 are considered indicative of acceptable model fit (Hu & Bentler, 1999). The fit indices Non-Normed Fit Index (NNFI) and the Comparative Fit Index (CFI) were included to test the relative fit of models. Values for the NNFI and CFI between 0.95 and 1.00 can be seen as indicative of a good fit and values of 0.90 or higher are indicative of an acceptable fit (Bentler, 1990). Furthermore, the fit of models was assessed by means of Akaike's Information Criterion (AIC) and the Consistent Akaike's Information Criterion (CAIC) (cf., Jöreskog & Sörbom, 2002). The relative fit of models that differ in restrictiveness can be assessed with these indices, with lower values of AIC and CAIC indicating better fit. Strictly speaking, the data are not normally distributed due to the use of Likert-scales. Nevertheless, the use of maximum likelihood estimation is warranted in light of its robustness to such a form of non-normality (cf., Dolan, 1994).

The assumption of measurement invariance is an essential aspect of longitudinal research (Meredith & Horn, 2001). In the comparison of measurements in both waves, it is important to establish that the items measure the same construct at Wave 1 and Wave 2, and that differences in item scores over time can thus be attributed to differences at the level of the constructs (i.e., the factors in the factor model). Measurement invariance can be investigated by testing for strong factorial invariance over time, which involves restricting measurement parameters to be equal over time. Meredith (1993) has shown that if factor loadings and measurement intercepts are time-invariant, measurement invariance holds (cf., Meredith & Horn, 2001). Note that, when testing for strong factorial invariance over time, factor (co)variances and factor means are allowed to differ over time.

Pooling

After creating 5 imputed versions of the data sets, and analysis of each separate data set, the pooling step consists of simply computing the average of the estimates in each
data file (Rubin, 1987). The uncertainty (variance) associated with the parameter in question has two components. The average within-imputation variance is the average of the parameter variances in the multiple data sets. The between-imputation variance is the variance of the parameter estimate over the multiple data sets. To assess the uncertainty due to the occurrence of missing values, the rate of missing information was studied. This estimate measures the increase in sample variance of a parameter due to missing values; it is determined by the rate and pattern of missing values and the ability of observed values to predict missing values successfully. It may be greater or smaller than the rate of missing values in any given problem (Schafer & Graham, 2002). For example, for data containing highly correlated variables, the missing information is expected to be lower than the actual rate of missing values and. For panel data suffering from definite dropout, missing information associated with variables in most recent waves is expected to be lower than the total rate of missing values. In practice, however, rates of missing values are often used as a base rate for the fractions of missing information (Schafer, 1997, p. 129). An MI-inference of all estimated parameters would go beyond the scope of this study. Therefore, to illustrate the impact of the occurrence of nonresponse on our outcomes, only the results of the MI-inference for the $\chi^2$ fit statistics of the theoretical models compared in this study are provided\(^7\). As advised by Schafer (1999), one imputed data file was randomly selected to form the empirically based ‘best-fitting’ model ($M_1$, described above). The resulting model was subsequently fitted on the four remaining imputed data sets.

**Results**

**RESPONSE**

Of the 950 dentists that had received a questionnaire at Wave 1, 735 responded (response rate 77%). At Wave 2, 22 of the 885 questionnaires that were sent proved undeliverable; 493 usable questionnaires were returned (57%). Each of these samples could be considered to be representative for the general population of dental

---

\(^7\) Results of the MI-inferences of all parameters are available when requested.
practitioners in The Netherlands, as is detailed elsewhere (see Gorter, Albrecht, Hoogstraten & Eijkman, 1999c; Te Brake, Bloemendal & Hoogstraten, 2003 for the 1997 and 2000 studies, respectively). At each of the studies, a number of dentists had removed their personal identification number, thereby assuring themselves of anonymity. However, in merging the datasets of the two studies, these respondents could not be included. As a result, a total of 30 respondents from the 1997 study (4.3%) and 18 respondents from the 2000 study (3.7%) where excluded. Burnout levels (mean EE, D and PA scores) within these excluded subgroups proved not to differ from the total means within studies, which makes the assumption tenable that deletion of these subjects will not influence the current analyses.

