Available online at www.sciencedirect.com
ScienceDirect
PERSONALITY AND INDIVIDUAL DIFFERENCES
Personality and Individual Differences 42 (2007) 1403–1413
www.elsevier.com/locate/paid

Modeling individual differences in contrapositive reasoning with continuous latent state and trait variables

Jean-François Bonnefon [a,*], Stéphane Vautier [b], Michael Eid [c]

[a] *Laboratoire Travail et Cognition, Centre National de la Recherche Scientifique, Maison de la Recherche, 5 allées A. Machado, 31058 Toulouse Cedex 9, France*
[b] *Centre d'Etudes et de Recherches en Psychopathologie, Université Toulouse-Le Mirail, Maison de la Recherche, 5 allées A. Machado, 31058 Toulouse Cedex 9, France*
[c] *Department of Psychology, Free University of Berlin, Habelschwerdter Allee 45, D-14195 Berlin, Germany*

Received 4 April 2006; received in revised form 4 October 2006
Available online 28 November 2006

Abstract

Recent developments in the psychology of judgment and reasoning have emphasized the importance of individual differences in responses to reasoning experiments. This shift in emphasis, from the *modal* performance to *variations* in performance, raises some methodological concerns about the reliability and validity of the scores obtained when measuring reasoning performance: Differences in these scores might reflect measurement errors or reactions to pragmatic and semantic contextualization rather than differences in reasoning ability. This article shows how latent state-trait modeling can be applied to the responses collected in standard reasoning experiments. A latent state-trait analysis of contrapositive reasoning ("if p then q; not-q; what follows?") revealed that the variables used to assess contrapositive reasoning ability showed remarkably good reliability and are only weakly affected by situation-specific influences. Consequently, they are appropriate for assessing stable interindividual differences in abilities. Moreover, the hypothesis of the existence of a general task-independent reasoning ability has to be rejected because the different task-specific abilities are only moderately correlated (between .27 and .63).

Keywords: Reasoning; Contrapositive; Latent state; Latent trait; Reliability; Stability; Specificity

* Corresponding author. Tel.: +33 5 61 50 36 11.
E-mail address: bonnefon@univ-tlse2.fr (J.-F. Bonnefon).

0191-8869/$ - see front matter
doi:10.1016/j.paid.2006.10.017

1. Introduction

Recent years have seen a promising rapprochement of two approaches to thinking (in general) and deductive reasoning (in particular). The first approach, rooted in cognitive psychology, uses experimental tasks and emphasizes general cognitive processes; the second approach, rooted in psychometrics, focuses on individual differences. The integration of these two approaches is still at a preliminary stage: In a typical study, participants engage in an experimental task, complete a psychometric test, and their responses to the task are correlated with their scores on the test. This approach offers valuable insights into the links between scores gathered by using experimental tasks and psychometric tests, but the correlations are often rather low and the reason why is not very clear. One explanation might be found in the insufficient psychometric properties of data from the experimental tasks measuring cognitive processes: These tasks might be appropriate for experimental studies, but not for the measurement of individual differences.

In this article, we show how modern psychometrics can help to obtain information about the reliability, the stability, and the specificity of scores from experimental tasks. In particular, we want to know whether these tasks capture enduring individual differences rather than situation-specific influences, and how strongly individual differences generalize across different tasks aimed at measuring the same cognitive operation. We first outline the recent attempts at linking experimental and psychometric approaches to reasoning with an emphasis on the questionable psychometric properties of responses to reasoning tasks. We then present the basic ideas of latent state-trait theory and discuss how it can be applied to the psychology of reasoning. To illustrate this, we focus on contrapositive reasoning (i.e., correctly answering "not-*p*" to such problems as "if *p* then *q*; not-*q*; what follows?"), and fit a latent state-trait model to data collected from 484 participants varying widely in age, occupation, and education.

2. Individual differences in reasoning, and the psychometric properties of reasoning tasks

The experimental approach to human reasoning has traditionally focused on the modal response to reasoning tasks, a response that numerous studies of deductive and probabilistic reasoning have shown to be biased, following the initial demonstrations of Wason (1966) and Kahneman et al. (1982). The fact that most participants in these experiments deviate from normative models started a raging debate on human rationality, which in turn served as a Trojan horse for the consideration of individual differences in reasoning performance.

