Training Evaluation With 360-Degree Feedback

Froukje A. Jellema
University of Twente, The Netherlands

Training evaluation is an important issue in many organizations. Even though organizations are becoming more interested to measure the effects of their training investments, organizations often lack the tools for the measurement of training effects, especially at the level of changed behavior. This paper will focus on the use of 360-degree feedback for evaluation of changed behavior as a result of training. The results of an experimental study in a large Dutch hospital will be described and discussed.

Keywords: Training Evaluation, 360-degree Feedback, Behavioral Change

Human resource development (HRD) is a process of developing and unleashing human expertise through organization development and personnel training and development, for the purpose of improving performance (Swanson, 2000 AHRD). Training evaluation can be useful to find out whether investments in training are contributing to performance improvement. One evaluation focus is changed behavior as a result of training (Kirkpatrick, 1994). Though change in behavior is in most cases not an end in itself, it is a necessary intermediary step to evaluate whether the training contributes to performance improvement. That is why in this paper the focus is on the evaluation of behavioral change.

Evaluation of changed behavior is done for only 11 percent of the training programs, though companies expect to increase this proportion in the future (1999 ASTD report). A main reason why training evaluation is not often focused on behavioral change of trainees, is the lack of appropriate methods. The specific aim of this paper is to examine one particular method that may be useful in this context. This method is 360-degree feedback, where employees receive feedback about behavior from a 'full circle' of co-workers such as the supervisor, peers, subordinates and clients. 360-degree feedback may be used to determine whether co-workers have experienced a difference in the behavior of the trainee. The main research question of this paper is whether 360-degree feedback is indeed useful for training evaluation.

Theoretical Framework

Though the use of 360-degree feedback for training evaluation is often mentioned in evaluation handbooks (see for example Brown and Seidner, 1998; Kirkpatrick, 1994; Robinson and Robinson, 1989), this method has rarely been studied in an evaluation context, though often in an employee developmental context.

One study in which 360-degree feedback is used to evaluate a training program, is the study by *McLean, Sytsma, and Kerwin-Ryberg (1995)*. This study included 73 ratees, who received 360-degree feedback both before and two years after a training program. Though participants' skills increased during these two years, it was difficult to attribute this behavioral change to the training program, due to the long period between pre- and posttest and the fact that a control group was not included. The authors
concluded that 360-degree feedback should be applied very cautiously as a tool to evaluate training.

Another study, in which 360-degree feedback is used for training evaluation, and in which a design with a control group was used, is the study of *Van Sandick and Schaap-Neuteboom (1993)*. This study was situated in a large Dutch supermarket chain and focused on a training program for department-managers. This study included 162 ratees, distributed over two training groups and one control group. In this study a training effect was measured (i.e. 0.2 on a 5-point scale). However, this study was limited by the fact that only one supervisor and one peer were involved in the feedback.

Another example is the study of *Rosti and Shipper (1998)*, where 360-degree feedback is used to study the impact of training in a management development program based on 360-degree feedback. Feedback was collected both before and after the program to enhance learning and evaluate training. Results, when analyzed by means of a series of four matched-pairs tests, were supportive that the change in the experimental group was significantly different than the change for the control group, and in the expected direction.

Since organizations are becoming more interested in measuring behavioral change as a result of training, it is interesting to study the usefulness of 360-degree feedback for this purpose. Since previous studies in this area are rare and often limited, it was decided to use as strict an experimental design as possible to again look at the usefulness of 360-degree feedback for training evaluation.

Research Questions

The main research question of this paper is whether 360-degree feedback can be useful for training evaluation. More specifically, the questions are:

1. Is training evaluation by means of 360-degree feedback reliable?
2. Is training evaluation by means of 360-degree feedback valid?
3. What training effects are found with 360-degree feedback?

Methodology and/or Research Design with Limitations

The methodology that is used to answer the questions, is a quasi-experimental study. This study is situated in a large Dutch hospital. Recently, this hospital has created a new professional job for nurses, the senior nurse, hierarchically situated below the head nurse. All senior nurses have to attend the course, that has been specifically developed for this function. This is an extensive course, consisting of 18 outdoor training days during a nine-month period. Coaching skills are an important part of the program.

