The value of response times in item response modeling

Molenaar, D.

DOI
10.1080/15366367.2015.1105073

Publication date
2015

Document Version
Final published version

Published in
Measurement

License
Article 25fa Dutch Copyright Act

Measurement, 13: 177–181, 2015

ISSN: 1536-6367 print / 1536-6359 online
DOI: 10.1080/15366367.2015.1105073

The Value of Response Times in Item Response Modeling

Dylan Molenaar
Department of Psychology, University of Amsterdam

A new and very interesting approach to the analysis of responses and response times is proposed by Goldhammer (this issue). In this approach, differences in the speed-ability compromise within respondents are considered to confound the differences in ability between respondents. These confounding effects of speed on the inferences about ability can be controlled for in experimental settings. As a result, the data for psychological or educational inferences consists of the response vectors only. The response time vectors are redundant as these are equal for all respondents (at least in the response signal paradigm as preferred by Goldhammer, this issue, and Goldhammer & Kroehne, 2014). To assess the merit of this promising approach by Goldhammer, a straightforward question that arises is: Why are we interested in differences in response times? Below, I will argue that the natural variability in response times can give valuable information for psychological and educational inferences about response processes and solution strategies but that the approach by Goldhammer is very valuable if a single process or strategy needs to be measured in isolation.

WHY ARE WE INTERESTED IN RESPONSE TIMES?

Traditionally, inferences about abilities have been based on the responses of respondents to the items of a psychological or educational test. Using a suitable measurement model, ability is represented as a latent variable θ that underlies the responses. The interest in the time needed by the respondents to solve the test items dates back many decennia (Thorndike, Bregman, Cobb, & Woodyard, 1926), and measurement models for the joint analysis of responses and response times are available for many years already (e.g., Roskam, 1987). However, it has only been recently that due to the increased popularity of computerized test administration, response times have become available in testing practices. The fact that the response times are now so easily available does not imply that they are useful. Perspectives about the value of response times differ among different approaches.

Measurement precision and response process

First, from a more statistically oriented perspective, response times are seen as an additional source of information about θ. That is, the response times are added to the traditional measurement models to improve measurement precision about θ, mostly resulting in an additional latent

Correspondence should be addressed to Dylan Molenaar, Department of Psychology, University of Amsterdam, Weesperplein 4, 1018 XA, Amsterdam, The Netherlands. E-mail: D.Molenaar@uva.nl

speed variable ζ. The degree to which the response times may add information to the measurement of θ depends on the cross loadings of the response times on θ (in models like that of Thissen, 1983, and Ferrando & Lorenzo-Seva, 2007, for instance) or similarly on the correlation between θ and ζ denoted ρ (in models like that of Van der Linden, 2007, and Klein-Entink, Fox, & Van der Linden, 2009).

From a second, more substantive oriented perspective, the responses are seen as the result of an underlying response process. The response time is then an indication of the time it took for the process to start, develop, and end. Using models like the Q-diffusion model (Van der Maas, Molenaar, Maris, Kievit, & Borsboom, 2011), the race model (Ranger, Kuhn, & Gaviria, 2015; Rouder, Province, Morey, Gomez, & Heathcote, 2015), and the proportional hazards model (Ranger & Kuhn, 2014), the response times together with the responses can be used to make inferences about the response process. Making inferences in terms of the underlying response process is valuable in testing practices; that is, test validity is ideally established by pinpointing the exact processes that transmit differences between respondents into differences in the responses (Borsboom, Mellenbergh, & Van Heerden, 2004).

Between-respondents and within-respondent differences

By conditioning on ζ, as proposed by Goldhammer (this issue), we thus loose statistical information about θ and substantive information about the response process. This hampers inferences about qualitative differences between respondents and within respondents.

Between-respondents differences

Response times include important information about qualitative differences between respondents in the response process. These differences are reflected statistically in θ and ζ. For instance, Van der Maas and Jansen (2003) showed that children differ in their solution strategies to solve the balance scale task. In this task, a picture of a balance scale is displayed with equally heavy weights placed at pegs situated at an equal distance from the fulcrum. The items differ in how the weights are configured across the balance scale. For some problems, 2 solution strategies can lead to the correct answer. However, as shown by Van der Maas and Jansen, the solution strategies differ in the time needed to apply them. Therefore, θ reflects whether a respondent used one of the correct strategies and ζ reflects whether the respondent used the efficient or the inefficient strategy. Inferences can then be made about what strategy a given respondent likely used. Molenaar, Tuerlinckx, and Van der Maas (2015) proposed a mixture extension of the model by Van der Linden (2007) to detect these between-respondents differences in strategy use.