Reacting to the widespread suggestion that the modal response to various tasks was being compared to ill-chosen normative models (e.g., Koehler, 1996; Oaksford & Chater, 1994), Stanovich and West (2000) made a compelling argument that normative models ought to concur with the responses of participants with higher intellectual ability, rather than with the responses of the majority of participants. And indeed, using the Scholastic Aptitude Test as a measure of participants' intellectual ability, Stanovich and West (1998) observed that high ability is related to normative answers on numerous tasks (e.g., syllogisms, selection task, covariation detection; see also Klaczynski, 2001, on the selection task, and Torrens, Thompson, & Cramer, 1999, on syllogisms). In a similar vein, Newstead, Handley, Harley, Wright, and Farelly (2004) correlated several reasoning tasks (e.g., selection task, syllogistic reasoning, conditional arguments) to a number of psychometric scales (e.g., the Rational Experiential Inventory for thinking style, Pacini & Epstein, 1999, or the AH5 group test of intelligence, Heim, 1968). While intellectual ability was a good predictor of syllogistic reasoning performance, its relation to performance on the selection task was not as clear. Intellectual ability correlated with normative responses to the deontic selection task in two studies but not in the third, whereas it correlated weakly with normative responses to the indicative selection task in the third study but not at all in the first and second.

Most interestingly from our perspective, Newstead et al. (2004) considered the possibility that these inconsistencies might be due to the poor psychometric properties of scores derived from the selection task itself – and it seems likely that this argument may also explain why contrapositive reasoning performance did not correlate with any test score. Stanovich (1999) also expressed concern that the modest correlations between scores on different reasoning tasks might be the consequence of modest reliability: In his words, reasoning tasks "are not the equivalent of the powerful psychometric instruments that are used in most research on individual differences in cognition and personality" (p. 37). Indeed, low correlations between the performance on an experimental task and the score on a psychometric test raises the question of whether the experimental responses have the same reliability and validity as the scores from tests of, say, intellectual ability or thinking style. Yet only if these psychometric properties are warranted can the correlations be reasonably interpreted.

Nevertheless, why would reasoning tasks not measure the same general reasoning ability as psychometric tests? In which way are these tasks different from standard psychometric tools? There are at least two reasons for these differences, which are due to the double contextualization of reasoning tasks as compared to psychometric tests.

Firstly, reasoning tasks are semantically contextualized, that is, they make use of everyday linguistic content rather than abstract material. This semantic content tends to evoke background knowledge which will influence respondents independently of their reasoning ability. For example, consider the argument: "If John studies hard, he does well on the test; John studies hard; does it follow that he does well on the test?" The logical answer is of course that he does. Nevertheless, many respondents will think of their experience with tests and think that studying hard is not enough to do well if the test itself is very hard, and they will reject the logically correct answer (De Neys, Schaeken, & d'Ydewalle, 2002).

Secondly, reasoning tasks are pragmatically contextualized, that is, unlike ability tests, they do not make clear whether or not everyday conversational assumptions should be disregarded while completing the task. This ambiguity has a wide array of consequences for reasoning tasks (e.g., Bonnefon & Hilton, 2004; Politzer & Bonnefon, 2006; see Politzer, 2004, for a review). To give only one example, it is common in everyday conversations that conditional statements come with an implicit assumption that their converse is true (Geis & Zwicky, 1971): When one says "If she is promoted, she will make more money", we are usually permitted to assume that one also meant "If she is not promoted, she will not make more money". With this assumption, it is legitimate to conclude "She will not make more money" from the premise "If she is promoted, she will make more money; she is not promoted". This conclusion is nevertheless logically unwarranted, and would not provide support to one's reasoning ability.