Within any two-wave longitudinal study, three types of ‘wave nonresponse’ (Schafer & Graham, 2002) can be distinguished: (1) subjects that responded at Wave 1, but not at Wave 2 (N=251); (2) subjects responding at Wave 2, but not at Wave 1 (N=67); and (3) subjects that responded at both measuring points (N=408). When CD methods are used, only respondents from the latter group would be included in analyses. The (incomplete) results of a total of 318 respondents would then be ignored whereas using MI, all 726 respondents are included in the analyses. MI reconstructs the missing values on the basis of observed values⁸. In Wave 1 9% (100×67/726) of the data was missing while in Wave 2 35% (100×251/726) of the respondents were missing. The overall rate of missing values in the data file was 22% (0.5×9%+0.5×35%).

MULTIPLE IMPUTATION

The EM algorithm converged in 42 iterations. It appeared that 42 cycles of DA sufficed to converge in distribution (Schafer, 1998). For an extra margin of safety, it was decided to carry out DA for 1000 cycles and generating imputations at every 200th cycle. All the estimates of means and (co)variances of the data showed good convergence behaviour.

⁸ Naturally, dentists who did not respond in either wave were omitted from the analysis because they contributed no information for the statistical inference (e.g., Schafer, 1997).
PSYCHOMETRIC RESULTS

Table 1 shows the interscale correlations for both measurement points for the imputed dataset. Interscale correlations show the same pattern as reported in the UBOS manual (Schaufeli & Van Dierendonck, 2000), although the high EE-D correlation both at Wave 1 and Wave 2 are higher (manual norm for EE-D is 0.49). This supports findings found earlier among dentists of relatively high correlations between EE and D (Gorter et al., 1999a). The internal consistencies, also reported in table 1, are consistent with figures reported in the UBOS manual.


<table>
<thead>
<tr>
<th>MBI subscale</th>
<th>Wave 1</th>
<th>Wave 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EE1</td>
<td>D1</td>
</tr>
<tr>
<td>Emotional exhaustion</td>
<td>0.89</td>
<td>0.60</td>
</tr>
<tr>
<td>Depersonalisation</td>
<td>0.70</td>
<td>0.59</td>
</tr>
<tr>
<td>Personal accomplishment</td>
<td>-0.20</td>
<td>-0.22</td>
</tr>
</tbody>
</table>

Note: All correlations are significant at the 0.01 level (2-tailed).

MISSING INFORMATION OF SEM MODELS

Table 2 reports the estimate of the chi-square fit statistic of the six structural models resulting from the matching MI-inferences. The fit statistics of the six models had percentages of missing information ranging from 37% to 41% with a mean of 39%. Consequently, these estimates are rather closer to the percentage of missing values in Wave 2 (35%) than to the overall percentage of missing values (22%). The disproportionate distribution of missing values over the two waves probably accounts for this. Many parameters of the estimated structural models are functions of variables in both waves. Therefore, it is very likely that these estimates (and consequently the fit statistics) suffered proportionally to the higher rate of missing values of the second wave. Even though the percentage of missing information is somewhat higher than expected, a value of 40% is generally considered to be a moderate rate (e.g., Schafer, 1997, p.137).
TABLE 2. TWO-WAVE LONGITUDINAL MODELS

<table>
<thead>
<tr>
<th>Model</th>
<th>$\chi^2$</th>
<th>df</th>
<th>RMSEA</th>
<th>NNFI</th>
<th>CFI</th>
<th>AIC</th>
<th>CAIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_0$</td>
<td>3040.30</td>
<td>739</td>
<td>0.0663</td>
<td>0.9482</td>
<td>0.951</td>
<td>3334</td>
<td>4010</td>
</tr>
<tr>
<td>$M_1$</td>
<td>3049.28</td>
<td>743</td>
<td>0.0662</td>
<td>0.9484</td>
<td>0.951</td>
<td>3341</td>
<td>3995</td>
</tr>
<tr>
<td>$M_2$</td>
<td>3051.98</td>
<td>743</td>
<td>0.0663</td>
<td>0.9483</td>
<td>0.951</td>
<td>3348</td>
<td>4001</td>
</tr>
<tr>
<td>$M_3$</td>
<td>3057.28</td>
<td>743</td>
<td>0.0664</td>
<td>0.9482</td>
<td>0.953</td>
<td>3350</td>
<td>4004</td>
</tr>
<tr>
<td>$M_4$</td>
<td>3057.10</td>
<td>743</td>
<td>0.0663</td>
<td>0.9482</td>
<td>0.951</td>
<td>3350</td>
<td>4003</td>
</tr>
<tr>
<td>$M_5$</td>
<td>3054.12</td>
<td>743</td>
<td>0.0664</td>
<td>0.9483</td>
<td>0.951</td>
<td>3353</td>
<td>4006</td>
</tr>
</tbody>
</table>

