Goldhammer, Frank; Zehner, Fabian

What to make of and how to interpret process data

Measurement 15 (2017) 3/4, S. 128-132, 10.1080/15366367.2017.1411651

Measurement: Interdisciplinary Research and Perspectives

ISSN: 1536-6367 (Print) 1536-6359 (Online) Journal homepage: https://www.tandfonline.com/loi/hmes20

What to Make Of and How to Interpret Process Data

Frank Goldhammer & Fabian Zehner

MEASUREMENT: INTERDISCIPLINARY RESEARCH AND PERSPECTIVES
2017, VOL. 15, NOS. 3–4, 128–132
https://doi.org/10.1080/15366367.2017.1411651

OPEN ACCESS Check for updates

What to Make Of and How to Interpret Process Data

Frank Goldhammer [a] and Fabian Zehner

[a]German Institute for International Educational Research (DIPF), Centre for International Student Assessment (ZIB) Centre for International Student Assessment (ZIB), Centre for Technology Based Assessment (TBA)

Maddox (2017) argues that respondents' *talk and gesture* during an assessment inform researchers how a response product has evolved. Indeed, how a task is performed represents key information for psychological and educational assessment. In an ancient example: Gideon was required by the Lord to select those men who lap the water with their tongues, but not those who kneel down to drink (Judges 7:5 New International Version).

In cognitive ability testing, process data can be defined as empirical information about the cognitive (as well as meta-cognitive, motivational, and affective) states and related behavior that mediate the effect of the measured construct(s) on the task product (i.e., item score). Thus, operationally, process data can be regarded as the empirical data reflecting the course of working on a test item. Recently, process data has increasingly gained attention in cognitive ability testing given the digitalization of measurement and the possibility of exploiting log file data. Other sources of process data are, for instance, concurrent think aloud protocols, screen capturing, eye tracking, facial expression, video-recorded behavior, and physiological sensor data (Azevedo, 2015).

As shown by Maddox for large-scale assessments, even talk and gesture can be regarded as useful process data. In this case, the process data is not only video-recorded but also observed by the interviewer in situ; the interviewer interactively uses it to influence the test-taking process and to reduce construct-irrelevant variance. Thus, like product data (e.g., scores), process data is used to draw inferences. We argue in the following that the interpretation and use of process data and derived indicators require validation, just as product data do (Kane, 2013). This theoretical background, including some examples about log file data, sets the ground for our comments on Maddox's use of "talk and gesture as process data."

Using and interpreting process data

Process data may be used to address substantial research questions, for instance, to learn about individual solution behavior and related cognitive processes (Greiff, Niepel, Scherer, & Martin, 2016; Molenaar, 2015). Measurement can be enhanced by using process data as evidence for process-oriented constructs, such as speed (Klein Entink, Fox, & Van Der Linden, 2009), for the detection of disengaged or aberrant response behavior (Kong, Wise, & Bhola, 2007; Van Der Linden & Guo, 2008), or for the validation of score interpretations (Borsboom, Mellenbergh, & Van Heerden, 2004; Kane & Mislevy, 2017).

Maddox considers talk and gesture as process data for formative use in the PIAAC interview situation. This is an interesting and novel extension in that the interviewer observes the respondent taking the test (e.g., verbal utterances), makes inferences based on these observations (e.g., feelings of failure), and derives adaptive interventions (e.g., face-saving support). This way, the interviewer contributes to reducing the confounding construct-irrelevant sources of variance for the test score. This is not in contradiction with standardized testing; rather, the interviewer represents a communicative and adaptive component of the measurement (Suchman & Jordan, 1990). However, this approach gives rise to the crucial question of whether interviewers are prepared to draw these

CONTACT Frank Goldhammer goldhammer@dipf.de German Institute for International Educational Research (DIPF), Centre for International Student Assessment (ZIB), Centre for Technology Based Assessment (TBA)

inferences in a valid and consistent way, particularly in cross-cultural assessments. If interviewers interact differently in the same situations, interviewer effects will contaminate the measure. For example, as shown by Ackermann-Piek and Massing (2014), PIAAC interviewers in Germany behaved differently than expected in many respects (e.g., how additional information was provided to the respondent on request). Furthermore, does Maddox's study imply that group assessments (e.g., PISA) are not trustworthy given the lack of individual supervision? It seems that high data quality can be obtained from computer-based assessments by a well-developed–user interface tailored to the respondents' needs. In this respect, Maddox gives some hints for potential improvement in the case of PIAAC, which we take up in our outlook.