Consequently, if responses to reasoning tasks are to be compared to responses to psychometric tests, individual differences in the observed scores should:

1. Mainly represent reliable individual differences and not measurement errors. This would address Stanovich's (1999) concern that low inter-task correlation is due to large error variance.
2. Generalize across different semantic contents rather than be specific to a single content. This would ensure that individual differences reflect a difference in reasoning ability rather than different reactions to semantic contextualization.
3. Represent temporally stable differences rather than situational differences. This would again ensure that individual differences reflect a difference in reasoning ability rather than different reactions to pragmatic contextualization. (Here we assume that situational differences applying systematically to all semantic content have to result from different attitudes toward the task at different points in time. Such an influence belongs to the domain of pragmatic contextualization.)

The analysis of these three requirements can shed new light on the reasons for the poor association between experimental reasoning tasks and psychometric reasoning tests. All these issues can be addressed through modern psychometric modeling, as we will show in the next sections. Although the integration of the experimental and the psychometric approaches is necessary for a deeper understanding of human reasoning, practically no study has yet applied modern psychometric models to the reasoning tasks of cognitive psychology (see Rijmen & De Boeck, 2003, for an exception). The present study will show how latent state-trait modeling can help researchers to obtain more information about the psychometric properties of data from experimental reasoning tasks.

3. Latent state-trait modeling of individual differences

Latent state-trait (LST) models are extensions of classical test theory's true-score models (Steyer, 1989; Steyer, Ferring, & Schmitt, 1992, 1999) that allow for an appropriate analysis of individual differences measured at different time points. The basic idea of LST modeling is that each observation consists of three components. The first component represents the stable ability of the individual as measured by a given elementary task (an item) and is called the latent trait variable or trait factor. This does not mean, however, that the model is only appropriate for personality traits. The term "trait" is only used to indicate that this variable indicates the stable part. Depending on the observed variables, the latent trait variables can represent different constructs (e.g., abilities as in the current case). The second component indicates the systematic influence of the situation in which the measurement takes place or the interaction between the individual and the situation. The third component consists of measurement error. To measure stable individual differences, the observed scores should depend more strongly on the stable ability component than on situational influences and measurement error. Furthermore, to separate the different components, it is necessary to have at least two items and two occasions of measurement.

Fig. 1 shows a Multistate-Multitrait (MSMT) model for a two-wave (test–retest) design with six items. Variables Y_{11} to Y_{61} refer to the first-wave measurements of items 1 to 6, and variables Y_{12} to Y_{62} refer to the second-wave measurements of items 1 to 6. The error components E_{11} to E_{62}

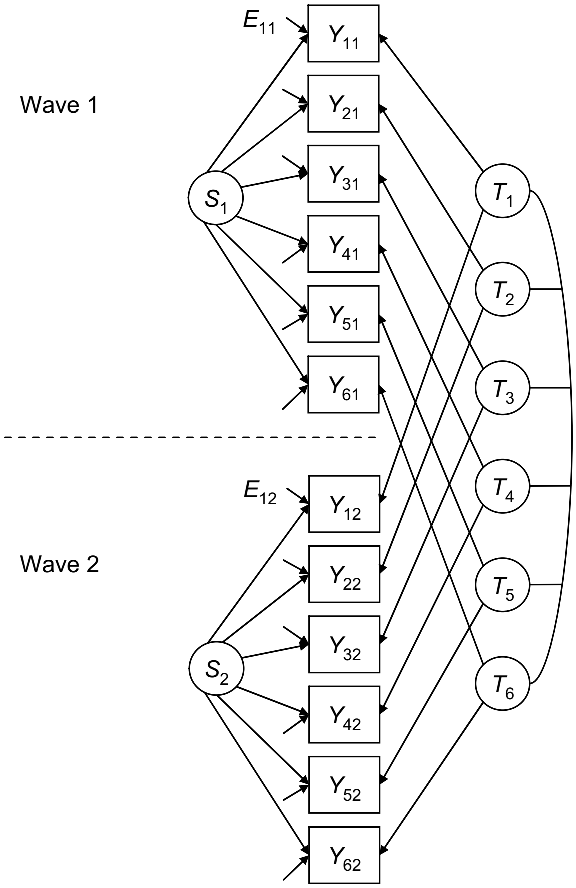

Fig. 1. Multistate-multitrait model for a two-wave design with six items.

