

UNPUBLISHED MANUSCRIPT<sup>1</sup> ACCEPTED BY HUMAN PERFORMANCE

(7 October 2011)

Enhancing criterion-related validity through bottom-up contextualization of  
personality inventories: The construction of an ecological conscientiousness scale  
for Ph.D. candidates

René Butter<sup>2</sup> and Marise Born<sup>3</sup>

*Abstract*

In this paper the concept of "ecological personality scales" is introduced. These are contextualized inventories with a high ecological validity. They are developed in a bottom-up or qualitative way and combine a relatively high trait specificity with a relatively high situational specificity. An ecological conscientiousness or time management scale for Ph.D. candidates was developed. It significantly predicted Ph.D. performance criteria and showed incremental validity beyond Big Five (Study 1) and narrow trait and frame-of-reference scales (Study 2). These findings suggest that an ecological approach may contribute to further improving the criterion validity of personality measures.

---

<sup>1</sup> refer as follows: Butter, R., & Born, M. (in press). Enhancing criterion-related validity through bottom-up contextualization of personality inventories: The construction of an ecological conscientiousness scale for Ph.D. candidates. Human Performance.

<sup>2</sup> René Butter Psychologisch Advies (RBPA), Van Brakelstraat 101, 3012 XW Rotterdam, tel. 0031627037777, [rene.butter@rbpa.nl](mailto:rene.butter@rbpa.nl), [www.rbpa.nl](http://www.rbpa.nl)

<sup>3</sup> Erasmus University Rotterdam, Institute of Psychology, POB 1738, 3000 DR Rotterdam, tel. 003110 40 88814/88799

The authors thank Thomas Faith, Miriam Heemskerk and Jojanneke Struijs for their help in collecting the data

A consistent finding across several meta-analyses is that conscientiousness proves to be the best personality predictor of performance criteria. Moreover, the added value above conscientiousness of the other four personality dimensions to the prediction of overall job performance is very low (e.g., Morgeson et al., 2007a). Poropat (2009) recently conducted a large meta-analysis in the context of predicting academic performance, namely student grades. He concluded that conscientiousness turns out to be as good a predictor as intelligence in predicting academic performance.

Dudley, Orvis, Lebiecki, and Cortina (2006) conducted a meta-analysis of the role of conscientiousness in the prediction of job performance. Confirming earlier meta-analyses, conscientiousness turned out to be the personality trait with the highest criterion validity ( $r_{corrected} = .22$  on average). Primary studies that focused on subfacets or narrow traits of the broad conscientiousness trait resulted in incremental validity over the use of the broad trait. Dudley et al. (2006) distinguish four key narrow traits of conscientiousness, namely achievement, dependability, order, and cautiousness. Achievement reflects the tendency to strive for competence and success in one's work, such as adopting high standards and working to accomplish one's goals. Dependability reflects the tendency to be a reliable, trustworthy worker who is accountable, self-disciplined, and respectful of laws, regulations and authority. Order reflects the tendency to apply structure to one's working environment, that is being well organized, planful, thorough, detail-oriented and methodical. Finally, cautiousness reflects the tendency to consider risks before taking a course of action. The study of Dudley et al. (2006) confirmed the importance of a multidimensional perspective on job performance in which

specific personality variables are used for the prediction of specific performance criteria. Nevertheless, for specific contexts, such as writing a Ph.D. thesis, the above narrow traits may still lack predictive power because they are defined across work situations and are not tailored to reflect a specific work context. It is unclear why, for example, an orderliness scale would be a relevant predictor for conscientiousness in research jobs.

In order to improve the predictive validity of personality inventories, other authors (e.g., Schmit, Ryan, Stierwalt, & Powell, 1995; Bing, Whanger, Davison, & VanHook, 2004) focus on the frame-of-reference from which test takers complete personality test items. Their assumption is that each person may respond to the same personality item from a different perspective. A person may, for example, respond from a private at home perspective or from a work perspective. These different perspectives increase the error variance of the scales, and thus reduce their predictive validity. To tackle this problem, Schmit, Ryan, Stierwalt, and Powell (1995), and Bing, Whanger, Davison, and VanHook (2004) added a specific context to each item. They used an altered conscientiousness scale in which a reference to school was appended to each item, usually at the beginning or at the end of the statement. For example, the item "I strive for excellence in everything I do" was modified to "I strive for excellence in everything I do at school". They demonstrated that such items, which they labeled "context-specific items", lead to incremental predictive validity beyond global conscientiousness in predicting school performance. Lievens, De Corte, and Schollaert (2008) provide a more recent example of the frame-of-reference approach using an experimental design. They showed that context-specific items lead to a higher predictive validity as a result of the reduction of between-person variability and within-person inconsistency. They noted that

in order to reach higher predictive validity, it is important to use a frame-of-reference that is conceptually relevant to the criterion in question.

