Original Article

Measuring Performance in Dynamic Decision Making

Reliability and Validity of the Tailorshop Simulation

Daniel Danner, Dirk Hagemann, Daniel V. Holt, Marieke Hager, Andrea Schankin, Sascha Wüstenberg, and Joachim Funke

Institute of Psychology, University of Heidelberg, Germany

Abstract. The Tailorshop simulation is a computer-based dynamic decision-making task in which participants lead a fictional company for 12 simulated months. The present study investigated whether the performance measure in the Tailorshop simulation is reliable and valid. The participants were 158 employees from different companies. Structural equation models were used to test τ-equivalent measurement models. The results indicate that the trends of the company value between the second and the twelfth month are reliable variables. Furthermore, this measure predicted real-life job performance ratings by supervisors and was associated with the performance in another dynamic decision-making task. Thus, the trend of the company value provides a reliable and valid performance indicator for the Tailorshop simulation.

Keywords: dynamic decision making, complex problem solving, Tailorshop, reliability, validity

Real-life decisions are complex, and sometimes there are no well-defined solutions for problems. A manager has to make decisions even if he or she does not have all relevant information, or an employer has to pursue the interests of his staff as well as the goals of his company, even if both views may be conflicting. Gonzalez, Yanyukov, and Martin (2005) call such decisions *dynamic decisions.* They are characterized by dynamics, complexity, opaqueness, and dynamic complexity. In a similar vein, Dörner (1980) characterizes such problems as *complex problems,* which means that their structure is complex, connected, dynamic, and nontransparent. Recently, dynamic decision-making tasks were also included in the Program for International Student Assessment (PISA; Wirth & Klieme, 2003). Since the ability to deal with such problems may impact important decisions in real life, it is an interesting question whether there are individual differences in dynamic decision making, and whether these differences can be measured reliably and validly (e.g., Baker & O'Neil, 2002; Rigas, Carling, & Brehmer, 2002; Süß, 1996, 1999; Strohschneider, 1986; Zaccaro, Mumford, Connelly, Marks, & Gilbert, 2000). Investigating these issues was the aim of the present study.

To investigate dynamic decision making, several authors suggested studying behavior in computer simulations. The *Tailorshop* is one such dynamic decision-making task that has been used for several decades (e.g., Barth & Funke, 2010; Putz-Osterloh, Bott, & Köster, 1990; Süß, Kersting, & Oberauer, 1993; Wittmann & Hattrup, 2004). The scenario simulates a small business that produces and sells shirts. The participants lead this business for 12 simulated months by manipulating several variables like the number of workers, the expenses for advertising, etc. (see Figure 1).

In total, the Tailorshop consists of 24 variables, 21 of which are visible to the participants, 3 variables being invisible to the participant. Twelve variables can be manipulated directly (e.g., the costs for advertising), whereas other variables can only be manipulated indirectly (e.g., the demand). The state of a variable in a given month influences the state of the same and other variables in a following month. Figure 2 shows schematically how the variables are connected (see Funke, 1983, for an algebraic definition of all system variables).

In order to use the performance in the Tailorshop for the investigation of individual differences or for individual assessment, the performance variable should be reliable and valid. The reliability of a performance variable is important in two ways.

In a research context, reliability considerations are important for an understanding of the *validity* of dynamic decision-making measures because the reliability of a variable affects its correlation with criterion variables. In an applied context, the Tailorshop may be used to measure a

DOI: 10.1027/1614-0001/a000055

Round 1 of 12

Variable	Value	Planning
Account status	165775	
Number of shirts sold	407	
Price of raw material	3.99	
Shirts in stock	81	
Workers 50	8	
Workers 100	0	
Salary	1080	
Price of shirts	52	
Shops	1	
Worker satisfaction %	57.7	
Loss of production %	0.0	

Variable	Value	Planning
Company value	250685	
Demand	767	
Raw material in stock	16	
Machines 50	10	
Machines 100	0	
Repair & service costs	1200	
Social costs per worker	50	
Advertising costs	2800	
Business location	suburb	suburb
Machine damage %	5.9	

Figure 1. Screenshot of the graphical user interface of the Tailorshop (labels translated).