Within-respondent differences

Differences within-subjects in ζ and θ may arise if the respondent does not work at a constant speed and constant ability through the test. Statistically, these differences are absorbed in the residuals of the measurement models for ζ and θ, resulting in conditional dependence between the responses and the response times. This poses no problems for the statistical model as approaches exists to test these violations (e.g., Bolsinova, & Maris, in press; Van der Linden & Glas, 2010) and model these violations (e.g., Meng, Tao, & Chang, 2015; Ranger & Ortner, 2014).

Although statistically unproblematic for the assessment of ζ and θ, the within-respondent differences may contain information about differences in the response process. That is, if the response times are corrected for the differences between respondents (i.e., ζ) and the differences between items (i.e., time intensity), the remaining residual differences may reflect differences in the response process; that is, a large residual response time may indicate that a respondent has chosen a different response strategy or response process on that item as compared to the other items (see Van der Linden & Guo, 2008, for a possible approach). This information is valuable in many situations. Three examples are given below.

Post-error slowing. Post-error slowing is the phenomenon that respondents tend to take more time to solve a new item if they know (or think) that they have made an error on the previous item (Rabbitt, 1979). This phenomenon is hard to see from the response vector only but it can be detected using the response times. That is, residual response times will be very large after an error. Detecting post-error slowing might for instance be interesting in educational settings. If a child masters a given subject but fails on the educational test because of post-error slowing (too many items are not reached toward the end of the test), it is valuable for a teacher to be able to see whether this is due to post-error slowing. If the within-respondent response time variability is not consulted, this phenomenon is hard to detect and it may appear that the child is not proficient in the subject while in fact the problem is more anxiety related.

Different use of solution strategies. As discussed above, respondents might differ in the exact solution strategy that they use. However, within respondents, there may also be such a difference. For instance, in a multiplication test, a respondent might answer most of the items from memory using the multiplication tables, but on some items the respondent needs to answer using finger counting as he or she does not know the right answer from memory. This will be reflected in large residual response times for specific items (as finger counting is assumed to take more time than memorizing). This is a within-respondent difference that might be interesting, again, in an educational setting in which the teacher can see from the response times that a child is answering an item using a suboptimal strategy. The raw response times cannot be used for this end as these are conflated by item properties (time intensity of an item) and person properties (the overall speed of the person).

Dynamic assessment. In dynamic assessment (Schneider Lidz, 1987), a respondent first completes part of the test (training phase) after which the respondent receives feedback from the test administrator (e.g., about the correct answers and the correct solution strategy). Then, the respondent completes the remainder of the test without feedback after which it is judged to what degree the respondent benefitted from the feedback. There are many advantages of dynamic assessment including increased external validity (Resing, 2006) and decreased vulnerability to test anxiety (Meijer, 2001). However currently, dynamic assessment is not suitable for large-scale administration as it requires an active role of the administrator for each individual respondent. In addition, the lack of standardization of the feedback given to the respondent is an important matter as differences between respondents in the quality of the feedback might influence the final conclusion about his or her ability. By implementing the residual procedure as described above, the dynamic assessment procedure can be computerized. That is, in the training phase, if the residual response time is large for a given item, a feedback page might be automatically displayed with information about the correct answer and the most optimal solution strategy. This avoids the need of a test administrator for each individual respondent, and it standardizes the feedback to a large degree.

DISCUSSION

A response to a test item can be the outcome of many different response strategies or processes. If the strategies differ in the time they take to implement, their use can be inferred from the response time data. Using experimental control as proposed by Goldhammer (this issue) will likely result in incorrect responses by respondents that use a correct but suboptimal strategy or process, a response that would have been correct if more time were allowed.

This strict control over the testing situation improves the unidimensionality of the test because only the optimal strategies and processes are being measured (respondents using suboptimal strategies or processes will fail). Next, it will restore local independences caused by speeding up and slowing down during the test (as ζ is held constant). Finally, as discussed by Goldhammer (this issue; see also, Wise & DeMars, 2006), the time pressure on the respondents evoked by the experimental control result in more-reliable response times as the effects of confounding factors like test-taking effort are reduced. Therefore, the approach by Goldhammer is very valuable if a single process or strategy needs to be measured in isolation. However, in many other settings of which some have been discussed above, the natural variability in response times can give valuable information for psychological and educational inferences.

FUNDING

The research by Dylan Molenaar was made possible by a grant from the Netherlands Organization for Scientific Research (Grant: VENI-451-15-008).