Adding such context information is useful to decrease the error variance in the measurements, and hence to increase the criterion validity. This approach will remain artificial, however, in the sense that the context information is added to standard test items regardless of whether these standard items are relevant for the specific context in the first place. It is doubtful, for example, whether a supposedly context-specific item such as “I always keep my things in a fixed place at work”, is a relevant predictor for conscientiousness in research settings.

Other scholars take an even more critical position towards the use of broad standard personality scales for predicting job performance than do the frame-of-reference researchers. For example, in Morgeson et al. (2007a, p. 715), Neal Schmitt states that he "would avoid published personality measures in almost all instances and would construct his own measures that are linked directly to job tasks in a face valid or relevant fashion". This strategy of a customized development of tests is propagated as a way to increase acceptance of personality tests, as tailor made tests may be perceived as more authentic and reflective of the realities of test-takers and client organizations (e.g., Morgeson et al., 2007b).

In line with the position taken by Neal Schmitt, we follow such a custom-developed test strategy. Based on the same measurement error reduction argument used by Schmit et al. (1995), one could also expect that better predictive validity levels might be reached through a customized development of personality scales. Like Lievens et al. (2008), we argue that apart from error reduction, also increased conceptual overlap with

the criteria plays an important role in the validity argument. We propose two basic ways of increasing conceptual overlap in this study. One can increase trait specificity and situational specificity.

Following this line of thinking, the purpose of this paper is to introduce so-called ecological personality scales. The term "ecological scale" reflects the qualitative, bottom-up development process of the items. This process guarantees a relatively high ecological validity, implying that the test items will be more job-related in the perception of respondents and therefore more motivating to complete than standard personality test items. Also, the ecological items will show more conceptual overlap with the criteria. As a result, enhanced criterion-related validity can be expected.

As an example of this approach, we present an ecological conscientiousness or time management scale for Ph.D. candidates.

*An ecological conscientiousness or time management scale for Ph.D. candidates*

The ecological conscientiousness scale in the present study was developed for the position of Ph.D. candidates in The Netherlands. A recent study for the Dutch government (Berger & De Jonge, 2005) reported that the output rates in terms of the number of Ph.D.s obtained at Dutch universities need much improvement. Only 4 to 8 percent of the Ph.D. candidates finish their Ph.D. within the generally prescribed four years. Between 17 and 33% need 5 years and about 10% drop out without a Ph.D. degree. It should be noted that, unlike in the United States, most Ph.D. candidates in The Netherlands work in full-time post-master research jobs at a university. They apply for a paid job in which they conduct a study to obtain their Ph.D. degree. A more or less

formal selection procedure is mostly utilized. The study by Berger and De Jonge (2005) indicated that personality attributes are seen as one of the important factors that determine the success of Ph.D. candidates. These authors furthermore concluded that the quality of coaching by Ph.D. advisors is perceived by Ph.D. candidates as the most important factor explaining the low output rates. Accordingly, by improving our understanding of the relation between personality characteristics and the performance of Ph.D. candidates, we also aim to stimulate that Ph.D. advisors use a more personalized coaching style.

A publication by Noordam and Gosling (2006) and their columns in *Science* (e.g., Noordam & Gosling, 2007) show that, also internationally, the attention for non-cognitive aspects of the performance of Ph.D. candidates has increased. In general, the international I/O-psychological literature hardly contains any references to the relation between personality factors and success of Ph.D. candidates. However, in studies focusing on bachelor and master levels, conscientiousness is consistently reported to be the most important non-cognitive predictor of academic success (Chamorro-Premuzic & Furnham, 2003, a; Chamorro-Premuzic & Furnham, 2003, b; Diseth, 2003; De Raad & Schouwenburg, 1996; Paunonen & Ashton, 2001).

We constructed a customized conscientiousness scale for Ph.D. candidates. Unlike the general "at-school" tag frame of reference that Schmit et al. (1995) added to each general personality item, we constructed items that each were fully derived from the specific context, in this case the "natural world of the Ph.D. candidate". Using an environmental parallel, such items may be labeled "ecological".

As was postulated by Lievens et al. (2008), further increasing the contextualization of personality scales than is achieved by frame-of-reference instruments may lead to even higher criterion-related validities. These authors also pointed out that adding an at-work or at-school tag is only the beginning of a process of further contextualizing personality inventories. In line with this view, we propose the construction of so-called ecological personality scales as a way of developing personality inventories with a higher degree of contextualization than frame-of-reference and narrow trait instruments. These last two types of instruments are both based on a top-down adaptation of generic test items in which the items constituting the scales are not directly derived from the specific context in question. In contrast, ecological scales are constructed in a bottom-up way. In order to obtain a scale that is both highly trait specific and highly situation specific, qualitative research needs to be conducted among all relevant stakeholders (for example successful job incumbents, employees who failed in a job, and co-workers and superiors of the target job holders) with respect to the behavior that is to be predicted by the test. Next, this qualitative material is translated into personality test items, such that each item directly reflects the context for which it is meant to be predictive.