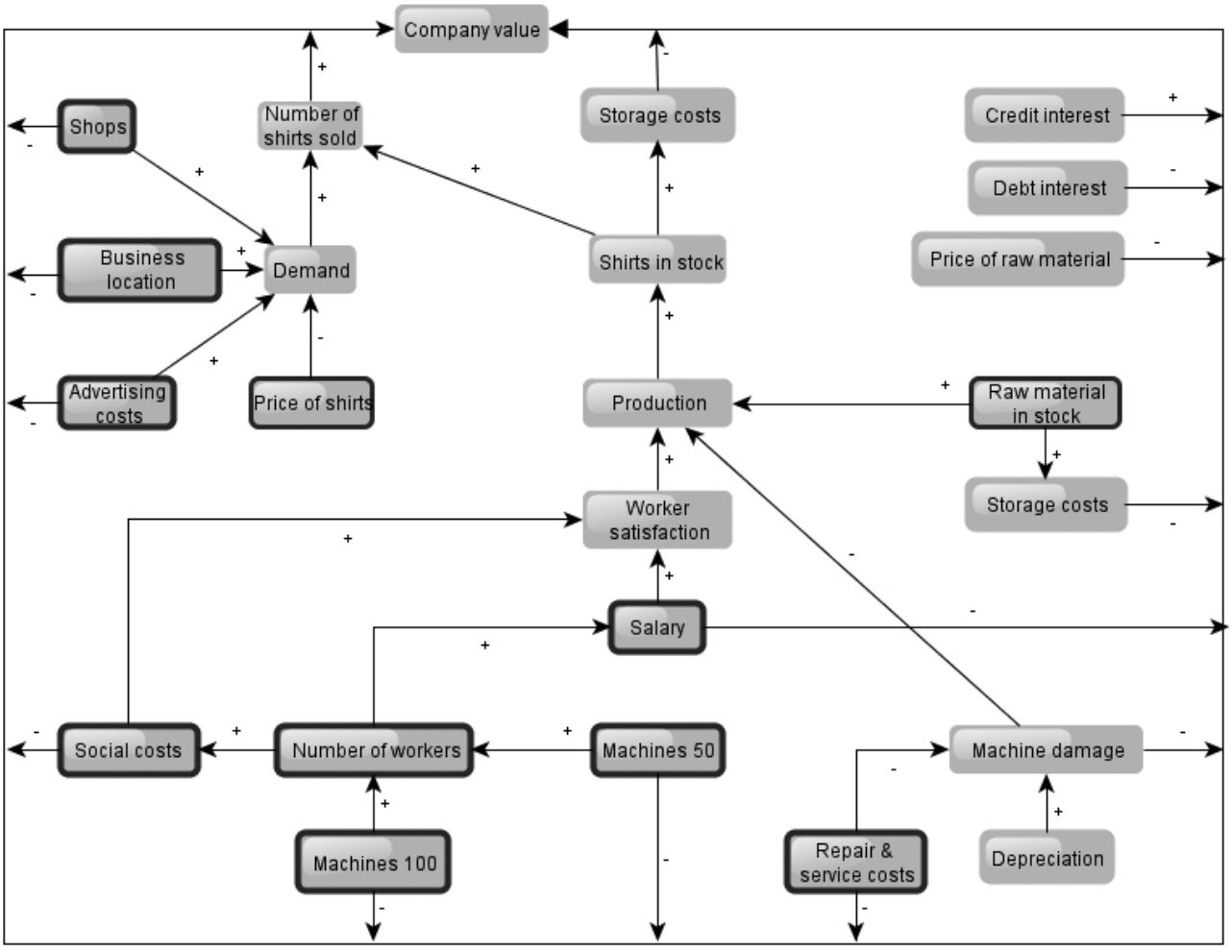

Figure 2. Schematic relation between the variables in the Tailorshop. The marked variables can be manipulated directly.

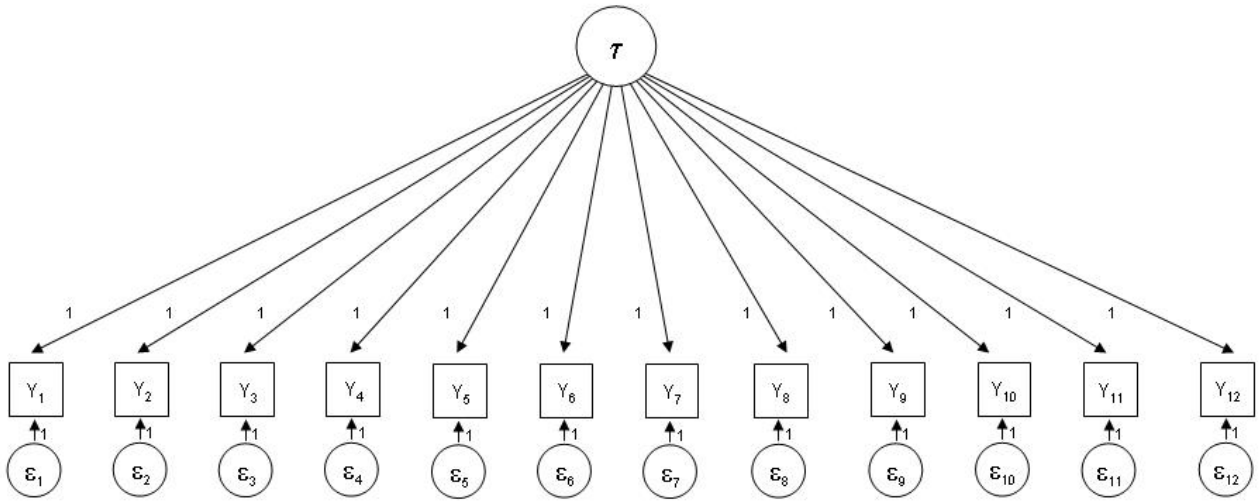

Figure 3. τ-equivalent measurement model. τ = true score variable, ε = measurement error variable.

single person's ability to solve complex problems, e.g., as part of an assessment center. This measurement is only useful if it is reliable because otherwise it yields incorrect decisions.