At the time of this study, about 40 senior nurses were working at the hospital. Though all senior nurses have to participate in the program, only 14 nurses could attend at the same time. That is why it was possible to study three conditions. Group 1 had already started with the program at the beginning of this study and received only a 360-degree feedback post-test. Group 2 received a pretest with 360-degree feedback before attending the program, and afterwards a posttest with 360-degree feedback. Group 3 received the feedback at the same time as group 2, but did not attend the program. All post-tests took place after ten training days, about seven months after the pretest. It would have been preferred if the post-measurement had taken place at the close of the training program, but that excluded the use of a control group, since group 3 would have started with the training by then.

The Dutch 360-degree feedback instrument Reflector was used in this study, consisting of eight behavioral competencies: dealing with stress, integrity, active listening, creating partnership, diagnostic skills, stimulation, giving feedback, and adapting. Each behavioral competency consists of five items, so the questionnaire included 40 items. The response scales that are used in Reflector are 5-point scales were raters select a position between opposite statements of behavior. The reliability and validity of Reflector has been examined in previous studies (see Jellema, Visscher and Mulder, 2000) and was found to be satisfactory.

Each senior nurse was allowed to select his or her own raters, but preferably their supervisor, two peers and four subordinates. Clients (in this case: patients) were not included. The ratees were encouraged to ask the same raters to provide feedback at the pre- and posttest, for as far as this was possible. All ratings could by given anonymously. After the results were analyzed, the results were summarized in reports and send to the home addresses of the ratees. All ratees were invited to join in a follow-up session.

The post-tests included a retrospective self-test and a questionnaire with questions regarding background, motivation for the training program, perceived support on-the-job, etc. Out of group 2 and 3, eight senior nurses were found willing to participate in a coaching simulation, that took place at the time of the post-test. Main purpose was to study correlation between 360-degree feedback (post-test) scores and scores of the simulation and thus to gain insight into the validity of 360-degree feedback. The research design is summarized in Table 1.

Table 1. *The research design of this study*

Group 1	training	360° & retrospective self & background	
Group 2	360°	training	360°& retrospective self & background & coaching simulation
Group 3	360°	no training	360°& retrospective self & background & coaching simulation

Response

At the start of the study each group consisted of 14 senior nurses. However, several senior nurses could not partake because they were ill (especially in group 3, the group that was on the waiting list for the program), left the organization or did not want to cooperate. At the posttest, again several nurses were ill and could not participate. Table 2 summarizes the response.

Table 2. *Response of the evaluation at the OLVG*

	Pre-test	Post-test		Post-test		
	360	360	Background & Retrospective	360	Background & Retrospective	Simulation
Group 1		n= 13 (95 raters)	n= 13			
Group 2	n=12 (86 raters)			n = 12 (78 raters)	n = 12	n=4
Group 3	n = 9 (69 raters)			n = 7 (48 raters)	n = 7	n=4

In this paper, the focus will be on group 2 and 3, since these groups received a pre- as well as a posttest. The pretest included 21 senior nurses, 12 in the experimental group and nine in the control group. A total of 155 raters were involved, an average of 7.4 raters for each ratee. Raters were divided as follows: self (21), supervisors (20), peers (62) and subordinates (52). At the posttest, a total of 18 senior nurses again participated in the 360-degree feedback. One senior (in group 2) participated in the posttest but had not participated in the pretest. A total of 126 raters were involved, an average of 6.6 raters for each ratee. Raters were divided as follows: self (19), supervisors (17), peers (50) and subordinates (40). About 35% of the seniors had asked exactly the same raters as at the pretest. In the other cases, one or more of the raters at the posttest had not been included in the pretest. All senior nurses filled in the questionnaire with background questions and the retrospective questionnaire.

Since this study was a quasi-experimental study, and the groups were not selected at random, they were compared on several variables to find out if they were different (see Table 3).

Table 3. *The two groups compared on several variables*

Variable	Group 2 (n=12)	Group 3 (n= 9)
Gender	53.8% female; 46.2% male	77.8% female; 22.2% male
Age	36 (4.6)	34 (8.1)
Working experience	12.2 (5.4)	12.7 (9.6)
Experience senior nurse	.8 (1.3)	2 (1.8)
Experience coaching	66.7%	88.9%
Perceived need to develop coaching skills	8.3% very high; 75% high	11.1% very high; 66.7% high
Motivation for program	25% very high; 66.7% high	22.2% very high; 66.7% high

As can be seen in Table 3, the groups are different on many of these variables. However, differences between groups were tested by means of chi-square (gender and experience with coaching), ANOVA (age and need to develop coaching skills) and Kruskal Wallis (experience and motivation), and none of the differences are significant. This is probably due to the limited number of respondents that was involved in this study.