It should be noted that the development process of ecological personality scales is similar to that of situational judgment tests (SJTs) in the sense that a qualitative bottom-up approach is taken. Contrary to ecological personality scales, however, SJTs typically consist of work-related scenarios. Accompanying each scenario are multiple possible ways to respond to the hypothetical situation (e.g., Chan & Schmitt, 1997, 2005; Dalessio, 1994; McDaniel, Hartman, Whetzel, & Grubb, 2007; Olson-Buchanan,

Drasgow, Moberg, Mead, Keenan, & Donovan, 1998; Weekley & Jones, 1997, 1999). The test taker then is asked to judge the possible courses of action. The situations, response alternatives and the scoring methods are mostly determined by experts in the field with the sole purpose of predicting a relevant criterion. Thus, the focus on underlying constructs is relatively small if not absent (e.g., McDaniel and Nguyen, 2001). Although ecological personality scales share a strong focus on criterion validity with SJTs, they also explicitly take into account construct validity. That is, they are developed from a personality testing perspective with the explicit goal to measure specific personality dimensions.

Our hypotheses were the following. In line with Dudley et al. (2006), who show that in order to maximize validity, traits must be selected on the basis of strong linkages to the criteria, we expected that ecological conscientiousness or time management would show significant positive correlations with performance measures that are closely related to time management (H1). In this case with *work progress*, *meeting deadlines*, and *completing the Ph.D. in time* (see the method section).

We furthermore expected that ecological conscientiousness or time management would provide incremental validity above Big Five conscientiousness (H2), because of its closer proximity to the work content which will lead to higher conceptual overlap with the criteria. H1 and H2 are addressed in Study 1.

Finally, we expected that the ecological conscientiousness scale shows incremental validity above narrow trait and frame-of-reference instruments because it combines the effects of increased trait specificity and increased situational specificity (H3). H3 is addressed in Study 2.

## STUDY 1

### Method

#### *Construction of the ecological conscientiousness or time management scale*

Following the critical incidents technique (Flanagan, 1954), qualitative interviews on success and failure factors were held with eight representatives of relevant stakeholders.

These included two Ph.D. candidates, two Ph.D. graduates, two Ph.D. drop-outs, and two Ph.D. advisors. The interviews were took place using a protocol. Thus, all interviewees answered the same questions. Example questions were

*What are the properties a Ph.D. candidate needs?, What properties can be a hindrance?, What is for you the big difference between coping with your Master's study and obtaining a Ph.D. degree?*

The resulting qualitative material was analyzed in a bottom-up way (see Miles and Huberman, 1994). This analysis yielded five patterns in the data that could be interpreted as contextualized personality traits, one of which is ecological conscientiousness or time management. Also, it turned out that the data reflected many dilemmas. For example, between getting acquainted with interesting topics versus working in a very structured way when writing a Ph.D. Next, scale items were constructed. In order to deal with the potential social desirability problem and at the same time further enrich the items with contextual clues, we used an item construction method that was similar to the procedure used by Hollenbeck (see Morgeson et al., 2007a, p. 411). Statements reflecting a high degree of conscientiousness were paired with opposite

statements reflecting a low degree of conscientiousness that were also based on the qualitative interviews. Similar bipolar formats were also used by Woods and Hampson (2005) and Duijsens and Diekstra (1995) who indicate as an advantage of the bipolar format that one pole of the item can be understood in the context of the counterpart pole. We used a four-point scale to indicate the degree of preference for either statement. The dots under the item reflect the response options. An example item is the following:

<i>I approach deadlines in a rather loose manner</i>	<i>As far as deadlines in my Ph.D. project are concerned, I am a reliable person</i>
○	○
○	○

The example item shows that the items combine a statement that reflects high conscientiousness with a statement for low conscientiousness. The high conscientiousness statements were randomly assigned to the right side or the left side of each item. Respondents need to indicate their degree of preference for the right-hand and left-hand statement by marking a dot. By marking the rightmost dot, they indicate that they definitely endorse the right-hand statement. By marking the leftmost dot, they indicate that they definitely endorse the left-hand statement. By marking the second dot from the right, they indicate that they prefer the right-hand statement, but also endorse the left-hand statement to some extent. Finally, by marking the second dot from the left, they indicate that they prefer the left-hand statement, but also endorse the right-hand statement to some extent. The conscientiousness items were embedded in a set of 62 items with identical format measuring various traits.

We constructed the items in English because many Ph.D. candidates in The Netherlands originate from other countries, and are therefore not fluent in Dutch. Moreover, also Dutch candidates need to master the English language and write their manuscripts and theses in English. Finally, some candidates working in foreign universities also participated in the study.

Sixty-two items, 10 of which supposedly measuring conscientiousness, were constructed and completed by 329 Ph.D. candidates through an e-survey. After factor analysis and item analysis, we constructed a scale of 7 items with a reasonable internal consistency ( $\alpha = .74$ ). As can be seen in the measures subsection, this reliability coefficient is somewhat lower than for the Big Five conscientiousness scale, most likely due to the lower number of items (7 vs. 10). Applying the Spearman-Brown formula (Brown, 1910; Spearman, 1910) leads to an alpha of .79, versus .80 for Big Five C. The content of the seven ecological items, which we show in the appendix, reflects time management in a research context, that is, *accurate and realistic time management in the Ph.D. work, performing one's research in a structured and self-propelled way, focusing on a long-term goal, and setting one's own priorities.*

Further evidence for the construct validity of the scale is provided by Table 1, which gives the correlations of the ecological conscientiousness or time management scale with the Big Five scales.