Note: $\chi^2$ = Chi-square goodness-of-fit index, df = degrees of freedom, RMSEA = Root Mean Square Error of Approximation, NNFI = Non-Normed Fit Index, CFI = Comparative Fit Index, AIC = Akaike's Information Criterion, CAIC = Consistent Akaike's Information Criterion. $M_0$ = stability model (see Figure 2), $M_1$ = EE $\rightarrow$ D & EE $\rightarrow$ PA, $M_2$ = EE $\rightarrow$ D $\rightarrow$ PA (Maslach & Jackson, 1981), $M_3$ = D $\rightarrow$ PA $\rightarrow$ EE (Golembiewski & Munzenrider, 1988), $M_4$ = PA $\rightarrow$ D $\rightarrow$ EE (Van Dierendonck et al., 2001), $M_5$ = PA $\rightarrow$ EE $\rightarrow$ D (De Vries, 2001).

FACTORIAL INvariance OVER TIME

To assess the assumption of measurement invariance, the tenability of strong factorial invariance was examined. Restricting factor loadings to be time-invariant, resulted in a significant increase in chi-square ($\Delta \chi^2(17)=50.0$). However, this restriction did not lead to a noticeable deterioration in any of the other fit measures. For instance, the RMSEA remained 0.065 over all models, while the NNFI showed a slight improvement from 0.948 in the model with no restrictions to 0.949 in the model with time-invariant factor loadings. Therefore, this restriction on factor loadings appears tenable. Restricting the item intercepts (and freeing the factor means on the second occasion) also resulted in a significant increase in chi-square ($\Delta \chi^2(17)=93.4$). But, again, the other fit indices indicated that this restriction was tenable. For instance, the NNFI of the strong factorial invariance model was identical to the NNFI value model without time-restrictions (0.948). The CFI showed a small drop (from 0.953 to 0.951), whereas the CAIC indicated that the restrictions for strong factorial invariance were tenable (CAIC changes from 4057 to 3968). Because restrictions on measurement parameters over time did not result in a clear deterioration in model fit, it was concluded that strong factorial invariance was tenable over time. Therefore, the model with time-invariant intercepts and time-invariant factor-loadings was used in all remaining analyses.
STRUCTURAL CONFIRMATIVE ANALYSIS

The iterative procedure used to exploratively produce a best fitting model \( (M_1) \) resulted in a model in which EE precedes both D and PA directly. Thus, high levels of EE lead to higher levels of D and PA independently. The goodness-of-fit indices for all models are presented in table 2. Although the RMSEA is somewhat higher than the rule-of-thumb value of 0.06, the CFI and NNFI indicate that all models show sufficient fit. The parameter estimates of the beta-weights (for the saturated model) are given in table 3. As can be seen by the Wald tests, three paths are not significantly different from zero (i.e., absolute Z-value smaller than 1.96, \( p>0.05 \)): PA\( \rightarrow \)EE, D\( \rightarrow \)EE, and D\( \rightarrow \)PA. This suggests that these paths do not add to the prediction of the components of burnout in 2000.