Reliability Estimation

In classical test theory, the reliability of a variable is defined as the proportion of the true score variance relative to the total variance of a variable (Lord & Novick, 1968). In the Tailorshop scenario, reliability is defined as the proportion of true individual performance differences relative to the total individual performance differences. The true score τ of a measurement i of a variable Y is defined as the expected value given a particular person P (Lord & Novick, 1968). In the Tailorshop scenario, the true score of a performance variable is defined as the expected performance given a particular person, $\tau_i := E(Y_i|P)$. In addition, the measurement error ε is defined as the deviation of the measured variable from the true score variable, $\varepsilon_i := Y_i - \tau_i$ (Lord & Novick, 1968). To estimate the reliability, multiple, experimentally independent measurements of a variable are necessary.

In addition, two assumptions have to be made that define the τ-equivalent measurement model. The first assumption is that the true score of a measurement i of a particular person is identical with the true score of another measurement j of this person, $\tau_i = \tau_j - \tau$. The second assumption is that the errors of the measurements are uncorrelated, $\text{cov}(\varepsilon_i,\varepsilon_j) = 0$, for all $i \neq j$. These assumptions may be tested with a structural equation model (Steyer, 1989) as shown in Figure 3. If the assumptions hold, then the variance of the true score may be estimated and the reliability may be computed by reliability $Y_i = \frac{var(\tau)}{var(Y_i)}$.

Validity Assessment

According to Dörner (1980) and Gonzalez et al. (2005) dynamic decisions are characterized by complexity, connectivity, nontransparency, and dynamics. Hence, the *content validity* of a performance variable may be evaluated regarding these four criteria. The *convergent validity* may be evaluated by the correlation with another dynamic decision-making task. Therefore, we expected a substantial correlation with the dynamic decision-making task Heidelberg Finite State Automaton (Wirth & Funke, 2005), which has also been used in the German PISA assessment in 2000 (Wirth & Klieme, 2003). The *predictive validity* may be evaluated by the correlation with real-life performance. We therefore expected that the performance in the Tailorshop can predict professional success. Finally, the *divergent validity* may be attested by a low correlation with another ability construct. Hence, we hypothesized that there is a low correlation between the performance in the Tailorshop and the performance on a standard intelligence test.

Performance Measurement

At the beginning of the simulation, the participants were instructed to maximize the company value. Thus, the success of dynamic decision making is measured by the achieved company value. The simplest approach would be to measure the company value after every month. However, the company value of a particular month depends on the company value of the previous month, company value$_i$ = company value$_{i-1}$ + change$_i$. Therefore, the company values are not experimentally independent, and the assumption of uncorrelated errors would be violated. On the other hand, there is no such relationship between the *changes of the company values*. Furthermore, the sum of the changes of the company values corresponds to the company value af-

ter 12 months because the company value at the beginning of the simulation is identical for all participants,

$$\text{company value}_{12} = \text{company value}_0 + \sum_{i=1}^{12} \text{change}_i.$$

Therefore, the *changes of the company values* after each simulated month may be taken as performance indicators for the Tailorshop simulation.

As an alternative, Funke (1983) suggested using the *trends of the company value* as performance indicators, which are binary variables. If the company value between two successive months increases, the trend is positive. If the company value decreases, the trend is zero.[1] This scoring may has several advantages. First, the trend measure is simple to interpret because each point corresponds to a month in which the given aim ("maximize company value") was achieved. Second, the trend measure is robust against outliers, whereas the change value may rise to extreme values (due to the nonlinear relationships between the variables). And finally, the measurement model for the trend measure makes fewer assumptions than the measurement model for the change measures on how the company value develops over the months. In particular, the τ-equivalent measurement model for the change measures states that the (true) change of a person is constant over time, $\tau_i = \tau_j$. On the other hand, the measurement model for the trend measures only states that a person who has a greater probability to make gain in a particular month, also has a greater probability to make gain in another month.

Aim of the Present Study

The present study investigates the reliability and the validity of (1) the change of the company value and (2) the trend of the company value. The reliabilities of these variables were investigated with τ-equivalent measurement models. Furthermore, the content, convergent, predictive, and divergent validities of these variables were evaluated.

Method

Participants

Participants were N = 158 employees (111 females, 47 males), who were recruited via newspaper announcement from various branches and companies around Heidelberg. The participants rated their jobs according to the International Standard Classification of Occupations (ISCO-88 COM): 6% rated themselves as member of a legislative body, senior officials, and managers, 25% as professionals, 11% as technical and associate professionals, 14% as clerks, 40% as service workers and shop and market sales workers, 1% as craft and related trade workers, 1% as plant and machine operators and assemblers, and 1% as unskilled occupations. The participants' mean age was M = 43.34 years (SD = 11.22).

Measures

Advanced Progressive Matrices

General intelligence was measured using the Advanced Progressive Matrices (Raven, Court, & Raven, 1994). The number of solved items in the second set was taken as a performance indicator. Cronbach's α for the 36 items was α = .85.