This limited number of respondents and the fact that respondents were not selected at random for each training group, is the main limitation of this study. Interpretation of results may be difficult due to this limitation. In addition, the ratees that were included at the pretest are not always (though often) the same individuals as the ratees at the posttest. In addition to this, it is a limitation of this study that it was not possible to train the raters in advance.

Results

In this section the following results will be presented: the psychometric properties of the 360-degree feedback instrument and the measurement of behavioral change as a result of the training.

Psychometric Properties of the 360-degree Feedback Instrument

In this section the focus will be on the internal consistency and inter-rater reliability and validity of the instrument.

Internal Consistency. Internal consistency is measured by means of reliability analysis of the scales. The analysis is done on the complete data set (pre- and posttest, including group 1). Table 4 summarizes the alpha coefficients.

Table 4. Internal consistency of the scales

Competency	Alpha
Active listening	.60 (n=357), .68 after deleting item 2
Creating partnership	.66 (n=341)
Diagnostic skills	.76 (n=333)
Stimulation	.70 (n=336)
Giving feedback	.74 (n=353)
Adapting	.81 (n=326)
Integrity (control competency)	.72 (n=356)
Dealing with stress (control competency)	.78 (n=349)

Two scales are below .7, i.e. active listening and creating partnership. The alpha of active listening can be increased to .68 by removing one item. The alpha of creating partnership and the other behavioral competencies can not be increased by removing items. When all items are considered as one scale ('coaching skills'), the alpha is .94 (n=272)

Inter-rater Reliability

Since most of the senior nurses have received feedback from more than one peer and/or subordinate, it is interesting to see if scores within each rater-source are more or less the same. Since many ratees have feedback from three or more peers and/or subordinates, it was not possible to use a correlation coefficient. That is why the standard deviation is used as a measure for within-source agreement. The data was first analyzed separately for each ratee after which an average score for all ratees was computed. The standard deviation is computed for all competencies and for an overall score of all competencies. Table 5 shows the average standard deviation for peers and subordinates.

Table 5. *Within-source standard deviation for peers and subordinates*

Competency	Peers	Subordinates
Dealing with stress	0.48 (n=44)	0.41 (n=39)
Integrity	0.33 (n=44)	0.39 (n=38)
Active listening	0.44 (n=44)	0.43 (n=40)
Creating partnership	0.48 (n=44)	0.49 (n=39)
Stimulation	0.43 (n=44)	0.48 (n=38)
Diagnostic skills	0.52 (n=43)	0.49 (n=38)
Adapting	0.48 (n=44)	0.59 (n=39)
Giving feedback	0.43 (n=44)	0.44 (n=39)
Overall score	0.45 (n=44)	0.46 (n=40)

The average overall standard deviation for peers is .45. This differs slightly for each behavioral competency, with most agreement between peers on integrity and less agreement on diagnostic skills. The average overall standard deviation for subordinates is .46, which again differs slightly for each behavioral competency. Most agreement between subordinates is on integrity and less agreement on adapting.

Since the main characteristic of 360-degree feedback is that several different rater sources are included, it is interesting to study agreement between rater-sources. Whenever a ratee has two or more peers and/or subordinates, an average score for these raters is computed. This average peer-score and average subordinate-score is used in the analysis. The analysis is based on the complete data set (both pre- and posttest, including group 1). For each ratee, a score was computed for self, supervisor, peer and subordinate. These scores were averaged for all ratees. In Table 6, the average scores are summarized.