are uncorrelated with each other as well as with the other factors. The factors S_1 and S_2 define wave-specific, item-unspecific sources of variance. They represent influences that are specific to an occasion of measurement. The trait factors T_1–T_6 define item-specific, wave-unspecific sources of variance. They represent stable individual differences in abilities, that means, individual differences that are neither due to situational influences nor to measurement error. By definition, the factors S_k and T_i are uncorrelated. The factors S_1 and S_2 are uncorrelated, a structural assumption that allows the interpretation of stability as the result of temporally stable individual differences. The MSMT model does not assume that each item measures the same individual differences, and the correlational structure of the trait factors is freely estimated. This allows the analysis of whether the different tasks measure the same ability or not. If the different latent trait variables are perfectly correlated, the different tasks measure the same ability. If the correlations between the trait variables are low, this indicates that the different tasks do not measure the same ability and that the semantic content plays an important role.

To evaluate whether the contrapositive reasoning tasks are appropriate for measuring individual differences, the reliabilities of the different tasks can be estimated. The reliability of a task is the proportion of variance explained by all latent variables, that is, the trait variables T_i and the occasion-specific variables S_k. The reliability coefficient is defined on the decomposition of an observed variable in the MSMT model

$$Y_{ik} = \lambda_{Sik} S_k + \lambda_{Tik} T_i + E_{ik},$$

where λ_{Sik} and λ_{Ti} are loading parameters. The reliability is the proportion

$$\mathrm{REL}(Y_{ik}) = \frac{\lambda_{Tik}^2 \mathrm{var}(T_i) + \lambda_{Sik}^2 \mathrm{var}(S_k)}{\mathrm{var}(Y_{ik})}$$

If contrapositive reasoning tasks are appropriate for analyzing individual differences, they should show high reliability. Based on the decomposition of the observed variables into the linear function of the latent variables, two other coefficients can be defined. The consistency coefficient

$$\mathrm{CON}(Y_{ik}) = \frac{\lambda_{Sik}^2 \mathrm{var}(T_i)}{\lambda_{Tik}^2 \mathrm{var}(T_i) + \lambda_{Sik}^2 \mathrm{var}(S_k)}$$

indicates the proportion of true (error-free) variance that is due to stable individual differences and not due to occasion-specific influences. The specificity coefficient, on the other hand, is the proportion of true variance due to variable occasion-specific influences

$$\mathrm{SPE}(Y_{ik}) = \frac{\lambda_{Sik}^2 \mathrm{var}(S_k)}{\lambda_{Tik}^2 \mathrm{var}(T_i) + \lambda_{Sik}^2 \mathrm{var}(S_k)}$$

The consistency and the occasion-specificity coefficients add up to one. If experimental contrapositive reasoning tasks assess reliable and stable individual differences with respect to a general reasoning ability they should show high reliability, high consistency, low specificity, and large intertrait correlations. Any deviation of these properties could indicate an important source of the poor associations between experimental reasoning tasks and psychometric tests of general reasoning abilities.

4. Application to contrapositive reasoning

Contrapositive reasoning amounts to deriving the conclusion "not-*p*" from premises "if *p* then *q*" and "not-*q*". The semantic content of propositions *p* and *q* varies from one problem to another, for example, "if a company's venture makes huge profits, its stocks go up", or "if a patient has malaria, he makes a quick recovery". Within the MSMT approach, the question arises of whether each content elicits its own specific trait. Besides the analysis of the reliability, consistency, and specificity, we are primarily interested in testing the hypothesis that all items measure the same trait factor (general ability) against the hypothesis that each item captures a specific trait factor (reaction to the specific semantic content).

5. Method

5.1. Participants, material, and design

Participants were recruited by third-year psychology students as a course credit requirement. Each student first made a list of several men and women that were older than 18, not studying psychology, and willing to take part in a two-wave survey on reasoning. Each student then randomly selected one man and one woman from his or her list, and asked them to take part in the study. It was expected that this recruitment procedure would promote variety in age, occupation, and education, while ensuring equal proportions of male and female participants.