Heidelberg Finite State Automaton

The Heidelberg Finite State Automaton (Wirth & Funke, 2005) was used as a second indicator for dynamic decision making. The scenario is computer based and simulates a space flight where the participants control a space ship and a ground vehicle via a graphical user interface (see Figure 4). The system variables are connected and dynamic. For example, the ability to fly with the spaceship depends on the state of the propulsion, the heatshield, the landing gear, and the state of the ground vehicle. The performance was measured with 22 items, where the participants have to reach a specified target (e.g., land the spaceship on a particular planet). The number of solved items was taken as the performance variable. Cronbach's α for the 22 items was α = .93.

Tailorshop

The participants were given information about the meaning of the variables in the Tailorshop (e.g., "The account status is the amount of money in your account that is available anytime. A negative value signifies that you took a loan."). Further, the participants were instructed to maximize the company value within 12 simulated months. For the purpose of the present study, we measured (1) the changes of the company value and (2) the trends of the company value after every simulated month. (English and German versions of the Tailorshop simulation software are available at www.atp.uni-hd.de/tools/tailorshop.)

Professional Success

The participants' professional success was measured by supervisor ratings (Higgins, Peterson, Pihl, & Lee, 2007) with

[1] Due to the complex relationships between the variables, it is very unlikely to obtain a change in the company value of exactly zero. In the present study, there was always either a positive or a negative change in the company value.

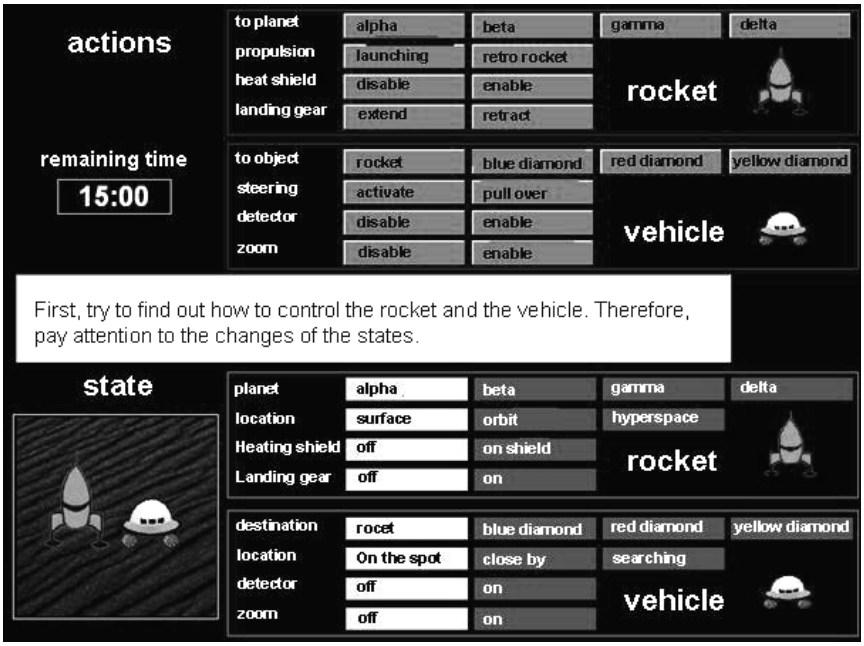

Figure 4. Screenshot of the graphical user interface of the Heidelberg Finite State Automaton (labels translated).

five items on a 6-point scale ("The employee achieves arranged and set objectives," "The employee demonstrates competence in all job-related tasks," "The employee meets all my expectations in his roles and responsibilities," "How do you rate the quality of his work?," "How do you rate the overall level of performance that you observe for this employee?"). Cronbach's α for these 5 items was α = .91. In addition, the participants' yearly income was measured in 13 categories (1 = "under EUR 2,500," 2 = "EUR 2,500 to EUR 5,000," 3 = "EUR 5,000 to EUR 7,500," 4 = "EUR 7,500 to EUR 10,000 EUR ," 5 = "EUR 10,000 to EUR 12,500," 6 = "EUR 12,500 to EUR 15,000," 7 = "EUR 15,000 to EUR 20,000," 8 = "EUR 20,000 to EUR 25,000," 9 = "EUR 25,000 to EUR 30,000," 10 = "EUR 30,000 to EUR 37,500," 11 = "EUR 37,500 to EUR 50,000," 12 = "EUR 50,000 to EUR 125,000," 13 = "over EUR 125,000").

Results

Measurement Models

The τ-equivalent measurement model was specified according to Figure 2. The measurement model for the *change variables* was estimated using the maximum likelihood procedure implemented in Mplus 5. The measurement model for the *trend variables* was estimated using the means and variance adjusted weighted least square estimator (WLSMV) implemented in Mplus 5 (Muthén & Muthén, 2007). In a first step, we estimated the measurement models for the performance indicators of all 12 months. However, the first assessment in a study may be unreliable and sometimes may not measure what is intended. Therefore, we also estimated the measurement models for the last 11 months, then for the last 10 months and so on.

