Strengthening a Comprehensive Model for Explaining HRD Effectiveness

A.A.M. (Ida) Wognum
Bernard P. Veldkamp
University of Twente

Item response theory (IRT) was used to investigate whether this method could be applied to measure HRD effectiveness. Attention was given to handling missing data, interpretation of IRT-scores, and comparison with a comprehensive model for explaining HRD effectiveness. An acceptable fit was shown. The IRT method provided much information about the performance of HRD programs with respect to components of the comprehensive model. It was concluded that the measurement of HRD effects should consist of more specific information, and should depend less on participant’s perception on HRD effects.

Keywords: HRD Effectiveness, Performance Improvement, Transfer

Problem Statement

Some years ago a comprehensive, multi-level model of HRD effectiveness was developed based on research into the quality or effectiveness of HRD programs and other learning interventions, and into factors that will enhance the effectiveness level of these programs (Wognum, 2001). In 2001 and 2002 two studies were performed to confirm and further fill this model. By means of these two studies, some of the variables of the HRD effectiveness model were confirmed or added to the model. Both studies, however, have some shortcomings.

One study was done in a large Dutch banking company. In this study the effects of a learning program about communication and cooperation on workplace behavior stood central as well as some effectiveness enhancing factors (Wognum, Veldkamp, Ankersmit & Van de Lagemaat, 2002b). Results, however, were based on a small number of respondents, which has consequences for the analysis techniques used. Also caution is called for when drawing conclusions. A large number of respondents would have underpinned the conclusions more.

The second study was concerned with the analysis of a large set of data on the effects of training programs and on effectiveness-enhancing characteristics (Wognum, Veldkamp, Ankersmit & Van de Lagemaat, 2002a). The study was based on a large sample size, i.e. 4,100 respondents, but the effectiveness measure in this study showed some disadvantages. A first problem has to do with this effectiveness measure. This measure, the Aggregated Impact Indicator (AII-score) is a weighted sum score of several components; i.e. the total satisfaction about the HRD intervention, the realization of expectations, the adjusted effect score calculated from the extent to which the objectives of the HRD assignment have been realized, the making of agreements and the extent to which these are adhered to, and satisfaction with the implementation of the program. The weights of the different components were calculated in a pilot study for a group of 400 respondents. The question remains, if these weights would have been the same when they were calculated for this new sample of 4,100 respondents. A second problem deals with missing data. In HRD research, missing data often occurs due to the practical setting of this kind of research. The AII-score, however, can only be calculated for complete cases, so either missing data have to be imputed or respondents who partly filled in the questionnaire cannot be taken into account in the analyses. A third problem is the amount of information provided by the AII-score. The AII-score expresses effectiveness in a single number. How should it be interpreted? Rankings of HRD programs are often based on these kinds of scores. They have been used for comparing the quality of these programs and the company of the suppliers of these programs, but these rankings have to be handled with care (Mulder, 1998).

To overcome these problems a valid effectiveness measure is needed to effectively determine HRD outcomes and investigate the appropriateness of the formerly mentioned comprehensive, multi-level model of HRD effectiveness. Developing such measure is important because work organizations are still making considerable investments in training and other HRD interventions (Van Buren, 2001), and managers will only be willing to opt for HRD programs that are expected to be effective and produce positive effects on the achievements not only of individual employees but also of separate departments and the entire organization. Research into HRD effectiveness has, as a result, been given significant impetus. Such research, however, has to be based on the use of valid effectiveness measures.

The problem statement thus runs as follows: Does the comprehensive HRD model for explaining the effectiveness of HRD programs fits when using an alternative effectiveness measure in order to investigate the
appropriateness of the model and to study the impact of effectiveness influencing factors on HRD effectiveness?

Some Theoretical Backgrounds

Comprehensive Model for Explaining HRD Effects

Based on research into the area of HRD effectiveness and effectiveness enhancing factors, Wognum has built a model of HRD effectiveness that was presented at the 2001 AHRD conference in Tulsa, Oklahoma (Wognum, 2001). The model of HRD effectiveness, which was adjusted to a certain extent during the presentation at the conference, was made to explain the effects of HRD initiatives, at learning, behavior and results level, from several categories of influencing factors (see Figure 1).