Table 1

*Correlations of the ecological conscientiousness or time management scale with the Big Five dimensions (N= 190)*

<i>Big Five Dimension</i>	<i>Correlation</i>
Extraversion	-.07
Agreeableness	.18*
Openness	.04
Conscientiousness	.59**
Emotional stability	.28**

\*  $p < .05$  (one-tailed)

\*\*  $p < .01$  (one-tailed)

Apart from conscientiousness ( $r=.59, p<.01$ ), significant correlations were found with Big Five dimensions emotional stability ( $r= .28, p<.01$ ) and agreeableness ( $r= .17, p<.05$ ). We note here that this correlation pattern should not necessarily be interpreted in terms of multidimensionality, as similar patterns are found among the Big Five scales in the sample.

#### *Participants and Procedure*

The criterion-related validity study was a online field survey in which the test-takers completed the ecological scale and the Big Five scales in one session in the first quarter of 2007. Additional validation data were collected in the second quarter of 2010. The ecological items were presented to each respondent in a random sequence to prevent order effects. The Big Five items were administered after the ecological items in a fixed

sequence. These items were completed by 329 Ph.D. candidates from a wide array of academic disciplines, varying from Psychology, Engineering, Medicine, Art Studies to Philosophy. The sample consisted of participants from both Dutch and non-Dutch universities (less than 5%). As indicated, all questions and items were phrased in English and no serious language problems were reported. Big Five scores were available for 190 participants and performance ratings for 128 participants. Both Big Five scores and performance ratings were available for 81 participants. The reason for this incomplete design is that not all participants allowed us to contact their advisors. Also, the Big Five items were not included in the earliest phase of the data collection.

In the first data collection step, we approached the participants through their institutions or through Ph.D. candidate networks. Next, we contacted them individually using standardized mail and telephone protocols. We also asked the participants in the e-survey for their permission to contact their Ph.D. advisors in order to collect performance ratings. We developed a concise email questionnaire with performance rating items for the advisors. The advisors who were indicated by their Ph.D. candidates were informed that a specific candidate had participated in the study and had given permission to contact the advisor in question. The advisor was then asked to complete the performance ratings.

The data collection phase for this study was demanding. At times, Ph.D. candidates or advisors were reluctant to participate. This is understandable given the potentially sensitive relationship between a Ph.D. candidate and his or her advisor in The Netherlands. Thus, many possible participants had to be contacted in order to obtain sufficient participants. There are reasons to believe that well-achieving Ph.D. candidates are overrepresented in the sample, potentially leading to reduced variance. Also, of the

128 Ph.D. candidates for whom we were able to collect performance ratings, about 25% were rated by the same advisors. The overlap appeared to be limited, however, as only one advisor rated four candidates and two advisors rated three candidates. Therefore, we did not take these dependencies into account in the data analysis.

### *Measures*

We used a 50 item Big Five instrument taken from the International Personality Item Pool (Goldberg, 1999) to examine the incremental validity of the ecological conscientiousness or time management scale. Each dimension of this instrument is measured by 10 items using a five point Likert scale ranging from 1 (very inaccurate) to 5 (very accurate) for each of the five dimensions. An example for conscientiousness is: "I follow a schedule". Three of the fifty items were replaced by alternative IPIP items: (*I get caught up in my problems* instead of *I often feel blue*) or were slightly changed (*Neglect my duties* instead of *Shirk my duties* and *Get tasks done right away* instead of *Get chores done right away*) because of potential language difficulties for the respondents. With respect to these alterations, we emphasize that we used an English questionnaire because it should also be suitable for non-Dutch Ph.D. candidates. However, most respondents are not native speakers of English and an item like "I feel blue", which might be evident for a native speaker of English, could easily be misinterpreted by a non-native English speaker. Also, a word like "chore" is unknown to most non-native English speakers and was therefore replaced by "task". The scale reliabilities for the IPIP scales in the sample are .89 for extraversion, .73 for agreeableness, .80 for conscientiousness, .85 for emotional

stability and .77 for intellect. These are very similar to those reported by Goldberg (1999).

### *Criterion measures*

Performance criteria were derived from competency profiles used by universities and directly from the "Ph.D. work itself". The following criteria were used. Academic quality, Performance with respect to publications, Research progress, Teaching, Meeting deadlines, Performance during conferences, Interpersonal functioning, Probability to obtain the Ph.D. in time, and General functioning. The criteria were measured using 10-point Likert scales ranging from (1 very poor) to 10 (excellent) in equal steps. We used this 10-point scoring format because it is widely used in the Dutch schooling system in general, from primary school up to university level. Therefore, the poles and the intermediate steps are familiar to respondents, also in organizational settings. Thus, we could be reasonably confident that all respondents attached a similar meaning to the score points.