Table 1. Model fit indices for the measurement models for the trend of the company value

Trend	χ^2	*df*	*p*	RMSEA	CFI
1 to 12	79.85	22	< .001	0.13	0.94
2 to 12	40.10	23	.015	0.07	0.98
3 to 12	38.08	21	.013	0.07	0.98
4 to 12	25.35	18	.116	0.05	0.99
5 to 12	17.77	16	.337	0.03	1.00
6 to 12	11.93	14	.612	0.00	1.00
7 to 12	8.51	11	.667	0.00	1.00
8 to 12	6.19	8	.626	0.00	1.00
9 to 12	2.29	5	.808	0.00	1.00
10 to 12	1.24	2	.538	0.00	1.00

Neither measurement model for the *change variables* fit the data, all $\chi^2 > 714.41$, all RMSEA > 0.71, all CFI < 0.45. However, the measurement models for the *trend variables* better fit the data. The results are reported in Table 1. As can be seen, the measurement model for the last 11 trend variables revealed an acceptable model fit, and the measurement models for the last nine or fewer trend variables fit even better. However, the fewer months included, the smaller the covariance matrix was and the fewer covariances had to be fit to the parameters of the model. Therefore, the better model fit might also be a consequence of the smaller covariance matrix. Furthermore, the dynamics during 12 months is greater than the dynamics in only the last few months. Therefore, the more months captured by a performance measure, the greater the content validity of the measure will be. Therefore, we decided to accept the

Table 2. Correlations (*p* values) between performance variables

	Change	Trend	HFA	Income	Supervisor rating
Change					
Trend	.13 (.098)				
HFA	.03 (.255)	.31 (< .001)			
Income	.01 (.923)	.08 (.323)	.05 (.561)		
Supervisor rating	.15 (.085)	.19 (.025)	.09 (.292)	–.02 (.801)	
APM	.19 (.020)	.31 (.001)	.55 (< .001)	.16 (.054)	–.03 (.706)

Notes. Change = sum of changes of the company value, Trend = sum of trends of the company value (between 2nd and 12th month), HFA = Heidelberg Finite State Automaton, Income = participants' yearly income, APM = Advanced Progressive Matrices.

measurement model for the last 11 trend variables and use it for reliability estimation.

The estimated variance of the latent τ-variable was 0.70, $p < .001$. Therefore, the reliability of each trend variable may be estimated by

$$\text{reliability trend}_i = \frac{var(\tau)}{var(trend_i)} = \frac{0.70}{1.00} = 0.70.$$

Applying the Spearman-Brown formula to estimate the reliability of the sum score of these 11 items reveals a reliability estimate of 0.96.

Correlation Between Performance in the Tailorshop and Other Variables

To evaluate the convergent, predictive, and divergent validity of (1) the change and (2) the trend of the company value, we computed the correlations between these performance variables and the performance in the Heidelberg Finite State Automaton, the participants' income, the participants' supervisor ratings, and the performance in the APM. The sum of the change variables was used as the performance indicator *change of the company value* and the sum of the trend variables (between the second and twelfth month) was used as the performance indicator *trend of the company value*.

The correlations between these variables are reported in Table 2. As can be seen, the correlation between the change variable and the trend variable was neither substantial nor significant, which suggests that both performance variables measure different performance aspects. The *change of the company value* correlated only significantly with the APM, which suggests a low overall validity of this performance variable.

On the other hand, there was a significant and substantial correlation between the *trend of the company value* and the Heidelberg Finite State Automaton, which points toward the convergent validity of the trend variable. Furthermore, there was a significant correlation between the trend variable and the supervisor ratings, which points toward the predictive validity of this measure.

There was also a substantial correlation between the trend of the company value and the APM. Therefore, we additionally computed partial correlations that were adjusted for the performance in the APM. The partial correlation between the trend variable and the Heidelberg Finite State Automaton was $r = .20$, $p = .023$, the partial correlation between the trend variable and the participants' income was $r = .05$, $p = .525$, and the partial correlation between the trend variable and the supervisor ratings was $r = .22$, $p = .010$.

Outlier Analysis

The measurement models for the trend values better fit the data than the measurement models for the change variables. One reason for this may be that the trend variables are less sensitive to outliers. To investigate the role of outliers in greater detail, we *z*-transformed the change variables for each month. There were $N = 7$ participants with $z > 3$ in at least one month. These *z*-values were trimmed to a maximum of $z = 3$ and a minimum of $z = –3$, and the measurement models were estimated again. However, the measurement model for the trimmed change values also did not fit the data, $\chi^2(65) = 1963.52$, $p < .001$, RMSEA = 0.43, CFI = 0.30.

In addition, we computed the correlations between the (sum of the) trimmed change values and the participants' scores of the Heidelberg Finite State Automaton, income, supervisor ratings, and APM. The correlation with the Heidelberg Finite State Automaton was $r = .24$, $p = .003$, the correlation with the participants' income was $r = .02$, $p = .807$, the correlation with the supervisor ratings was $r = .14$, $p = .102$, and the correlation with the APM was $r = .38$, $p < .001$. Meng, Rosenthal, and Rubin's (1992) method for comparing correlated correlations revealed that none of these correlations was significantly greater than the correlation with the trend variable.