Of the 484 participants who returned two fully completed questionnaires (49% men, 51% women, mean age = 31, SD = 12.5), 20% had completed graduate school or an equivalent school form, 40% had the equivalent of an undergraduate education, 25% graduated from high school only, and the educational level of 15% was lower than high school. The sample included a large proportion of students (37%), but the remaining 63% came from practically all professional perspectives (including 10% unemployed).

The contrapositive reasoning task consisted of six problems, all in the following form: (a) the conditional rule is introduced, embedded in a simple context; (b) the consequent of the conditional is negated; and (c) participants are asked whether the antecedent of the conditional is true in that situation. Response categories were "Yes", "No", and "Maybe". Here is a complete example:

> You are a doctor in a tropical country. According to your experience, *if a patient has malaria, he makes a quick recovery*. You observe the following situation: A patient is not making a quick recovery. Does the patient have malaria?

The other five conditionals (all taken from Thompson, 2000, and translated into French) were:

1. If there is a low pressure system, then it will rain.
2. If a restaurant sells liquor, then it must have a liquor license.
3. If someone has broken an item in the store, then they must pay for it.
4. If a company makes a big profit, then the price of their shares will go up.
5. If the content of the bottle is poisonous, then it must be labeled "poison".

Participants completed the tasks a first time, and then a second time after a minimum time period of three weeks had passed (mean = 26 days, SD = 4.6). The survey included other tasks in addition to the reasoning task that we also address in this article (e.g., a well-being questionnaire), and the contrapositive reasoning questions did not appear in one block, but were interwoven among other conditional reasoning questions.

5.2. Analyses

We examined the statistical fit of two MSMT models. Model 1 is the MSMT model presented in Fig. 1. Performance on a contrapositive reasoning item can be coded as normatively correct (1) or incorrect (0), making it a dichotomous manifest variable Y_{ik}. Because the observed variables are binary variables, we specified a MSMT model for binary variables. In this model, each observed binary variable Y_{ik} is assumed to be linked to a latent, item-specific, continuous, and normally distributed variable Y^*_{ik} by the following measurement model:

$$Y_{ik} = \begin{pmatrix} 0 & \text{if } Y^*_{ik} \leqslant \tau_{ik} \\ 1 & \text{if } Y^*_{ik} > \tau_{ik} \end{pmatrix},$$

where the threshold parameter τ_{ik} cut the continuous variable into two parts. The latent variable Y^*_{ik} can then be decomposed into the trait, occasion-specific, and error variables as the observed variables in the MSMT model for continuous variables. All coefficients are defined in an analogous way.

In addition, we assumed measurement invariance of the loading and threshold parameters belonging to the same task across waves. Furthermore, the variances of the components S_1 and S_2 are also assumed to be invariant across waves. To test the hypothesis of a multitrait vs. a single latent trait structure, we compared Model 1 to a model with only one latent trait factor (Model 2), which implied that all trait factors are perfectly correlated. The correlational structure of the trait factors was examined to assess their discriminant and convergent validities. Even though the present study was not geared toward providing information about the external validity of the reasoning scores, we introduced three covariates in the model: gender, age (young vs. old), and education. All analyses were performed using the weighted least square means and variance adjusted (WLSMV) estimator and theta parameterization as implemented in the Mplus software (Muthén & Muthén, 2004).

6. Results

Model 1 fitted the data well, $\chi^2(30, N = 484) = 38.76$, $p = .13$. Table 1 displays the consistency, specificity, and reliability estimates. Because of the time-invariant loading parameters, the coefficients belonging to the same task do not differ between the two occasions of measurement. The observed variables are, therefore, noted as $Y^*_{i\bullet}$, using the dot to replace the occasion of measurement. The reliability (of the variables $Y^*_{i\bullet}$) ranges from .63 to .91, a finding which shows that the reliabilities are rather good given that the observed variables are single binary items. This means that the items are capable of measuring individual differences in contrapositive reasoning reliably.