**Table 3. Parameter Estimates of the Beta-Weights for the Saturated Model (M_0)**

<table>
<thead>
<tr>
<th></th>
<th>EE 1997</th>
<th>D 1997</th>
<th>PA 1997</th>
</tr>
</thead>
<tbody>
<tr>
<td>EE 2000</td>
<td>beta estimate</td>
<td>0.832</td>
<td>-0.212</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.056</td>
<td>0.146</td>
</tr>
<tr>
<td></td>
<td>Z-value</td>
<td>14.870**</td>
<td>-1.448</td>
</tr>
<tr>
<td>D 2000</td>
<td>beta estimate</td>
<td>0.115</td>
<td>0.478</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.033</td>
<td>0.086</td>
</tr>
<tr>
<td></td>
<td>Z-value</td>
<td>3.448**</td>
<td>5.563**</td>
</tr>
<tr>
<td>PA 2000</td>
<td>beta estimate</td>
<td>-0.065</td>
<td>0.077</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.032</td>
<td>0.082</td>
</tr>
<tr>
<td></td>
<td>Z-value</td>
<td>-2.028*</td>
<td>0.915</td>
</tr>
</tbody>
</table>

Note: EE = emotional exhaustion; D = depersonalisation; PA = personal accomplishment. *\( p<0.05 \); **\( p<0.01 \)

Notwithstanding the evident fit of all models, any clear conclusions concerning the best-fitting model are precluded because of the minute differences of the fit indices between models. An explanation for these small differences can be found in the large number of parameters included in the complete model. Since the measurement invariant longitudinal factor model fits the data quite well, it is warranted to reduce the number of parameters in the compared models by using the scale scores instead of the raw item scores.

The results of the path-analysis using scale scores are reported in table 4. In terms of RMSEA the only model that comes close to the acceptable fit (RMSEA<0.06) is model \( M_1 \). Nevertheless, for all models, the relative fit indices NNFI and CFI indicate acceptable (i.e., >0.90) to good (>0.95) fit. What is more, fit
indices are far less homogeneous between the models, which allows for a distinction between models. Not surprisingly, model $M_1$ shows the overall best fit. Of the remaining models, however, models $M_2$ and $M_5$ stand out in having the relative best fit, whilst showing a good fit according to the NNFI and CFI measures. Of these two models, $M_5$ should be preferred; for this model, all fit indices indicate the best fit.

**Table 4. Two-wave longitudinal models, using scale means for analyses**

<table>
<thead>
<tr>
<th>Model</th>
<th>$\chi^2$</th>
<th>df</th>
<th>RMSEA</th>
<th>NNFI</th>
<th>CFI</th>
<th>AIC</th>
<th>CAIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_1$</td>
<td>18.57</td>
<td>4</td>
<td>0.0689</td>
<td>0.9742</td>
<td>0.9928</td>
<td>53</td>
<td>147</td>
</tr>
<tr>
<td>$M_2$</td>
<td>26.42</td>
<td>4</td>
<td>0.0870</td>
<td>0.9596</td>
<td>0.9894</td>
<td>60</td>
<td>155</td>
</tr>
<tr>
<td>$M_3$</td>
<td>57.41</td>
<td>4</td>
<td>0.1334</td>
<td>0.9206</td>
<td>0.9742</td>
<td>90</td>
<td>185</td>
</tr>
<tr>
<td>$M_4$</td>
<td>38.58</td>
<td>4</td>
<td>0.1062</td>
<td>0.9382</td>
<td>0.9836</td>
<td>72</td>
<td>167</td>
</tr>
<tr>
<td>$M_5$</td>
<td>24.11</td>
<td>4</td>
<td>0.0824</td>
<td>0.9640</td>
<td>0.9904</td>
<td>58</td>
<td>153</td>
</tr>
<tr>
<td>$M_{1,\text{REV}}$</td>
<td>2.73</td>
<td>3</td>
<td>0.0082</td>
<td>1.001</td>
<td>0.9998</td>
<td>39</td>
<td>139</td>
</tr>
</tbody>
</table>