Discussion

The present study evaluates the reliability and the validity of performance variables in the Tailorshop simulation. Therefore, we investigated (1) the change of the company value and (2) the trend of the company value.

Reliability and Measurement Models

The measurement models for the changes of the company value did not fit the data. This suggests that the single change values are not suitable for the reliability estimation. One reason for this may be that the τ-equivalent measurement model makes rather strong assumptions about how the company value develops over the months. In particular, the model states that the "true" change of the company value in the month *i* is the same than the "true" change in the month *j*, $\tau_i = \tau_i$.[2] However, this assumption may be violated because different persons may use different strategies to maximize their company value. For example, one participant may make large investments in the first month and therefore have little gain first and great gain later. Another participant may make constant investments and therefore have a constant gain across time. Hence, investigating individual differences in dynamic decision-making *processes* may be a worthwhile issue for future research. Nonetheless, the structural equation model analysis of the present study revealed that the sum of the trends between the second and twelfth month is a reliable performance variable.

Content Validity

The Tailorshop was developed according to Dörner's (1980) definition of dynamic decision making. In particular, the simulation may be seen as *complex* and *connected* because it consists of many variables that are connected. The tasks may also be seen as *nontransparent* because the participants do not know how the variables in the simulation are connected, and the tasks may be seen as *dynamic* because each intervention in the simulation influences the following state of the simulation. Therefore, the structure of the present dynamic decision-making task can be seen as a valid representation of general dynamic decision-making demands. Furthermore, the participants were instructed to maximize their company value, so that the changes in the company value as well as the trends of the company can be seen as content-valid performance measures.

Convergent Validity

The correlation between the trend of the company value and the performance in the Heidelberg Finite State Automaton was substantial and significant, which indicates the convergent validity of this variable. Furthermore, this correlation remained significant when adjusted for general intelligence, which indicates that the relationship between both dynamic decision-making tasks is incremental to the overlap with general intelligence.

On the other hand, the correlation between the change of the company value and the performance in the Heidelberg Finite State Automaton was close to zero and not significant. After controlling for outliers, this correlation increased. However, controlling for outliers may be difficult, especially in small samples or in individual assessments. Furthermore, none of the correlations with the trimmed change variable was significantly greater than the correlation with the trend variable.

Predictive Validity

The correlation between the change of the company value and the participants' supervisor ratings was not significant. However, there was a significant correlation between the trend of the company value and the supervisor ratings, which remained significant even after controlling for individual differences in general intelligence. This indicates the incremental predictive validity of the trend measure. This replicates the findings of Kersting (2001), who also reported an incremental predictive value of a dynamic decision-making measures on participants' superior ratings. Furthermore, this result points toward the practical value of dynamic decision-making measures and suggests that they may provide insights into aspects of professional success, which cannot be predicted by general intelligence.

There was no relationship with participants' income.[3] This may be due to two reasons. First, income may measure a different aspect of professional success than supervisor ratings. This is supported by the low and nonsignificant correlation between income and supervisor rating. Second, income may just be a valid indicator for professional success within an occupational category and not between. For example, a priest may earn less than a broker, even if the priest does his job better than the broker.

Divergent Validity and the Relationship Between Dynamic Decision Making and General Intelligence

Dörner and colleagues (e.g., Dörner, 1980; Dörner & Kreuzig, 1983), who introduced the construct of dynamic decision making (or complex problem solving, respectively),

[2] We additionally investigated the change variables with a τ-congeneric measurement model, which makes weaker assumptions than the τ-equivalent measurement model. In particular, the model states that the "true" change of the company in a month *i* can be linearly transformed into the true score of another month *j*, $\tau_i = \gamma^*\tau_i$. (Lord & Novick, 1968; Steyer, 1989). However, the τ-congeneric measurement model fit neither the nontrimmed change variables ($\chi^2(54) = 4582.79$, $p < .001$, RMSEA = 0.73, CFI = 0.16) nor with the trimmed change variables ($\chi^2(54) = 1605.81$, $p < .001$, RMSEA = 0.43, CFI = 0.42).

[3] Some studies (e.g., Roszkowski & Grable, 2010) report that women earn less than men. Therefore, we additionally calculated this correlation separately for women and men: There were no significant differences.

proposed that general intelligence and dynamic decision making are independent abilities. They reported several studies in which low relationships between measures of general intelligence and dynamic decision making were observed (Dörner, Kreuzig, Reither, & Staudel, 1983; Putz-Osterloh, 1981; Putz-Osterloh & Lüer, 1981). However, following studies revealed rather heterogeneous findings. Kluwe, Misiak, and Haider (1991) presented an overview of early studies and reported a broad range of correlation (between $r = –.52$ and $r = .46$), whereas subsequent studies found stronger associations (Kröner, Plass, & Leutner, 2005; Wittmann & Hattrup, 2004). One study even found a correlation between a latent intelligence and a latent dynamic decision making variable of $r = .84$ (Wirth & Kliеme, 2003).