Table 1
Consistency, specificity, and reliability coefficients for the six tasks based on the MSMT model

	CON($Y_{i\bullet}$)	SPE($Y_{i\bullet}$)	REL($Y_{i\bullet}$)
$Y^*_{1\bullet}$	0.95	0.05	0.69
$Y^*_{2\bullet}$	0.92	0.08	0.63
$Y^*_{3\bullet}$	0.93	0.07	0.86
$Y^*_{4\bullet}$	0.85	0.15	0.91
$Y^*_{5\bullet}$	0.93	0.07	0.71
$Y^*_{6\bullet}$	0.91	0.09	0.80

Table 2
Correlational structure of the trait factors estimated by the MSMT model (standard errors of the estimates appear in parentheses)

	T_1	T_2	T_3	T_4	T_5
T_2	.56 (.08)				
T_3	.27 (.08)	.36 (.09)			
T_4	.28 (.09)	.35 (.10)	.43 (.08)		
T_5	.35 (.08)	.40 (.09)	.63 (.06)	.60 (.08)	
T_6	.35 (.08)	.31 (.09)	.31 (.08)	.55 (.08)	.48 (.07)

Traits T_1 to T_6 correspond to the *shares*, *rain*, *license*, *pay*, *poison*, and *recovery* problems, respectively (in reference to the consequent of the conditional).

The consistency coefficients range from .85 to .95, and the specificity coefficients range from .05 to .15. The interpretation of these values denotes a small contribution of the wave-specific variance to the explained variance compared to the contribution of wave-unspecific variance. In other words, the response to a contrapositive reasoning task does not depend much on the occasion of measurement. Instead, individual differences in reasoning tasks mainly represent stable abilities. Table 2 displays the correlations between the trait factors, along with their standard errors. The correlations ranged from .27 to .63. Since these correlations are corrected for attenuation due to measurement error and occasion-specific influences, they show that the different responses do not reflect one single trait. A multidimensional structure is necessary to represent the similarities and differences between the tasks appropriately.

Finally, the model supplemented with the gender, age, and education variables fitted the data well, $\chi^2(38, N = 484) = 39.60$, $p = .40$. In this model, the six trait factors are the dependent variables and the gender, age, and education variables are the independent variables. In four tasks, men ($n = 237$) exhibited a higher mean than women ($n = 247$) with the standardized regression coefficients ranging from .12 to .36 – yet the reverse was true for two tasks, with standardized regression coefficients of $-.19$ and $-.16$. The age variable revealed a mixed pattern, with four positive but small standardized regression coefficients ranging from .06 to .22 and two negative standardized regression coefficients of $-.45$ and $-.22$. The unique contribution of the education variable to the variance of the six dependent variables did not exceed 2.6%.

7. Discussion

The high reliability and consistency coefficients indicate that the different reasoning tasks measure stable individual differences in abilities with sufficient precision. They are not only useful for experimental research but also good measures for assessing and exploring individual differences. The high consistency coefficients show that contrapositive reasoning tasks have trait-like character because the responses they elicit are only slightly affected by situational variability. With respect to these properties, contrapositive reasoning tasks are similar to items of psychometric test assessing general abilities such as intelligence.

However, the results downplay the role of a general (task-unspecific) contrapositive reasoning ability, and hint at the multidimensional nature of the contrapositive reasoning tasks. About two thirds of the inter-trait correlations reported in Table 1 are below .40, even though they are corrected for attenuation. In other terms, Stanovich's (1999) suggestion that low inter-task correlations are due to large error variances appears ill-founded: Correcting for error influences does not seem to significantly improve the correlations. This result is a strong argument against the idea that all six tasks measure the same (and only the same) construct, even when correcting for random and systematic transient error. Moreover, although the age and gender covariates revealed only small effect sizes, their directions are inconsistent: If trait variables reflected only one construct, age and gender should show consistent associations with all six trait variables. Hence, the six tasks are clearly not unidimensional.

Even though different responses to the contrapositive reasoning task mainly represent reliable individual differences rather than measurement error, and even though these differences are stable across time (i.e., respondents do not appear to change their pragmatic assumptions toward the task across different occasions), individual differences in the observed scores seem to reflect different reactions to different semantic contents rather than differences in general reasoning ability. As a consequence, what is measured by cognitive psychologists in reasoning experiments might, to a large extent, be different from what is measured by tests of standard ability.