REFERENCES

Bolsinova, M., & Maris, G. (in press). A test for conditional independence between response time and accuracy. *British Journal of Mathematical and Statistical Psychology*. doi:10.1111/bmsp.12059

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071. doi:10.1037/0033-295X.111.4.1061

Ferrando, P. J., & Lorenzo-Seva, U. (2007). An item response theory model for incorporating response time data in binary personality items. *Applied Psychological Measurement*, *31*, 525–543. doi:10.1177/0146621606295197

Goldhammer, F., & Kroehne, U. (2014). Controlling individuals' time spent on task in speeded performance measures: Experimental time limits, posterior time limits, and response time modeling. *Applied Psychological Measurement*, *38*(4), 255–267. doi:10.1177/0146621613517164

Klein Entink, R. H., Fox, J.-P., & Van der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika*, *74*, 21–48. doi:10.1007/s11336-008-9075-y

Meijer, J. (2001). Learning potential and anxious tendency: Test anxiety as a bias factor in educational testing. *Anxiety, Stress & Coping*, *14*(3), 337–362. doi:10.1080/10615800108248361

Meng, X.-B., Tao, J., & Chang, H.-H. (2015). A conditional joint modeling approach for locally dependent item responses and response times. *Journal of Educational Measurement*, *52*(1), 1–27. doi:10.1111/jedm.2015.52.issue-1

Molenaar, D., Tuerlinckx, F., & Van der Maas, H. L. J. (2015). A bivariate generalized linear item response theory modeling framework to the analysis of responses and response times. *Multivariate Behavioral Research*, *50*(1), 56–74. doi:10.1080/00273171.2014.962684

Rabbitt, P. (1979). How old and young subjects monitor and control responses for accuracy and speed. *British Journal of Psychology*, *70*, 305–311. doi:10.1111/bjop.1979.70.issue-2

Ranger, J., & Kuhn, J.-T. (2014). An accumulator model for responses and response times in tests based on the proportional hazards model. *British Journal of Mathematical and Statistical Psychology*, *67*(3), 388–407. doi:10.1111/bmsp.2014.67.issue-3

Ranger, J., Kuhn, J. T., & Gaviria, J. L. (2015). A Race Model for Responses and Response Times in Tests. *Psychometrika*, *80*, 791–810.

Ranger, J., & Ortner, T. (2012). The case of dependency of responses and response times: A modeling approach based on standard latent trait models. *Psychological Test and Assessment Modeling*, *54*(2), 128.

Resing, W. C. M. (2006). Using children's "ability to learn" in diagnosis and assessment. *Educational & Child Psychology*, *23*, 6–11.

Roskam, E. E. (1987). Toward a psychometric theory of intelligence. In E. E. Roskam & R. Suck (Eds.), *Progress in mathematical psychology* (pp. 151–171). Amsterdam, The Netherlands: North Holland.

Rouder, J. N., Province, J. M., Morey, R. D., Gomez, P., & Heathcote, A. (2015). The lognormal race: A cognitive-process model of choice and latency with desirable psychometric properties. *Psychometrika*, *80*, 491–513. doi:10.1007/s11336-013-9396-3

Schneider Lidz, C. (1987). *Dynamic assessment: An interactional approach to evaluating learning potential*. New York, NY: Guilford Press.

Thissen, D. (1983). Timed testing: An approach using item response testing. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing* (pp. 179–203). New York, NY: Academic Press.

Thorndike, E. L., Bregman, E. O., Cobb, M. V., & Woodyard, E. (1926). *The measurement of intelligence*. New York, NY: Teachers College Bureau of Publications.

Van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*, 287–308. doi:10.1007/s11336-006-1478-z

Van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, *73*(3), 365–384.

Van der Linden, W. J., & Glas, C. A. (2010). Statistical tests of conditional independence between responses and/or response times on test items. *Psychometrika*, *75*(1), 120–139. doi:10.1007/s11336-009-9129-9

Van der Maas, H. L., & Jansen, B. R. (2003). What response times tell of children's behavior on the balance scale task. *Journal of Experimental Child Psychology*, *85*(2), 141–177. doi:10.1016/S0022-0965(03)00058-4

Van der Maas, H. L. J., Molenaar, D., Maris, G., Kievit, R. A., & Borsboom, D. (2011). Cognitive psychology meets psychometric theory: On the relation between process models for decision making and latent variable models for individual differences. *Psychological Review*, *118*, 339–356. doi:10.1037/a0022749

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement*, *43*(1), 19–38. doi:10.1111/jedm.2006.43.issue-1