The model fits into the performance paradigm and the theory of HRD effectiveness. According to the performance paradigm, performance improvement is the ultimate outcome of HRD programs and other learning interventions: HRD efforts have to improve the capabilities of individual working in the organization and enhance the organizational systems in which they perform their work. The primary outcome of HRD is, therefore, not just learning, but performance at various levels (Holton, 2000). Employees must put what they have learned into practice in their working situation. In other words, there must be some transfer (Gielen, 1995; Holton, Bates, Ruona & Leimbach, 1998). Improved behavior leads, in principle, to better individual, departmental or company results. Several authors have expressed their concern over the actual accomplished degree of transfer, which have led to various research on factors influencing transfer of HRD interventions and to a number of models attempting to clarify relationships between these factors and transfer (Baldwin & Ford, 1988; Holton, 1996). More research, however, was suggested into the operationalization and impact of transfer enhancing factors (Holton, 1996).

Figure 1. A Comprehensive HRD Effectiveness Model

The HRD effectiveness model, which basically is comparable with the formerly mentioned transfer models, thus has a means-goals ordaining between ultimate effect criteria or outcomes, and supportive, effectiveness enhancing criteria, or drivers (Kaplan & Norton, 1996; Scheerens & Bosker, 1997). It here concerns a systems model of an organizationally and contextually embedded HRD or learning process, with antecedent conditions classified in terms of inputs, processes and contexts of HRD.

Based on this, Wognum distinguished various categories of effectiveness influencing factors. One category concerns Organization and HRD related contextual factors such as the problem situation that serves as starting point for HRD, the company HRD climate, position of the HRD department, and the form the HRD program takes. Strategic HRD alignment, as part of another category of organization processes, leads to decisions on strategic choices concerning on, among others, the intended goals of HRD interventions. Input characteristics include specifications of the target population, and characteristics of individual participants in the learning program. The HRD process itself has to do with learning materials, the trainer, and specifications of the learning program. The model has also a multi-level structure, where the HRD function is embedded in organizations, and individual employees who want to develop themselves are embedded in the HRD function.
By means of the formerly mentioned two studies (Wognum c.s. 2002a; 2002b), some of the variables of the HRD effectiveness model were confirmed or added to the model. For instance, based on Wognum et al. (2002a) the setting or non setting of objectives, which is an outcome of the strategic HRD alignment process, impacted the effectiveness scores. The setting of specific objectives leads to higher scores than the setting of general objectives. Also the function of those who are involved in the studied HRD program, somehow or other, showed a difference in effectiveness scores. Those who are more directly involved in the programs, such as HRD officers, heads of HRD departments and HRD coordinators within HRD departments are more positive about HRD effectiveness than are respondents who are not directly involved in programs, such as directors/works managers, line/departmental/divisional managers, heads of personnel, personnel officers and respondents from the category ‘others’. It was confirmed that the content field of HRD programs (i.e. language versus others) impacted the effectiveness scores, which is probably due to the fact that the content of language programs is more concrete and directly applicable than the content of programs such as management, communication, commerce, and marketing.

The second study (Wognum c.s. 2002b) mainly showed the importance of the decision-making or HRD alignment process prior to the learning program. In the studied process in a large Dutch banking company each participant in the program and his or her supervisor selected three competencies to which more attention should be paid during the learning process. Progress in performance on the selected competences was found to be significantly higher than for competencies that were not selected for specific attention. In this study the importance of management support, which is a contextual factor, was also found as important for encouraging transfer and with that, HRD effectiveness. This supports Xiao’s finding that it is particularly the supervisor’s behavior that is of major importance for transfer (Xiao, 1996). Also the learning program was found to be an important factor contributing to a better performance by participants of the program. The latter study also revealed that participants in the HRD program were motivated to take part in the learning program and that motivation increased further during the program. However, more research is needed to find important factors influencing HRD effectiveness. Besides, both studies had some shortcomings that will be explained below.

Item Response Theory and the Effectiveness Level of an HRD Program

In educational measurement the Item Response Theory (IRT) has proved to be a useful statistical model for estimating abilities of students (Hambleton & Swaminathan, 1985). An IRT model supposes that an overall score on a test can be predicted (or explained) in terms of one characteristic, denoted by $\theta$. In HRD measurement, this characteristic is effectiveness of the HRD program or learning event. Several components could provide information about HRD effectiveness. Information on these components could be obtained by means of, for instance, items in a questionnaire. Formally stated, an IRT model specifies a relationship between the observable answers of the respondent on the item, the properties of the item, and the effectiveness of the HRD program. Based on all the information on these items the effectiveness score can be estimated.