Only with respect to the estimated probability to obtain the Ph.D. in time, an open format was used. Respondents could indicate a percentage between 0 and 100 %.

Criterion ratings were obtained for 128 candidates. As was indicated above, the relation between Ph.D. candidates and their supervisors is sensitive. This might explain the relatively low variance in the criterion measures, which was possibly caused by an overrepresentation of well-achieving participants.

Dudley et al. (2006) demonstrate the importance of selecting performance criteria with an a priori linkage to the narrow traits used as predictors. In line with the narrow

trait predictor "time management", we used a set of three performance criteria, that is, 1) research progress ( $M = 7.53$ ,  $SD = 1.07$ ), 2) meeting deadlines ( $M = 7.53$ ,  $SD = 1.42$ ) and 3) estimated probability to obtain the Ph.D. within the prescribed period (mostly four years) ( $M = 81\%$ ,  $SD = 17\%$ ). These three performance criteria also turned out to be related on the basis of a principal components analysis with varimax rotation. The analysis yielded a solution with two components that explain 62% of variance. On the first component, work progress, meeting deadlines and the probability to obtain the Ph.D. within the prescribed period clearly have the highest loadings. On the second component, performance during conferences, teaching and interpersonal functioning load highest. As can be seen in Table 2, time management is significantly related to Component 1, but not to Component 2. This result underlines the importance of trait-criteria matching, as is advocated by Dudley et al. (2006).

## Results

Table 2 gives the entire correlation matrix of predictor and criterion measures.

Table 2

*Correlations between all variables in Study 1*

<i>Variables</i>	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>	<i>7</i>
<i>Criteria</i>							
1 Research progress	1						
2 Meeting deadlines (N=125)	.70**	1					
3 Estimated probability to obtain the Ph.D. in time (N=121)	.66**	.57**	1				
4 Component 1 (N=128)	.87**	.83**	.84**	1			
5 Component 2 (N=128)	.21*	.14	.13	0	1		
<i>Predictors</i>							
6 Time management (ecological C) (N=128)	.25**	.31**	.38**	.35**	-.13	1	
7 Big Five conscientiousness (N=81)	.06	.09	.12	.08	-.09	.59**	1

Note. Component 1 reflects the variables (performance criteria) 1, 2 and 3. Component 2 reflects the criteria performance during conferences, teaching and interpersonal functioning.

\*  $p < .05$  (two-tailed)

\*\*  $p < .001$  (two-tailed)

Correlations with research progress ( $r = .25, p < .01$ ), meeting deadlines ( $r = .31, p < .01$ ), estimated probability to obtain the Ph.D. within the prescribed period ( $r = .38, p < .01$ ) and Component 1 ( $r = .35, p < .01$ ) are significant. Thus, H1 could be confirmed.

Inspection of Table 2 also shows that the correlations of Big Five conscientiousness with the same criteria are lower than those for time management and all non-significant.

Apparently, this scale, which is devoid of context, shows no predictive power.

Table 3 presents the results of the hierarchical regression analyses used to test the incremental validity of ecological conscientiousness or time management above Big Five conscientiousness.

Table 3

*Hierarchical regression results of the performance criteria on Big Five conscientiousness (first step) and time management or ecological C (second step)*

Performance criteria	Beta	R <sup>2</sup>	ΔR <sup>2</sup>	N
Research progress	.28*	.06	.06*	81
Meeting deadlines	.27*	.06	.05*	79
Estimated probability to obtain the Ph.D. within the prescribed period	.34**	.10	.09**	77
Combined criterion (Component 1)	.35**	.10	.09**	76

\*  $p < .05$  (two-tailed)

\*\*  $p < .001$  (two-tailed)

On average, we found an significant ΔR<sup>2</sup> of .07 for the four criteria, which means that H2 could be confirmed.

## STUDY 2

## Method

Study 1 showed that the ecological conscientiousness scale has incremental validity above the Big Five, which is a broad type of personality inventory. In this sense, Study 1 was only a first step in demonstrating the added value of the ecological scale. Thus, in line with Dudley et al. (2006), we also needed to demonstrate its incremental validity in direct competition with more narrowly defined scales. Accordingly, Study 2 compares the ecological scales to more specific instruments that are mentioned in the literature, that is, frame-of-reference scales that are more situation-specific than the Big Five and narrow trait scales that are more trait-specific than the Big Five. We used the 10 Big Five C items used in Study 1 preceded by an "at work tag" as frame-of-reference (FOR) items. Next, we chose achievement from the narrow traits described by Dudley et al. (2006) because it seemed the most relevant (or least irrelevant) narrow trait for predicting Ph.D. success. The ten narrow trait (NT) items were also taken from the International Personality Item Pool (Goldberg, 1999). Example items for both scales are respectively "In my work I pay attention to details" and "Excel in what I do".