Table 6. *Average score 360-degree feedback*

Competency	Self (n=53)	SV (n=50)	P (n=52)	SO (n=50)
Dealing with stress	4.03 (.61)	4.00 (.75)	3.96 (.55)	4.10 (.52)
Integrity	4.51 (.38)	4.57 (.38)	4.48 (.33)	4.44 (.51)
Active listening	4.01 (.44)	3.87 (.60)	3.90 (.38)	3.99 (.48)
Creating partnership	3.96 (.42)	3.96 (.67)	3.92 (.43)	3.90 (.58)
Stimulation	3.96 (.46)	4.01 (.64)	3.97 (.40)	4.04 (.43)
Diagnostic skills	4.00 (.43)	4.00 (.57)	4.05 (.41)	4.08 (.52) (n=49)
Adapting	3.97 (.57) (n=52)	3.87 (.74)	3.94 (.50)	3.98 (.58)
Giving feedback	4.08 (.45)	4.06 (.60)	4.15 (.41)	4.22 (.44)
Overall score	4.06 (.35)	4.04 (.54)	4.05 (.33)	4.09 (.45)

SV=supervisor; P=peers; SO = subordinates

As can bee seen, the overall scores of all rater-sources are very similar. The standard deviation in Table 6 indicates the differences between ratees. When the assumption is that ratees have different levels of performance, this coefficient should be large; it is, however, rather low.
Correlation between these sources was studied by means of Ttests, since the scores appear to be normally distributed. In Table 7 the coefficients are summarized. The marked coefficients are significant (at .95 level).

Table 7. *Correlation between rater-sources*

Competency	S-SV (n=50)	S-P (n=52)	S-SO (n=50)	SV-P (n=49)	SV-SO (n=47)	P-SO (n=49)
Dealing with stress	0.39*	0.20	0.08	0.44*	0.30	0.31*
Integrity	0.11	-0.04	0.05	0.15	0.27	0.00
Active listening	0.27	-0.03	0.09	0.07	0.12	0.04
Creating partnership	0.22	0.20	-0.13	0.17	0.26	0.06
Stimulation	0.18	-0.02	0.15	0.12	0.30*	0.07
Diagnostic skills	0.27	0.05	0.02	0.23	0.24	-0.16
Adapting	0.35*	0.23	0.11	0.41*	0.29*	0.01
Giving feedback	0.32*	0.03	-0.02	0.28	0.16	-0.01
Overall score	0.34*	0.00	0.03	0.29*	0.32*	0.03

SV=supervisor; P=peers; SO = subordinates

The fact that correlation appears to be moderate for most competencies and rater-sources, implicates that scores should by presented for each rater-source separately. This is in accordance with what you might expect for 360-degree feedback. Correlation between peers-subordinates, peers-self and self-subordinates, is in some cases even negative. When the correlation of the overall score is used as a measure of rater-correlation, the correlation between supervisor-self, supervisor-peer, and supervisor-subordinates is significant. These correlation coefficients are in agreement with the low correlation that is often found in other studies (see for example Harris and Schaubroeck, 1988). However, correlation between self-peer, self-subordinate and peer-subordinate are much lower than have been found in previous studies. Interesting to note is that the correlation between the supervisor and any of the other rater sources is significant, while correlation between the other sources is not significant.

Validity

To study the validity, the 360-degree feedback scores were compared to another, more objective, measure of the same behavior. Eight senior nurses were found willing to participate in a coaching simulation that consisted of a coaching conversation that the senior nurse had with an employee, played by an actress. The simulation focused on some of the competencies that had been included in the 360-degree feedback, i.e. active listening, creating partnership, diagnostic skills and stimulation. The eight conversations were recorded on videotape and two masters students evaluated the videotapes by means of the same questionnaire that was used for the 360-degree feedback. The students were trained in advance (with an extra videotape) to use the instrument and assess the coaching behavior. Each student rated each ratee, after which they discussed the results and tried to come to agreement. Since each student-score correlated significantly with the agreement-score, this agreement-score is used in the analysis. Because the simulation was in the same period as the 360-degree feedback posttest, the posttest scores are used. Table 8 shows the correlation that was found.

Table 8. *Correlation between 360-degree feedback and coaching simulation*

	Sim-360 Self (n=8)	Sim-360 SV (n=8)	Sim-360 peers (n=8)	Sim-360 SO (n=8)
Active listening	-.16	-.58	.31	-.70
Creating partnership	.17	-.33	.25	-.31
Diagnostic skills	-.25	-.72*	.40	-.44
Stimulation	-.44	-.60	.29	-.58

SV=supervisor; SO = subordinates

As can be seen, the correlation between the 360-degree feedback and simulation is in most cases negative. Only peers have a positive (though not significant) correlation with the simulation-score.