Note: Scale mean scores for EE, D and PA were used in analyses. $\chi^2$ = Chi-square goodness-of-fit index, df = degrees of freedom, RMSEA = Root Mean Square Error of Approximation, NNFI = Non-Normed Fit Index, CFI = Comparative Fit Index. AIC = Akaike's Information Criterion, CAIC = Consistent Akaike's Information Criterion. $M_0$ = stability model (see Figure 2), $M_1$ = EE$\rightarrow$D & EE$\rightarrow$PA, $M_2$ = EE$\rightarrow$D$\rightarrow$PA (Maslach & Jackson, 1981), $M_3$ = D$\rightarrow$PA$\rightarrow$EE (Golembiewski & Munzenrider, 1988), $M_4$ = PA$\rightarrow$D$\rightarrow$EE (Van Dierendonck et al., 2001), $M_5$ = PA$\rightarrow$EE$\rightarrow$D (De Vries, 2001). $M_{1,\text{REV}}$ = alternative explorative model, based on scale means: EE$\rightarrow$D & EE$\rightarrow$PA$\rightarrow$D.

As a final explorative exercise, the iterative procedure used to produce $M_1$ was repeated using the scale means as input values. The procedure resulted in a revised version of $M_1$, which combined EE$\rightarrow$D with EE$\rightarrow$PA$\rightarrow$D (effectively adding the PA$\rightarrow$D path to the original $M_1$). The model fit indices for this model ($M_{1,\text{REV}}$) are also included in table 4. Results on $M_{1,\text{REV}}$ not only show an extremely good absolute fit, but also an exceptionally better fit to the data relative to the other models.

**Discussion**

The first purpose of this study was to examine the sequence of the three MBI subscales: EE, D and PA in a longitudinal setting. Two methods were used to find the ‘best sequence’. Firstly, a number of models specifically named in literature were compared, including a model that can only be tested in a longitudinal setting the model proposed by Van Dierendonck et al. (2001). Secondly, a completely
empirically based, best-fitting model was examined. An additional purpose of this study was to describe a more elaborate handling of missing values than is commonly used in longitudinal research. Results indicate that the models proposed by Golembiewski and Munzenrider (1988) and Van Dierendonck et al. (2001) do not adequately fit the data. In contrast, both the model proposed by Maslach and Jackson (1981), as well as an alternative found among physiotherapists (De Vries, 2001) were found to have a good fit on the data. Furthermore, the best-fitting models indicated that among dentists, EE is most appropriate aspect to be considered an early sign of burnout.

Of the two models that showed a good fit, current results indicate a slight preference for the PA→EE→D model. De Vries (2001), when originally proposing this model, did not explicitly name any theoretical or empirical reasons for including this model, other than it being a variation of the Van Dierendonck et al. (2001) model, which also places PA as the first aspect to be affected in the process of burnout. Regardless of its exact origin, it is interesting to see that this model, originally found among physiotherapists, is also found to have the best fit among dentists. In a number of work-related aspects, dental practitioners can be considered comparable to physiotherapists. Within general health care, both professions can be considered extraordinary because of their entrepreneur-like position. Also, both dentists and physiotherapists have to deal with patients on a daily basis, often within a relatively confined working environment. Finally, both rely to a large degree on manual abilities in daily work. It is therefore of interest to find that in both professions evidence was found for PA as a firstly affected burnout component.

PA has been compared to feelings of competence, mastery and goal orientation: “Personal Accomplishment is defined as the evaluation of the relational skills in handling recipients, which may influence self-efficacy beliefs regarding future performance” (Van Dierendonck et al., 2001, p. 49). It is imaginable that such (general) beliefs translate to (more specific) feelings about one's manual competence. However, given the central position of manual abilities in dental education it seems more realistic to assume that these manual competences are a relatively stable aspect within dental work, while, indeed, the relational aspects are much more susceptible to
outside influences. Moreover, the items on the PA scale do not relate to any specific manual skills (which are very important in daily work for dentists and physiotherapists alike). In contrast, the PA items have much more to do with social and emotional skills (De Vries, 2001), aspects for which the average dentist is often less adequately trained. In this line of thought, it is to be assumed that PA is influenced in the early stages of the onset of burnout.