First, a small rating study was conducted in which 12 independent judges (Ph.D. candidates who were unfamiliar with the ecological items) rated each item of the three instruments in terms of 1) trait specificity (i.e., the extent to which it measures a narrow aspect of personality rather than a broad characteristic) and 2) situational specificity (i.e., the extent to which it covers the Ph.D. context). The ratings were made on a 10-point scale ranging from 1 "not at all" to 10 "very much so".

For "trait specificity", the average ratings (across judges and items) for the narrow trait, the frame-of-reference scale and the time management scale were respectively 3.83, 4.23 and 5.70. For "situational specificity", the averages were respectively 5.12, 5.18 and 6.10. Thus, we can conclude that the time management items were perceived to be more "trait specific" than narrow trait items ( $t=4.03$ ,  $df=11$ ,  $p<.01$ ) and more "situation specific" than frame-of-reference items ( $t=3.55$ ,  $df=11$ ,  $p < .01$ ).

### *Procedure*

To investigate whether the increased specificity of the ecological items would also lead to enhanced prediction, the incremental validity of the ecological scale incremental validity above narrow trait and frame-of-reference scales was investigated using an additional online study. The same items as in the specificity study were used. The ecological scale (EC) items were presented with the described 4-point response scale. For the other items, the same 5-point scale as for the Big Five items was used. Preceded by the performance ratings, the 27 personality items were completed in a random order by 99 Ph.D. candidates (55 female and 44 male) who were not included in the earlier studies. Their mean age was 27.59 years ( $SD=3.20$ ). All scales showed a reasonable internal consistency ( $\alpha = .72$ ,  $.69$  and  $.82$ , for EC, FOR and NT respectively).

*Criterion measures*

Respondents rated their own performance with respect to research progress and meeting deadlines using the same format as for the supervisor ratings in Study 1, that is, research progress ( $M = 6.86$ ,  $SD = 1.37$ ), 2) meeting deadlines ( $M = 7.32$ ,  $SD = 1.50$ ) and 3) estimated probability to obtain the Ph.D. within the prescribed period (mostly four years) ( $M = 73\%$ ,  $SD = 31\%$ ).

## Results

Table 4 gives the entire correlation matrix of predictor and criterion measures.

*Correlations between all variables in Study 2*

<i>Variables</i>	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>
<i>Criteria</i>						
1 Research progress	1					
2 Meeting deadlines (N=96)	.68**	1				
3 Estimated probability to obtain the Ph.D. in time (N=79)	.37**	.33**	1			
<i>Predictors</i>						
4 Time management (ecological C) (N=87)	.38**	.45**	.40**	1		
5 Frame-of-reference C (N=85)	.23*	.34**	.33**	.56**	1	
6 Narrow trait C (N=83)	.59**	.49**	.23*	.53**	.30**	1

\*  $p < .05$  (two-tailed)\*\*  $p < .001$  (two-tailed)

Tables 5 and 6 present the results of the hierarchical regression analyses used to test the incremental validity of ecological conscientiousness above frame-of-reference or narrow trait conscientiousness.

Table 5

*Hierarchical regression results of the performance criteria on frame-of-reference C (first step) and ecological C (second step)*

Performance criteria	Beta	R <sup>2</sup>	ΔR <sup>2</sup>	N
Research progress	.36**	.14	.09**	85
Meeting deadlines	.38**	.21	.10**	85
Estimated probability to obtain the Ph.D. within the prescribed period	.33*	.18	.07*	70

\*  $p < .05$  (two-tailed)

\*\*  $p < .001$  (two-tailed)

Table 6

*Hierarchical regression results of the performance criteria on narrow trait C (first step) and ecological C (second step)*

Performance criteria	Beta	R <sup>2</sup>	ΔR <sup>2</sup>	N
Research progress	.11	.35	.01	83
Meeting deadlines	.27*	.29	.05*	85
Estimated probability to obtain the Ph.D. within the prescribed period	.39**	.16	.11**	72

\*  $p < .05$  (two-tailed)

\*\*  $p < .001$  (two-tailed)

On average we found an significant  $\Delta R^2$  of .07 for the three criteria, implying that H3 was confirmed.

### *Discussion*

Summarizing the results of Studies 1 and 2, we see that the ecological conscientiousness scale correlates highest with the Big Five conscientiousness scale. Furthermore, it is significantly related to directly relevant performance criteria and unrelated to less relevant criteria. In predicting these performance criteria, the ecological scale shows incremental validity above Big Five, narrow trait and frame-of-reference scales. Thus, we may conclude that the ecological conscientiousness (or time management) scale has an adequate construct validity and criterion validity.

In line with Lievens et al. (2008), we believe that the road to improving the criterion validity of personality inventories lies in further increasing the contextualization of these instruments. With that objective in mind, the ecological approach to personality scale development is based on qualitative interviews with relevant stakeholders which leads to richer and more valid measurements than can be obtained by using uncontextualized, standardized instruments only (see also Butter, 2011).