Measuring Behavioral Change as a Result of Training

In this section the focus is on the gain scores between pre- and posttest. First, it is important to focus more specifically on the self-scores, since these may be sensitive to 'response-shift bias': because ratees in group 2

receive intensive training of coaching skills, their ideas about coaching may change. This may result in the fact that self-scores at the posttest are based on another perspective of coaching, which makes them difficult to compare with the pretest scores. Self-scores could be lower at the posttest, even though performance has improved, because trainees have learned how difficult coaching actually is and are more demanding of their own performance. The same process may happen for ratees in group 3, who received no training but who did receive a feedback-report after the pretest, joined in a follow-up meeting and were encouraged to make a personal development plan. These individuals may also have changed their ideas about coaching. That is why a retrospective self-test was included with the posttest, where ratees were asked to again rate their coaching skills at the time of the pretest. Correlation between pretest self-scores and retrospective scores was studied by means of a T-test (see Table 9).

Table 9. *Correlation between 360-degree pre self-score and retrospective self-score*

Competency	Pretest self-score	Retrospective self-score	Correlation
Dealing with stress (n=31)	4.01	3.81	.68*
Integrity (n=31)	4.46	4.41	.40*
Active listening (n=31)	3.93	3.55	.43*
Creating partnership (n=30)	3.89	3.65	.09
Stimulation (n=31)	3.92	3.55	.31
Diagnostic skills (n=31)	3.94	3.59	.42*
Adapting (n=30)	3.95	3.64	.53*
Giving feedback(n=31)	3.93	3.78	.43*

Though the retrospective self-scores are indeed lower than the pretest self-scores, this is for most competencies not significant. That is why the pretest scores will be used in the further analysis.

In Table 10, the pre- and posttest scores are summarized. Again, the score for peers and subordinates is based on the average score. A paired samples T-test analysis was executed, to see if differences between pretest and posttest are significant (the significant differences are again marked in Table 10). Whenever there is a (-) in the table, this indicates that the posttest scores are lower.

Table 10. *Pre- and posttest scores of group 2 and 3*

Group 2								
Competency	**Pretest**				**Posttest**			
	S	**SV**	**P**	**SO**	**S**	**SV**	**P**	**SO**
Dealing with stress	3.82	3.70	3.84	4.17	3.96	4.06	4.22	3.98 (-)
Integrity	4.42	4.26	4.46	4.40	4.48	4.66	4.52	4.34 (-)
Active listening	3.87	3.58	3.80	4.08	4.03	4.10	4.08*	4.03 (-)
Creating partnership	3.90	3.50	3.96	3.98	4.03	4.16	3.91 (-)	3.74 (-)
Stimulatio n	3.81	3.64	3.96	4.05	4.00	4.32*	4.09	3.97 (-)
Diagnostic skills	3.78	3.56	3.97	4.18	4.05	4.20	4.25*	3.98 (-)
Adapting	3.69	3.36	3.96	4.03	3.95	4.36*	3.94 (-)	3.80 (-)
Giving feedback	3.90	3.56	4.10	4.33	4.15*	4.32*	4.33	4.24 (-)
Overall score	**3.90**	**3.64**	**4.00**	**4.15**	**4.08**	**4.27***	**4.17**	**4.01 (-)**
Group 3								
Competency	**Pretest**				**Posttest**			
	S	**SV**	**P**	**SO**	**S**	**SV**	**P**	**SO**
Dealing with stress	4.09	4.10	3.83	4.04	4.26	4.63*	3.81 (-)	4.12
Integrity	4.34	4.43	4.33	4.48	4.80*	4.87	4.49	4.34 (-)
Active listening	3.97	3.53	3.76	3.94	4.26	4.33*	3.87	3.79 (-)
Creating partnership	3.72	3.63	3.71	4.06	4.14*	4.27*	3.89	3.89 (-)
Stimulation	3.74	3.43	3.86	4.15	4.00	4.50*	4.02	4.07 (-)
Diagnostic skills	3.90	3.70	3.92	4.17	4.17	4.47*	3.99	3.99 (-)
Adapting	3.81	3.10	3.73	4.22	4.14	4.17*	4.01	3.89 (-)
Giving feedback	3.83	3.67	3.89	4.31	4.46*	4.63*	4.17	4.00 (-)
Overall score	**3.93**	**3.70**	**3.88**	**4.17**	**4.28**	**4.48***	**4.03**	**4.01 (-)**

SV=supervisor; P=peers; SO = subordinates; (-) = the posttest lower than the pretest

As can be seen, both group 2 and 3 have higher posttest scores than pretest scores for most behavioral competencies. Supervisors of both groups, have often significantly higher scores at the posttest, as well as the ratees of group 3 themselves. Peers in group 2, are also significantly more positive at the posttest than at the pretest for two competencies (active listening and diagnostic skills). However, subordinates of both groups are more negative (though not significantly) at the posttest.