Although intuitively intriguing, this interpretation is not supported by the empirically found ‘best fitting’ models, which indicate that EE not only leads to an increase in D, but also directly leads to a decrease in PA (as opposed to the ‘route’ via D), thereby indicating a relation between the two. This contradicts suggestions made elsewhere, where it was suggested that PA has a relatively independent role compared to EE and D (e.g., Lee & Ashforth, 1996; Leiter, 1993; Maslach et al., 2001). A possible explanation for this unusual outcome may be found in the fact that only causal relations are tested. Synchronous development of EE and D are not explicitly tested but such a (mutual) development could be at work. Such inference is supported by the relatively high correlation between EE and D that was found among dentists in this study (see table 1) as well as in earlier research (Gorter et al., 1999a). Another explanation can perhaps be found in the time-window of 3 years that was used in this study. It should be stressed that the time between measurement points probably can be of great importance for the process that is at work. The hypothesis that different processes are working in the short term and in the long term cannot be discarded. In the same vein, the current study made use of two measuring points. In future research, three measuring point would be advisable for the fitting of these models. Furthermore, no so-called third variables were included in the design. Inclusion of such variables would generally strengthen the causal inference (Zapf et al., 1996), but they could also shed more light on the relation between EE and PA.

Of the theoretically proposed models, the results point to, but do not clearly differentiate between, the EE→D→PA model by Maslach et al. (1996) and the PA→EE→D model by De Vries (2001). Nevertheless, some important conclusions can be drawn that should be considered relatively exclusive for the setting of health care entrepreneurs (e.g., dentists and physiotherapists). Both early signs of exhaustion, as
well as feelings of reduced personal accomplishment should be taken as early warning signs of burnout. However, findings on both best-fitting models indicate exhaustion to be the first aspect of the burnout process, and that exhaustion, in turn, is followed by reduced feelings of personal accomplishment and, at the same time, an attitudinal response in the form of depersonalisation. Thus on these results, it seems that exhaustion is most likely to be the first effect of a pending burnout. As EE is traditionally seen as an orthodox response to actual work related stress, specific stress related training programs are the most appropriate aspect to be addressed with professional entrepreneurs in health care.

The second purpose of this study beholds an additional practical implication of this study, and lies in the methodological issues concerning longitudinal research. For one, the (often implicitly made) implicit assumption of measurement invariance was explicitly tested. It was found that the MBI was indeed measurement invariant over time. This suggests that the items of the MBI measure the same attributes in an equivalent manner in the two waves of this study. This clearly adds to the validity of the MBI. For instance, measurement invariance over time of the MBI suggests the absence of any re-test effects. Secondly, even though nonresponse is omnipresent, many recent longitudinal studies on burnout and work-related stress do not provide information on the exact handling of missing values (e.g., Burke & Greenglass, 1995; Deary, Watson & Hogston, 2003; Savicki & Cooley, 1994; Van der Ploeg & Kleber, 2003). Others resort to CD (e.g., Bakker, Schaufeli, Sixma, Bosveld & Van Dierendonck, 2000b; Houkes, Janssen, De Jonge & Bakker, 2003), probably because CD is the default method for dealing with incomplete data in many statistical software packages. However, CD may only yield unbiased estimates when the missing data comply with the strict assumption of MCAR (e.g., Smits, Mellenbergh & Vorst, 2002), and even when MCAR holds, CD should still be considered inefficient (Schafer & Graham, 2002). Moreover, CD does not properly account for the uncertainty that is associated with the occurrence of missing data (e.g., Little & Schenker, 1995).

The application of the MI-inference ensured better parameter estimates than when applying naïve methods like CD. In an extra analysis, to specifically test the
assumption of MCAR, Little’s test for MCAR (Little, 1988) showed that the data where not MCAR ($\chi^2(950)=1054.73$, $p=0.010$). Thus, in the current research, the assumption of CD would be violated and would have led to biased estimates. For future longitudinal research it is recommended also to check for MCAR; when this assumption is not met, the use of CD should be avoided. Furthermore, the MI-inference gave insight in the effect of the occurrence of missing values on the uncertainty associated with the estimated models: 40% of the variance of the fit statistics was a result of the nonresponse. This percentage was closer to the percentage of missing values in the second wave than the overall percentage of missing values. This outcome clearly signifies that when studying the phases of burnout (i.e., modelling processes through time) not only the use of a longitudinal design should be stressed, but also every effort should be taken have equal rates of (non)response in each wave.