Eliciting process data and synthesizing it to process indicators

Process data is quite often regarded as collateral information. However, for being able to make the desired inferences, researchers need to plan thoroughly what kind of process data should be collected and what the temporal and spatial resolution should be (Kroehne, Roelke, Kuger, Goldhammer, & Klieme, 2016; Oranje, Gorin, Jia, & Kerr, 2017). Following the evidence-centered design approach (Mislevy, Almond, & Lukas, 2003), the intended inferences primarily determine the empirical evidence to be elicited. Thus, depending on the claims about the latent process (e.g., strategy use, test-taking engagement), observable evidence for the targeted construct needs to be identified (e.g., selection, sequence, and duration of behavioral steps), and finally, situations and tasks that evoke the desired behavior need to be designed.

Table 1 shows two examples for the chain of inferences: from process data to behavior during the assessment to the latent (e.g., cognitive or motivational) process. For Example 1, the task design would only require observing response times by item. A suitable task design for Example 2 would be a web environment including tools for highlighting and annotating text. Thus, it is necessary to design a task that elicits the desired behavior during task completion and to design a system that records all required events, including time information.

By combining process data, process indicators are derived that reflect the behavior during the assessment with respect to the targeted latent cognitive process. In Example 1, this indicator could represent whether the response time was below a threshold; in Example 2, this could be a measure of semantic transitions between different web pages.

The process data in the PIAAC interview, as described by Maddox, is entirely different. There are no mechanisms to systematically evoke talk or gesture. Instead, the interviewer serves as support and primarily responds to body language and verbal utterances to address the respondent's questions and problems. Here, process data *is* collateral information; that is, in a given situation (e.g., the respondent did not fully get the instruction), process data may or may not be available for adapting the interview process, depending on respondent characteristics among others (e.g., extraversion). Thus, one needs to design the interview in a way that respondents communicate their questions and problems related to the assessment comparably. In other words, the observed behavior of respondents is comparable if, and only if, it was evoked in the same manner.

Table 1. Illustration of the chain of inferences from log file data to a latent process.

	Level of Inference		
	(1) Latent Process	(2) Empirical Behavior	(3) Process Data
Example A *Rushing Through the Test*	Test-taking disengagement	Responding to an item quickly (i.e., below response time threshold)	Log file data (i.e., click events of next-item button and time stamps)
Example B *Thoroughly Comparing Information*	Contrasting information in reading multiple documents	Switching back and forth and (re)reading related information presented across web sites	Log file data (i.e., navigation events, highlighted text, or typed annotations)

Validating the interpretation of process indicators

Inferring latent (e.g., cognitive) processes from process indicators needs to be justifiable. Similarly to the validation of test-score interpretation (AERA, APA, NCME, & Joint Committee on Standards for Educational Psychological Testing, 2014; Kane, 2013), both theoretical and empirical evidence is needed to make the link from the process indicator to the latent process well-founded. Regarding Example 1, one may wonder whether a response time falling below a certain threshold would allow one to conclude that the respondent shows no test-taking engagement.

How can the intended interpretation of a process indicator be challenged? This requires the falsification of hypotheses as proposed, for instance, by theories of cognitive processing. Failed falsifications would provide evidence for the intended interpretation. Evidence for supporting a claim such as "indicator Y measures the latent process X" can be collected using correlational and experimental approaches. Thus, the claim may be justifiable if Y is correlated with Z (e.g., an external criterion obtained from video data) or if Y is affected by an experimental manipulation as expected.

For instance, one can assume the rate of correct responses to be around chance level for item completions that are classified as disengaged (Goldhammer, Martens, & Lüdtke, 2017; Lee & Jia, 2014). An experimental strategy could be based on the assumption that high-stakes testing causes a lower rate of disengaged completions than a low-stakes condition. The validity of the interpretation of the indicator would be supported if the accuracy of disengaged completions were only at chance level and if the indicator were able to reveal the assumed difference between testing conditions.

In principal, the talk and gesture used as process data by Maddox also face the problem of a clear interpretation. Thus, evidence is needed that interviewers are able to draw the correct inference about latent processes, such as fatigue. However, talk is certainly less ambiguous information than the response time or navigation behavior that is extracted from log file data. From this perspective, talk and gesture could serve as external criteria to validate, for instance, the interpretation of response time as an indicator of engagement. This would be a fruitful combination of small-scale qualitative and large-scale quantitative studies.