These results are the basis for a more refined analysis of the association between contrapositive reasoning tasks and psychometric ability tests in future studies. According to our results, the poor association is neither due to measurement error nor to situation-specific influences. The major reason for this finding might be the context-specificity of reasoning tasks. Future studies have to explore the reasons for this context dependency. It might also be important to analyze the correlations between psychometric ability scales and single contrapositive reasoning tasks in order to find out whether there are contexts that are more prone to the influence of general ability. Moreover, it would be interesting to see whether single contrapositive reasoning tasks would show incremental validity beyond psychometric ability tests when it comes to the prediction of behavior. All these questions can be analyzed in the methodological frame presented here, and answers to these questions would shed more light on the important, but not yet sufficiently explored, field of research on individual differences in cognition.

References

Bonnefon, J. F., & Hilton, D. J. (2004). Consequential conditionals: invited and suppressed inferences from valued outcomes. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 30*, 28–37.

De Neys, W., Schaeken, W., & d'Ydewalle, G. (2002). Causal conditional reasoning and semantic memory retrieval: a test of the semantic memory framework. *Memory and Cognition, 30*, 908–920.

Geis, M. L., & Zwicky, A. (1971). On invited inferences. *Linguistic Inquiries, 2*, 561.

Heim, A. W. (1968). *AH5 group test of intelligence [manual]*. London: National Foundation for Educational Research.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: heuristics and biases*. Cambridge: Cambridge University Press.

Klaczynski, P. A. (2001). Analytic and heuristic processing influences on adolescent reasoning and decision making. *Child Development, 72*, 844–861.

Koehler, J. J. (1996). The base rate fallacy reconsidered: descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences, 19*, 153.

Muthén, L. K., & Muthén, B. O. (2004). *Mplus 3.11 [Computer software]*. Los Angeles (CA): Muthén and Muthén.

Newstead, S. E., Handley, S. J., Harley, C., Wright, H., & Farelly, D. (2004). Individual differences in deductive reasoning. *Quarterly Journal of Experimental Psychology A, 57*, 3360.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review, 101*, 608–631.

Pacini, R., & Epstein, S. (1999). An evaluation of dual-process theories of reasoning. *Journal of Personality and Social Psychology, 76*, 972–987.

Politzer, G. (2004). Reasoning, judgement, and pragmatics. In I. Noveck & D. Sperber (Eds.), *Experimental pragmatics* (pp. 94–115). London: Palgrave.

Politzer, G., & Bonnefon, J. F. (2006). Two varieties of conditionals and two kinds of defeaters help reveal two fundamental types of reasoning. *Mind & Language, 21*, 484–503.

Rijmen, F., & De Boeck, P. (2003). A latent class model for individual differences in the interpretation of conditionals. *Psychological Research, 67*, 219–231.

Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*. Mahwah (NJ): Lawrence Erlbaum.

Stanovich, K. E., & West, R. F. (1998). Individual differences in rational thought. *Journal of Experimental Psychology: General, 127*, 161–188.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: implications for the rationality debate? *Behavioral and Brain Sciences, 23*, 645–665.

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

Steyer, R., Ferring, D., & Schmitt, M. J. (1992). States and traits in psychological assessment. *European Journal of Psychological Assessment, 8*, 79–98.

Steyer, R., Schmitt, M. J., & Eid, M. (1999). Latent state-trait theory and research in personality and individual differences. *European Journal of Personality, 13*, 389–408.

Thompson, V. A. (2000). The task-specific nature of domain-general reasoning. *Cognition, 76*, 209–268.

Torrens, D., Thompson, V. A., & Cramer, K. M. (1999). Individual differences and the belief bias effect: mental models, logical necessity, and abstract reasoning. *Thinking and Reasoning, 5*, 128.

Wason, P. (1966). Reasoning. In B. M. Foss (Ed.), *New horizons in psychology* (pp. 106–137). Harmandsworth (UK): Penguin.