When the One-Parameter Logistic Model (OPLM) is applied, the relationship can be described as:

$$P_i(\theta) = \frac{\exp(a_i(\theta + b_i))}{1 + \exp(a_i(\theta + b_i))}$$

where $P_i(\theta)$ is the probability that the respondent who participated in an HRD program with effectiveness level $\theta$ gives a positive judgment to item i of the questionnaire. The parameters $a_i$ and $b_i$ are the item discrimination, and difficulty parameter respectively. The discrimination parameter denotes how will the item discriminate between high effectiveness level and low effectiveness level of HRD involvements. The difficulty parameter denotes how high the effectiveness level of the HRD involvement should be before a respondent will answer positively to the question. Based on the answers of the respondents the effectiveness level $\theta$ is estimated. This estimate is reported as the effectiveness level of an HRD program.

Research Question

The main research question in the here presented study is, whether the Item Response Theory procedure for data analysis is applicable in an HRD context and could be used to estimate HRD effectiveness. More specifically the question is whether the comprehensive model for explaining the effectiveness of HRD efforts fits to the data when the IRT procedure is used. Attention is paid to comparison of AII-scores used in former research, and the newly developed IRT-scores. Attention is also paid to the handling of missing data in the IRT procedure, and to the interpretation of IRT-scores.
Method

Sample. To investigate the appropriateness of the model in an applied context, an existing data file was used. This file contains data from 4,100 questionnaires that were completed in the period from 1993 up to and including 2000 in the context of an evaluation commissioned by the Association of Training and Educational Organizations in the Netherlands (Vetron). Vetron is an association of about 40 training companies that satisfy a number of requirements relating to professional competence, continuity, quality and method of working. They offer all kinds of management training, social skills programs, language courses, HRM and other programs to various profit and non-profit organizations in the Netherlands. The data in the existing data file came from clients of the companies that are affiliated to the Vetron. The number of these companies fluctuates, but in the period from 1993 up to and including 2000 an average of forty companies is included in the evaluation. The data were gathered by means of a questionnaire sent to a random sample from the client base of the Vetron companies, on average about 6 months after the HRD activity had been completed. This is because effects of HRD programs on working behavior and results level only become visible after some period. Respondents were informed about the investigation and could fill in the form anonymously. On average, the response percentage was 35%. The main reason for non-response was the fact that respondents were no longer in the employ of the organization involved, due to high turnover rates in the companies.

Data collection. The data of the existing data file were collected by means of a questionnaire, which had been specially developed at that time for the evaluation project (Mulder, 2000). This questionnaire was used to determine the quality of HRD assignments by commercial Vetron companies. The questionnaire contains questions about the respondent, about characteristics of the HRD assignment, about the setting of objectives, about the division of responsibilities, about the attribution of the results to the training company and, finally, questions about making agreements and adhering to these (Mulder, 2000). In accordance with the HRD effectiveness model, the questionnaire measures both characteristics that have a direct impact on HRD effectiveness and characteristics that could possibly be used as explanatory variables. At the level of strategic alignment the setting of general and specific goals were asked. At input level, the target population, learning priorities, and design of the HRD involvements were investigated. It was checked whether agreements were made about these topics, and whether the agreements were satisfied. At the level of the HRD process, the learning materials, the trainer, the length of the training, and the transfer of the training were topic. At rounding off level, evaluation procedures and transfer were evaluated. Finally, three levels of perceived HRD effects were evaluated, i.e. at learning level, working behavior, and results level, and the overall perceptions of the respondents about HRD effectiveness were checked. Based on these questions HRD effectiveness was measured. To validate the instrument, internal consistency, test-retest reliability, and inter-rater reliability were taken into account. For details about validity and reliability of the instrument, see Mulder, Van Ginkel and Nijhof (1994).

Missing data. Although 4,100 respondents returned the questionnaire, only few of them filled in all the questions. For example, one of the characteristics by which HRD effectiveness is calculated is the realization of expectations. Approximately a third of the respondents had not completed this question. A different topic was the effect of the HRD program in the working environment. Only one quarter of the respondents answered this question. The length of the questionnaire and the period between the training and the evaluation is a possible reason for these low response percentages. There are different ways of dealing with missing values. Firstly, the decision can be made to use only complete cases, which means that the data from incomplete cases are not taken into consideration. Implementing this decision would result in 200 complete cases. When the missing data were investigated, no pattern of missingness was found. Therefore it was decided to include cases where at most four out of sixteen questions that measured HRD effectiveness were missing.

Preparation of the data. Before the analyses were conducted, the data file was adjusted on a number of points in order to be able to answer the research question. In the questionnaire, most questions were answered on a 5-point Likert scale, or scored on a scale ranging from 1 to 10. However, when the data were examined it turned out that the categories 1 till 3 hardly contained any respondents. Also for the ten-point scale, the respondents were not evenly distributed. Because of this the responses were dichotomized. For the 5-point scales, the recoded scores were 0 in case the initial score was 1-4 and 1 in case the initial score was 5. The respondent now either scored unsatisfied or totally satisfied. For the ten-point scale the scores were dichotomized by recoding the scores 1-7 in 0 and 8-10 in 1.