Both narrow trait and frame-of-reference instruments can be considered as relatively top down or off-the-shelve approaches. They are based on standard personality items with or without a general context tag added. This means that they are susceptible to

the criticism that the existing test items that are used might be suboptimal predictors for the specific context in the first place. In other words, they might not have enough conceptual overlap with the criteria. As was mentioned by Lievens et al. (2008), it is the amount of conceptual overlap that determines the criterion validity of personality test items. There are two ways to increase conceptual overlap. One can increase trait specificity and situational specificity. The rating study confirmed that ecological items are indeed perceived to be more trait specific than narrow trait items and more situation specific than frame-of reference items.

#### *Implications for practice*

Opinion surveys among Ph.D. candidates indicate that they experience a high need for a more personalized coaching relation with their advisors (Berger & De Jonge, 2005; Meijer, 2002). This suggests that lack of coaching may be a factor that contributes to the high drop-out rates in The Netherlands, and thus to personal, social and economic losses. We believe that the situation in other countries will be rather similar.

Accordingly, the practical aim of this study was to improve personality-based prediction of the performance of Ph.D. candidates, which is important to improve selection and competency management in general for Ph.D. candidates. We focused on conscientiousness in this study because this construct is the most consistent predictor of job performance over a wide range of jobs (e.g., Behling, 1998). With respect to selection, the interviews indicate there is reason to believe that the personal competencies needed for specific projects are not always examined when selecting Ph.D. candidates for projects. Ecological personality scales can play a role here.

As for coaching and training purposes, for example, time management training could be offered to candidates scoring low on ecological conscientiousness at an early stage, that is, before any delay in the progress of the research work actually occurs. Also, we think that our findings may help to fine-tune the management style by Ph.D. advisors. Candidates scoring low on time management might make more research progress when a clear form of structure is provided to them by their advisors, for example by setting out milestones in the first year of the research work. Until now, not much was known on this topic.

Finally, our results suggest that ecological scales are worth considering as a strategy to improve the validity of personality instruments in all situations where the economies of scale that are needed to enable and justify the development of tailor-made instruments can be realized.

#### *Limitations and future research directions*

The respondents were job incumbents, that is, Ph.D. candidates and university professors who work under considerable time pressure. Obtaining a reasonable sample size was difficult and it was essential to minimize the task load for respondents.

Therefore, frame-of-reference items and narrow trait items were not included in Study 1 and Big Five items were dropped in Study 2.

Also, Study 1 hinged on supervisor ratings of performance, whereas in Study 2 self-ratings were used.

The ecological conscientiousness scale was constructed following a bottom-up process. Personality items were derived from qualitative interviews with relevant

stakeholders who are experts with respect to the context in question. The purpose of these interviews was to search for information that is directly relevant for predicting the criteria. Because of this bottom-up construction process, the ecological approach combines a trait specificity and situational specificity strategy. This strategy most likely explains the increased criterion validity of the ecological scale above narrow trait and frame-of-reference scales. Further experimental research is needed, however, to disentangle the effects of trait specificity and situational specificity. Finally, future research should also address the issue of fakeability of ecological scales. If it can be shown that these scales are more resistant to faking than narrow trait and frame-of-reference scales, their practical utility will be further increased.

## References

- Berger, J., & De Jonge, J. (2005). *Rendement verkend. Succes en faalfactoren van promotietrajecten aan Nederlandse Universiteiten* [Output explored. Success and failure factors in Ph.D. trajectories at Dutch Universities]. Zoetermeer, The Netherlands: EIM Onderzoek voor Bedrijf en Beleid.
- Behling, O. (1998). Employee selection: Will intelligence and conscientiousness do the job? *Academy of Management Executive*, 12, 77-86.
- Bing, M. N., Whanger, J. C., Davison, H. K., & VanHook, J. B. (2004). Incremental validity of the frame-of-reference effect in personality scale scores: A replication and extension. *Journal of Applied Psychology*, 89, 150-157.
- Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, 3, 296-322.
- Butter, R. (2011). Dancing with structure: research in ecological pedagogy. *Quality & Quantity*, 45, 739-741.
- Chamorro-Premuzic, T., & Furnham, A. (2003a). Personality traits and academic examination performance. *European Journal of Personality*, 17, 237-250.
- Chamorro-Premuzic, T., & Furnham, A. (2003b). Personality predicts academic performance: Evidence from two longitudinal university samples. *Journal of Research in Personality*, 37, 319-338.
- Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perception. *Journal of Applied Psychology*, 82(1), 143-159.