In the present study, there was a significant correlation of $r = .31$ between the performance in the APM and the performance in the Tailorshop. In addition, there was a significant correlation of $r = .57$ between the performance in the APM and the performance in the Heidelberg Finite State Automaton. Thus, general intelligence could explain 10% (or 32%, respectively) of the variance in dynamic decision-making performance, which suggests that there is a partial but not a complete overlap between the constructs.

However, our results do not allow us to draw final conclusions about the relationship between general intelligence and dynamic decision making. In particular, Wittmann (1988; Wittmann & Süß, 1999) suggested that the relationship between two indicators only allows conclusions about the relationship between underling constructs if the indicators are symmetric. For example, the APM may be seen as an intelligence test that particularly captures individual differences in figural reasoning. In a similar vein, the Tailorshop may particularly capture individual differences in economy-related dynamic decision making. Therefore, both measures may contain not only systematic construct variance (e.g., general intelligence variance), but also "unwanted" but reliable and specific variance (e.g., specific figural reasoning variance in the APM). However, investigating the symmetry of the variables would require measuring each construct with several indicators and at several measurement occasions. Following this reasoning, the present findings cannot provide a final answer to the question on how general intelligence and dynamic decision making are related.

Performance Differences Between Men and Women

Wittmann and Hatrupp (2004) reported that men showed a better performance in the Tailorshop than women ($d = 0.70$). This finding was replicated in the present study. The number of months with a positive trend in the company value (between the second and the twelfth month) was greater for men ($M = 3.60$) than for women ($M = 2.25$), $t(156) = 2.49$, $p = .014$, $d = 0.46$. According to this, Wittmann and Hatrupp (2004) suggested that women may behave more risk-aversive than men and therefore construct for themselves a less favorable learning environment in the Tailorshop and accordingly show a lower performance. Furthermore, there were no significant performance differences between women and men in the Heidelberg Finite State Automaton or the APM, which suggests that these differences are task specific for the Tailorshop.

Conclusion

The sum of the trends between the second and the twelfth month is a reliable and valid performance indicator in the Tailorshop simulation. Hence, this score may be used for the study of individual differences as well as for individual assessments. For example, dynamic decision-making tasks may be a useful complement for the selection of job applicants as suggested by Kersting (2001).

Acknowledgments

This research was funded by German Research Foundation Grand DFG, Ha 3044/7–1. We gratefully thank Andreas Neubauer, Anna-Lena Schubert, and Katharina Weskamp for conducting the assessment and three anonymous reviewers for helpful comments on an early draft of this manuscript.

References

Baker, E. L., & O'Neil, H. F. (2002). Measuring problem solving in computer environments: Current and future states. *Computers in Human Behavior, 18*, 609–622. doi: 10.1016/S0747-5632(02)00019-5

Barth, C. M., & Funke, J. (2010). Negative affective environments improve complex solving performance. *Cognition and Emotion, 24*, 1259–1268. doi: 10.1080/02699930903223766

Dörner, D. (1980). On the difficulty people have in dealing with complexity. *Simulation and Gaming, 11*, 87–106.

Dörner, D., & Kreuzig, H. W. (1983). Problemlösefähigkeit und Intelligenz [Problem solving ability and intelligence]. *Psychologische Rundschau, 34*, 185–192.

Dörner, D., Kreuzig, H. W., Reither, F., & Stäudel, T. (1983). *Lohausen. Vom Umgang mit Unbestimmtheit und Komplexität* [Lohausen. On dealing with uncertainty and complexity]. Bern: Hans Huber.

Funke, J. (1983). Einige Bemerkungen zu Problemen der Problemlöseforschung oder: Ist Testintelligenz doch ein Prädiktor? [Issues in problem solving research: Is test intelligence a predictor after all?] *Diagnostica, 29*, 283–302.

Gonzalez, C., Vanyukov, P., & Martin, M. K. (2005). The use of microworlds to study dynamic decision making. *Computers in*

Human Behavior, 21, 273–286. doi: 10.1016/j.chb.2004.02.014

Higgins, D. M., Peterson, J. B., Pihl, R. O., & Lee, A. G. M. (2007). Prefrontal cognitive ability, intelligence, Big Five personality, and the prediction of advanced academic and workplace performance. *Journal of Personality and Social Psychology, 93*, 298–319. doi: 10.1037/0022-3514.93.2.298

Kersting, M. (2001). Zur Konstrukt- und Kriteriumsvalidität von Problemlöseszenarien anhand der Vorhersage von Vorgesetztenurteilen über die berufliche Bewährung [On the construct and criterion validity of problem-solving scenarios based on the prediction of supervisor assessment of job performance]. *Diagnostica, 47*, 67–76. doi: 10.1026//0012-1924.47.2.67

Kluwe, R. H., Misiak, C., & Haider, H. (1991). The control of complex systems and performance in intelligence tests. In H. A. H. Rowe (Ed.), *Intelligence: Reconceptualization and measurement* (pp. 227–244). Hillsdale, NJ: Erlbaum.