Data analysis. First, the parameters of the IRT model were estimated in order to make sure that one latent variable, the effectiveness level, could account for all the answers of the respondents. A statistical software package for estimating IRT models, WOPIN (Citigroup, 2000), was applied. Then, the resulting IRT scores were compared with the AI scores. Rank order statistics indicate the different orderings for the effectiveness of the Vetron companies based on both methods. SPSS was applied for conducting these analyses. Finally, the resulting IRT
estimates were analyzed to find which components provide most information about the effectiveness of the programs, and which components provide fewest.

**Results**

*Fitting the IRT model.* In this research, the focus was on questions that had direct impact on HRD effectiveness. Therefore questions about the respondent, about setting of objectives, and about division of responsibilities were not taken into account. Items from the remaining questions are shown in Table 1. Five overall items; overall perception of HRD effectiveness, realization of expectations, and questions about the three output levels of Kirkpatrick, and eight more specific items about input level, HRD process, rounding off, and administration level were used. The more specific questions dealt with parts of the programs. For example, at input level, questions were posed about definition of target population, definition of learning priorities, and design of HRD program.

Table 1. Item Parameters for Different Items.

<table>
<thead>
<tr>
<th>Item</th>
<th>Discrimination</th>
<th>Difficulty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perception of HRD effectiveness</td>
<td>3</td>
<td>0.8</td>
</tr>
<tr>
<td>Realization of expectations</td>
<td>3</td>
<td>1.6</td>
</tr>
<tr>
<td>Learning results</td>
<td>2</td>
<td>0.3</td>
</tr>
<tr>
<td>Changes in working behavior</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
<tr>
<td>Organization program</td>
<td>2</td>
<td>0.3</td>
</tr>
<tr>
<td>Organization at input level</td>
<td>4</td>
<td>1.0</td>
</tr>
<tr>
<td>Performance perception at input level</td>
<td>6</td>
<td>0.9</td>
</tr>
<tr>
<td>Organization of HRD process</td>
<td>7</td>
<td>1.1</td>
</tr>
<tr>
<td>Performance perception during HRD process</td>
<td>6</td>
<td>1.0</td>
</tr>
<tr>
<td>Organization at rounding off level</td>
<td>5</td>
<td>1.2</td>
</tr>
<tr>
<td>Performance perception at rounding off level</td>
<td>4</td>
<td>1.2</td>
</tr>
<tr>
<td>Organization of program administration</td>
<td>5</td>
<td>1.4</td>
</tr>
<tr>
<td>Performance perception of program administration</td>
<td>4</td>
<td>1.2</td>
</tr>
</tbody>
</table>

These questions were grouped in two questions at input level: one about the organization and one about the perceived performance of these parts. Besides, the question of performance perception at output level is composed from perceived results on the one hand, and on enhancing these results on the other hand. It here concerns so-called transfer enhancing measures. The computer program WOPIN (Citogroup, 2000) was used to estimate the parameters of the individual questions. The OPLM model was applied. The best values for these parameters should be found interactively. The initial attempts of fitting the model did not succeed. One general item about change in working behavior had to be removed, because it did not show acceptable fit. But after that, the remaining items showed an acceptable fit. The effectiveness level can account for the scores that were obtained in the survey. The item parameters for the questions are also shown in Table 1. The estimation of IRT parameters was only based on the cases without missing data. In this way, it was guaranteed that imputation of missing data did not influence the estimation of the model.

*Interpreting the IRT parameters.* From Table 1 it can be deduced that, for instance, questions concerning the organization of the HRD process highly discriminate between HRD programs with a high and lower effectiveness score (discrimination parameter = 7). Questions concerning learning results are more problematic. Because the discrimination parameter is low (2), it can be said that these kinds of questions hardly discriminate between good and bad programs. The same is the case for questions on the perceived HRD effectiveness or marks of HRD programs (discrimination parameter = 3) and the organization of HRD programs (discrimination parameter = 2). From the analysis of the discrimination parameters, it can also be concluded that the eight specific questions have higher discriminating power. That is, these questions are more able to distinguish the programs with low effectiveness scores from programs with high effectiveness scores. In other words, overall scores about effectiveness are less precise than scores that are based on specific levels of the programs.