Next, group 2 and 3 were compared to see if the gain scores between these two groups are different from each other. Since group 2 is the training-group, it was expected that the gain scores of this group are higher than the scores of group 3. The results are summarized in Table 11 (marked coefficient is significant).

Table 11. Gain scores of group 2 and 3

Competency	Group 2				Group 3			
	S	SV	P	SO	S	SV	P	SO
Dealing with stress	0.14	0.36	0.38	-0.19	0.17	0.53	-0.02	0.09
Integrity	0.07	0.40	0.07	-0.06	0.46*	0.43	0.16	-0.14
Active listening	0.17	0.52	0.28	-0.05	0.29	0.80	0.11	-0.16
Creating partnership	0.13	0.66	-0.05	-0.23	0.42	0.63	0.18	-0.17
Stimulation	0.19	0.68	0.13	-0.07	0.26	1.07	0.16	-0.07
Diagnostic skills	0.27	0.64	0.28	-0.20	0.27	0.77	0.07	-0.18
Adapting	0.26	1.01	-0.02	-0.23	0.33	1.07	0.28	-0.32
Giving feedback	0.25	0.76	0.23	-0.08	0.63	0.97	0.28	-0.31
Overall score	0.18	0.63	0.17	-0.14	0.35	0.78	0.15	-0.16

SV=supervisor; P=peers; SO = subordinates

As can be seen in Table 11, most rater sources indicate positive gain scores, with an exception for subordinates and for peers (regarding some competencies). A first quick look seems to indicate that the gain scores in group 3 are, contrarily to what might be expected, higher than those of group 2. However, in only one case is this difference significant, and more positive, for group 3.

Discussion

This study focused on the use of 360-degree feedback in a training evaluation context. First the focus was on the psychometrical properties of 360-degree feedback. Results indicate that the psychometrical properties with regard to internal consistency of the scales and inter-rater reliability are more or less in accordance with previous research. Alphas of most scales are sufficient (.70 or higher). Within-source agreement, measured by means of the standard deviation, was found to be high (.45 for peers and .46 for subordinates). However, this may be caused by the fact that the scores of most raters are very positive, as can be seen in the overall scores for each rater source (ranging from 4.04 to 4.09 on a 5-point scale). Subordinates appear to be slightly more positive than the ratees themselves, who are in their turn slightly more positive than peers are. Supervisors are of all rater-sources the least positive. However, note that this is still a very positive score. Interesting is also the fact that supervisors are the only rater-source that significantly correlate with the other rater sources. All other combinations have extremely low correlation.

Next, the 360-degree feedback scores were compared with scores of a 'more objective' method, i.e. a coaching simulation. The 360-degree feedback scores were negatively correlated, though except for one case not significant, with the coaching simulation. Only peers have a positive, though not significant, correlation with the simulation score. This is interesting since subordinates were expected to have the most relevant view in this case, since they are the individuals that are being coached by the senior nurse. However, the peers in this case are other senior nurses, who work closely with the ratee so they may also be a relevant rater-source. Nevertheless, these results do not support 360-degree feedback as a valid method.

Next, the focus was on the ability of 360-degree feedback to measure behavioral changes. First, pretest self-scores were compared with retrospective self-scores, to study response-shift bias. It appeared that retrospective self-scores are lower than pretest self-scores, for all competencies. The fact that retrospective scores are lower, though except for creating partnership and stimulation not significantly lower, may imply that response-shift bias actually took place. However, lower retrospective scores may also have different reasons, such as the fact that ratees lower their scores after they have seen the pretest feedback report, for example because they were 'over-raters' at the pretest.