Kröner, S., Plass, J. L., & Leutner, D. (2005). Intelligence assessment with computer simulations. *Intelligence, 33*, 347–368. doi: 10.1016/j.intell.2005.03.002

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Oxford: Addison-Wesley.

Meng, X., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin, 111*, 172–175. doi: 10.1037/0033-2909.111.1.172

Muthén, L. K., & Muthén, B. O. (2007). *Mplus user's guide* (5th ed.). Los Angeles, CA: Muthén & Muthén.

Putz-Osterloh, W. (1981). Über die Beziehung zwischen Testintelligenz und Problemlöseerfolg [The relationship between test intelligence and problem solving success]. *Zeitschrift für Psychologie mit Zeitschrift für Angewandte Psychologie, 189*, 79–100.

Putz-Osterloh, W., & Lüer, G. (1981). The predictability of complex problem solving by performance on an intelligence test. *Zeitschrift für Experimentelle und Angewandte Psychologie, 28*, 309–334.

Putz-Osterloh, W., Bott, B., & Köster, K. (1990). Modes of learning in problem solving: Are they transferable to tutorial systems? *Computers in Human Behavior, 6*, 83–96. doi: 10.1016/0747-5632(90)90032-c

Raven, J. C., Court, J. H., & Raven, J. (1994). *Manual for Raven's Progressive Matrices and Mill Hill Vocabulary Scales. Advanced progressive matrices*. Oxford: Oxford Psychologists Press.

Rigas, G., Carling, E., & Brehmer, B. (2002). Reliability and validity of performance measures in microworlds. *Intelligence, 30*, 463–480. doi: 10.1016/s0160-2896(02)00121-6

Roszkowski, M. J., & Grable, J. E. (2010). Gender differences in personal income and financial risk tolerance: How much of a connection? *The Career Development Quarterly, 58*, 270–275.

Steyer, R. (1989). Models of classical psychometric test theory as stochastic measurement models: Representation, uniqueness, meaningfulness, identifiability, and testability. *Methodika, 3*, 25–60.

Strohschneider, S. (1986). Zur Stabilität und Validität von Handeln in komplexen Realitätsbereichen [On the stability and validity of complex problem-solving behavior]. *Sprache & Kognition, 5*, 42–48.

Süß, H.-M. (1996). *Intelligenz, Wissen und Problemlösen* [Intelligence, knowledge and problem-solving]. Göttingen: Hogrefe.

Süß, H.-M. (1999). Intelligenz und komplexes Problemlösen: Perspektiven für eine Kooperation zwischen differentiell-psychometrischer und kognitionspsychologischer Forschung [Intelligence and complex problem solving. Perspectives for a cooperation between differential-psychometric and cognition-psychological research]. *Psychologische Rundschau, 50*, 220–228. doi: 10.1026//0033-3042.50.4.220

Süß, H.-M., Kersting, M., & Oberauer, K. (1993). Zur Vorhersage von Steuerungsleistungen an computersimulierten Systemen durch Wissen und Intelligenz [On the predictability of control terformance on computer-simulated systems by knowledge and intelligence]. *Zeitschrift für Differentielle und Diagnostische Psychologie, 14*, 189–203.

Wirth, J., & Funke, J. (2005). Dynamisches Problemlösen: Entwicklung und Evaluation eines neuen Messverfahrens zum Steuern komplexer Systeme. In E. Klieme, D. Leutner, & J. Wirth (Eds.), *Problemlösekompetenz von Schülerinnen und Schülern* (pp. 55–72). Wiesbaden: Verlag für Sozialwissenschaften.

Wirth, J., & Klieme, E. (2003). Computer-based Assessment of Problem Solving Competence. *Assessment in Education: Principles, Policy and Practice, 10*, 329–345. doi: 10.1080/0969594032000148172

Wittmann, W. (1988). Multivariate reliability theory: Principles of symmetry and successful validation strategies. In J. R. Cattell (Ed.), *Handbook of multivariate experimental psychology* (2nd ed., pp. 505–560). New York: Plenum.

Wittmann, W., & Hattrup, K. (2004). The relationship between performance in dynamic systems and intelligence. *Systems Research and Behavioral Science, 21*, 393–409. doi: 10.1002/sres.653

Wittmann, W. W., & Süß, H.-M. (1999). Investigating the paths between working memory, intelligence, knowledge, and complex problem-solving performances via Brunswik symmetry. In P. L. Ackerman, P. C. Kyllonen, & R. D. Roberts (Eds.), *Learning and individual differences: Process, trait, and content determinants* (pp. 77–108). Washington, DC: American Psychological Association.

Zaccaro, S. J., Mumford, M. D., Connelly, M. S., Marks, M. A., & Gilbert, J. A. (2000). Assessment of leader problem-solving capabilities. *The Leadership Quarterly, 11*, 37–64. doi: 10.1016/S1048-9843(99)00042-9

Accepted for publication: February 14, 2011

Daniel Danner

Institute of Psychology
University of Heidelberg
Hauptstraße 47–51
69117 Heidelberg
Germany
Tel. +49 6221 547354
Fax +49 6221 547325
E-mail daniel.danner@psychologie.uni-heidelberg.de