From analysis of the difficulty parameters, it can be concluded that realization of expectations is very hard (difficulty = 1.6). In this study, it was the most difficult question to obtain a positive evaluation. On the other hand, the difficulty parameter of the ‘Learning results’ (difficulty = 0.3) was very low. Even poor programs had high scores for learning results. Based on analysis of these questions, it can be concluded that all Vetron companies
provide programs with high learning results, but the respondents expect more. Further analysis at item level reveals that differences between difficulty parameters at specific levels are only small.

*Figure 2. Scatterplot of IRT Score and AII Score.*

Comparison of IRT and AII-scores. In the theoretical framework, some disadvantages of the AII-scores were discussed. The IRT model is proposed, in order to overcome these problems. The question remains, whether the new method results in different effect scores. Does, for instance, the order in effectiveness level of the Vetron companies change, or does the order remain the same? A scatter plot shows the relationship between both scores for individual participants (Figure 2). The correlation between the IRT-scores and the AII-scores was significant (\(\rho = 0.71, p = .000\)). However, this correlation also indicates a far from identical ordering for both methods. Therefore, the remarks in Mulder (1998) about weaknesses of rankings based on effectiveness scores, can also be made in this study.

An explanation for these differences can be found in the way the components of the Comprehensive model of HRD effectiveness (see Figure 1) are implemented in the measurement model. For the AII-score, a weighted sum of the scores on the different components was calculated, where the IRT model applied a summation logistic function to model the relationships.

IRT and missing data. One of the theoretical advantages of IRT is the possibility to handle missing data. When missing is random, an IRT model can be used to calculate the effectiveness levels, that is the IRT-score, based only on items the respondent answered to, without imputing the missing data. This property of IRT models is also being applied in the context of the Vetron data. In Figure 3, the distributions of the scores based on cases without missings (n=181) and cases with less then 5 missings (n=1082) are shown. Although the scales of both histograms differ due to different numbers of cases, it can be seen that the distributions of the scores are more or the less the same. In other words, calculation of scores based on at least 9 out of 13 questions hardly influences de score distributions. As a result, the number of cases that can be used without additional data imputations increases from 181 to 1082 cases. This implies that more accurate conclusions can be drawn about effectiveness when IRT models are applied.

*Figure 3. IRT Score Distributions Based on Complete Cases (n=181) and Cases with Less Then Five Missings (n=1082).*
Conclusion and Discussion

In this paper a comprehensive model for explaining HRD effectiveness was described (see Figure 1), and a statistical advanced method for measuring HRD effectiveness was introduced (IRT). Both the model and the effectiveness measure were applied in the context of evaluating the HRD programs of Vetron associated training institutes. The data that were collected in several surveys and were based on the several categories of the comprehensive HRD effectiveness model, showed an acceptable fit to the IRT model. Because of this, it can be concluded that IRT is applicable in an HRD context, moreover, that it is applicable in combination with the comprehensive model for explaining HRD effectiveness. Application of the IRT method resulted in more specific information about the different components of the comprehensive HRD effectiveness model. Moreover, questionnaires that are not completely filled in can also be used in the analysis without further imputation of missing data. These findings imply that measurement of HRD effectiveness should consist of more specific information on all components of the comprehensive model.

When the new effectiveness measure was applied, it was shown that overall questions about HRD effectiveness, i.e. questions on a total effect score, or mark, of an HRD program or on output levels of Kirkpatrick, are less informative than questions about specific components such as the organization of the HRD process, and the performance perception during the HRD process. Most evaluation studies into the effects of HRD programs, however, ask for respondents' perception on the overall effect of the programs in which they participated, and a mark or overall score has been asked to rate the quality of the program. Based on the results, a recommendation for evaluating the effects of HRD programs can be, that questions should ask for so-called hard evidence of the gained effects, for instance on increased figures, or performance results, in stead of depending on participant perception on HRD effects and effectiveness.

Another conclusion that could be made concerns the learning results of the respondents. In Table 1, the difficulty of this item is reported to be very small. Even Vetron companies with a low effectiveness score, could obtain high scores on learning results. Because of this, Vetron companies can be encouraged not only to focus on the learning results, but also to focus on other parts/components of the training. Focusing on issues at input level, like definition of learning of priorities, issues during the HRD process, or at rounding off level can further improve the effectiveness.

In this project, a dataset was used that has been collected in several surveys between 1993 and 2000. The instrument has not been altered in this period of time. In the past decade, new ideas and insights have been gained in the area of HRD effectiveness research. For example, the importance of transfer enhancing measures and factors influencing HRD effectiveness, and the need for hard measurement of effect levels, instead of perception measures. A next step in the development of a comprehensive model for explaining HRD effectiveness is to develop new measurement instruments, based on these new insights, which can be used for the collection of new datasets.

References