- Chan, D., & Schmitt, N. (2005). Situational judgment tests. In A. Evers, N. Anderson, & O. Voskuil (Eds.), *Handbook of personnel selection* (pp. 219-246). Oxford: Blackwell.
- Dalessio, A. T. (1994). Predicting insurance agent turnover using a video-based situational judgment test. *Journal of Business and Psychology, 9*(1), 23-32.
- De Raad, B., & Schouwenburg, H. C. (1996). Personality in learning and education. *European Journal of Personality, 10*, 303-336.
- Diseth, A. (2003). Personality and approaches to learning as predictors of academic achievement. *European Journal of Personality, 17*, 143-155.
- Dudley, N. M., Orvis, K. A., Lebiecki, J. E., & Cortina, J. M. (2006). A meta-analytic investigation of conscientiousness in the prediction of job performance: Examining the intercorrelations and the incremental validity of narrow traits. *Journal of Applied Psychology, 91*, 40-57.
- Duijsens, I. J., & Diekstra, R. W. F. (1995). The 23bb5: a new bipolar big five questionnaire. *Journal of Personality and Individual Differences, 19*(5), 753-755.
- Flanagan, J. C. (1954). The critical incidents technique. *Psychological Bulletin, 51*(4), 327-358.
- Goldberg, L. R. (1999). A broad band-width, public domain, personality inventory measuring the lower-level facets of several five factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorp (Eds.), *Personality Psychology in Europe* (Vol. 7, pp. 7-28). Tilburg, Netherlands: Tilburg University Press.
- Hough, L. M., & Oswald, F. L. (2005). They're right, well mostly right: Research evidence and an agenda to rescue personality testing from the 1960s

- insights. *Human Performance*, 18, 373-387.
- Lievens, F., De Corte, W., & Schollaert, E. (2008). A closer look at the frame-of-reference effect in personality scale scores and validity. *Journal of Applied Psychology*, 93, 268-279.
- McDaniel, M. A., & Nguyen, N. T. (2001). Situational judgment tests: A review of practice and constructs assessed. *International Journal of Selection and Assessment*, 9(1/2), 103-113.
- McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L., III (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel Psychology*, 60(1), 63-91.
- Meijer, M. M. (2002). Behoud Talent!: *Een rapportage over de verschillende aspecten die een rol spelen bij de begeleiding van promovendi*. [Retain Talent!. A report on several aspects playing a role in advising Ph.D. students]. Utrecht: LAIOO.
- Miles, M. B., & Huberman, A. M. (1994). *An expanded sourcebook. Qualitative data analysis*. London: Sage.
- Morgeson, P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007a). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology*, 60, 683-72.
- Morgeson, P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007b). Are we getting fooled again?: Coming to terms with limitations in the use of personality tests for personnel solution. *Personnel Psychology*, 60, 1029-1049.

- Noordam, B., & Gosling, P. (2006). *Mastering your Ph.D.: survival and success in the doctoral years and beyond*. New York: Springer.
- Noordam, B., & Gosling, P. (2007). Mastering your Ph.D.: relating to your co-workers' personality types. Retrieved on August 31, 2007 from [http://sciencecareers.sciencemag.org/career\\_development/previous\\_issues/articles/2007\\_07\\_20/caredit\\_a0700104/\(parent\)/68](http://sciencecareers.sciencemag.org/career_development/previous_issues/articles/2007_07_20/caredit_a0700104/(parent)/68).
- Olson-Buchanan, J. B., Drasgow, F., Moberg, P. J., Mead, A. D., Keenan, P. A., & Donovan, M. A. (1998). An interactive video assessment of conflict resolution skills. *Personnel Psychology*, 51(1), 1-24.
- Paunonen, S. V., & Ashton, M. C. (2001). Big Five predictors of academic achievement. *Journal of Research in Personality*, 35, 78-90.
- Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin*, 135, 322-338.
- Schmit, J., Ryan, A. M., Stierwalt, S. L., & Powell, A. B. (1995). Frame-of-reference effects on personality scale scores and criterion-related validity. *Journal of Applied Psychology*, 80, 607-620.
- Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, 3, 271-295.
- Weekley, J. A., & Jones, C. (1997). Video-based situational judgment testing. *Personnel Psychology*, 50(1), 25-49.
- Weekley, J. A., & Jones, C. (1999). Further studies of situational tests. *Personnel Psychology*, 52(3), 679-700.

Woods, S. A., & Hampson, S. E. (2005). Measuring the big five with single items using a bipolar response scale. *European Journal of Personality, 19*, 373-390.

APPENDIX

Items of the ecological conscientiousness scale

<p><i>1) I approach deadlines in a rather loose manner</i></p>	<p><i>As far as deadlines in my Ph.D. project are concerned, I am a reliable person</i></p>
<p><i>2) My time management in my research is very accurate and realistic</i></p>	<p><i>I like working with rather global time-limits; more detailed time-limits turn out to be unattainable most of the time</i></p>
<p><i>3) I can keep myself going</i></p>	<p><i>I need input from my surroundings to keep my research going</i></p>
<p><i>4) I am able to get my priorities right in my research work</i></p>	<p><i>I need others helping me out to get my priorities right in my research work</i></p>
<p><i>5) I keep appointments precisely</i></p>	<p><i>Keeping appointments is less important than the substantial quality of my Ph.D. work</i></p>
<p><i>6) I like working for a long-term goal in my research</i></p>	<p><i>I prefer working with short time horizons</i></p>
<p><i>7) Working on a Ph.D. project means operating in a very structured way</i></p>	<p><i>Working on a Ph.D. project means living through a series of unplanned and disordered events</i></p>