Gain scores for the experimental group and the control group were compared to see if, according to what was expected, the experimental group improved more than the control group. However, the opposite seemed to be the case. Though ratees in both groups have higher posttest scores than pretest score from themselves, their supervisors, and their peers (except for some competencies), it is striking that the subordinates' gain scores of both groups are negative. This is interesting, since subordinates are expected to have the most relevant view in this case because they are the ones that are being coached by the senior nurse. An explanation could be that the pretest has raised the consciousness of subordinates of what to expect from a coach. A similar process as the 'response-shift bias' of trainees, could also have happened to subordinates, resulting in different expectations of their coach at the time of

the posttest. In a follow-up study interviews with participants will be used to answer the question why subordinates have lower posttest scores. Interesting is also that the supervisors have the highest gain scores. Though supervisors of both ratees in group 2 and group 3, have the lowest pretest scores for almost all competencies, they have the highest posttest scores for almost all competencies. Again, interviews will be used to try to explain this result.

The results further seem to indicate that the gain scores in group 3 are, contrarily to what might be expected, higher than those of group 2. This is especially the case for self-scores and supervisor-scores. For peers, the results are more mixed, with sometimes higher scores for group 2 (dealing with stress, active listening, and diagnostic skills and overall score) and sometimes higher scores for group 3 (integrity, creating partnership, stimulation, adapting and giving feedback). For subordinates the results are also mixed, sometimes higher (or better to say: less negative) for group 2 (integrity, active listening, adapting, giving feedback and overall score) and sometimes for group 3 (dealing with stress, creating partnership, and diagnostic skills). However, in only one case is this difference in gain scores significant, and more positive, for group 3.

As table 3 has shown, the groups were different on several variables. Respondents in group 2 are more often male and on average two year older than respondents in group 3. In addition, respondents in group 3 indicate that they have more experience as a senior nurse and more experience with coaching. It may be the case that senior nurses must have some experience in their new job as a senior, and especially with coaching, before they are able to develop their coaching skills. Another explanation may be that the nurses that have been to the training need more time before they are ably to change their coaching behavior. Note again that at the time of the posttest, the training program was not yet concluded. Further analysis of the data, especially combination of the gain scores with background variables, as well as interviews with some of the participants, may clarify this interesting result.

Conclusions, Recommendations and How This Research Contributes to New Knowledge in HRD

Organizations are becoming more and more interested in evaluating their training programs at the level of behavioral change, for which 360-degree feedback is often suggested as a possible method. However, critical reflections on this purpose of 360-degree feedback can also be found in literature (see for example McLean et al, 1993). In this paper the possibility of 360-degree feedback to measure behavioral change was the object of study. Contrarily to what was expected, this study indicated more behavioral change for the control group than for the experimental group. Still, this study has provided more insight in the psychometrical properties and the usefulness of 360-degree feedback. However, further experimental research is necessary, especially with a larger number of respondents than was the case in this study and with the possibility to randomly select ratees, to answer the question whether 360-degree feedback is indeed useful for training evaluation.

References

Brown, S. and C. Seidner (1998). *Evaluating corporate training: models and issues*. Kluwer Academic Publishers.

Harris, M. and J. Schaubroeck (1988). *A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings.* Personnel Psychology, no 48, p. 35-62.

Jellema, F., A. Visscher and M. Mulder (2000). An evaluation of the quality of 360-degree assessment instruments. *Annual conference of The Academy of Human Resource Development.*

Kirkpatrick, D. (1994) *Evaluating training programs: the four levels*. San Fransisco: Berrett-Koehler.

McLean, G., M. Sytsma and K. Kerwin-Ryberg (1995). Using 360-degree feedback to evaluate management development: new data, new insights. *Annual conference of The Academy of Human Resource Development.*

Robinson, D. and J. Robinson (1989). *Training for impact: how to link training to business needs and measure the results*. San Fransisco: Jossey-Bass.

Rosti, R. and F. Shipper (1998). A study of the impact of training in a management development program based on 360 feedback. *Journal of Managerial Psychology*, no1/2, 77-89.

Sandick, A. Van and A. Schaap-Neuteboom (1993). *Rendement van een bedrijfsopleiding: een instrument voor het bepalen van het financiële rendement van trainingen* [The return on investment of a corporate training: an instrument to measure the financial results of training], PhD Thesis University of Twente.

Swanson, R. (2000). Strategic roles of human resource development in the new millennium. *Annual conference of The Academy of Human Resource Development.*