Looking ahead: Suggestions for the large scale

While Maddox demonstrates the usefulness of talk and gesture for small-scale studies, consideration needs to be given on how this information could be harvested on a large scale. When Maddox asks, "Can a computer do that?" (p. 14), we are bold and answer, "Yes, to some extent." If the task is to react appropriately and in a standardized manner to the verbal utterance and facial expression of a respondent, modern technologies could do the job.

Several indicators could be engineered from video and audio streams. Audio recordings would be transformed into text by automatic speech recognition. This is challenging, but most problems induced by the setting—such as multi-party dialogue or noise—are or are about to be solved (Huang, Baker, & Reddy, 2014; Traum & Rickel, 2002). With talk transcribed to text, natural language processing techniques, centering around discourse processes, could be utilized. For example, the semantics of a verbal exchange could be classified by latent semantic analysis (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990), or response cries in combination with a facial expression would reveal a lot about the respondent's affective state.

With the video information, facial expressions of the test person could be automatically classified by the Facial Action Coding System (Ekman & Rosenberg, 1997), similarly to how the system is used in the intelligent tutoring system AutoTutor (D'Mello & Graesser, 2013). This way, frustration, engagement, or fatigue could be identified (Grafsgaard, Wiggins, Boyer, Wiebe, & Lester, 2013; Gu & Ji, 2004).

With smart ideas on how to exploit the additional information, large-scale assessment could benefit significantly from the approach proposed by Maddox. Coming back to this section's opening, virtual embodied conversational agents acting as test administrators could draw on the features described here and, thus, appropriately adapt the testing situation in a manner that is still standardized. Obviously, the scenario described here involves an expensive development, and its limitations and implications for intervening with the assessment would need to be investigated. However, the attempt seems worth the effort.

ORCID

Frank Goldhammer http://orcid.org/0000-0003-0289-9534
Fabian Zehner http://orcid.org/0000-0003-3512-1403

References

Ackermann-Piek, D., & Massing, N. (2014). Interviewer behavior and interviewer characteristics in PIAAC Germany. *Methods, Data, Analyses*, *8*(2), 199–222. doi:10.12758/mda.2014.008

AERA, APA, NCME, & Joint Committee on Standards for Educational Psychological Testing. (2014). *Standards for educational and psychological testing. Washington, DC: American Educational Research Association.*

Azevedo, R. (2015). Defining and measuring engagement and learning in science: Conceptual, theoretical, methodological, and analytical issues. *Educational Psychologist*, *50*(1), 84–94. doi:10.1080/00461520.2015.1004069

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

D'Mello, S., & Graesser, A. (2013). AutoTutor and Affective AutoTutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. *ACM Transactions Interact Intelligent Systems*, *2*(4), 23:21–23:39. doi:10.1145/2395123.2395128

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391. doi:10.1002/(sici)1097-4571(199009)41:6\391::aid-asi1.3.0.co

Ekman, P., & Rosenberg, E. L. (Eds.). (1997). *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. New York, NY: Oxford University Press.

Goldhammer, F., Martens, T., & Lüdtke, O. (2017). Conditioning factors of test-taking engagement in PIAAC: An exploratory IRT modelling approach considering person and item characteristics. *Large-Scale Assessments in Education*, *5*(1), 18. doi:10.1186/s40536-017-0051-9

Grafsgaard, J., Wiggins, J. B., Boyer, K. E., Wiebe, E. N., & Lester, J. (2013). Automatically recognizing facial expression: Predicting engagement and frustration. In *Proceedings of the Sixth International Conference on Educational Data Mining*. Memphis,: TN: International Educational Data Mining Society.

Greiff, S., Niepel, C., Scherer, R., & Martin, R. (2016). Understanding students' performance in a computer-based assessment of complex problem solving: An analysis of behavioral data from computer-generated log files. *Computers in Human Behavior*, *61*(Suppl. C), 36–46. Seoul: IEEE. doi:10.1016/j.chb.2016.02.095

Gu, H., & Ji, Q. (2004). An automated face reader for fatigue detection. *Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition* (pp. 111–116). doi:10.1109/afgr.2004.1301517

Huang, X., Baker, J., & Reddy, R. (2014). A historical perspective of speech recognition. *Communications of the ACM*, *57*(1), 94–103. doi:10.1145/2500887

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. doi:10.1111/jedm.12000

Kane, M. T., & Mislevy, R. (2017). Validating score interpretations based on response processes. In K. Ercikan & J. W. Pellegrino (Eds.), *Validation of score meaning for the next generation of assessments* (pp. 11–24). New York, NY: Routledge.

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*(1), 21–48. doi:10.1007/s11336-008-9075-y

Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Educational and Psychological Measurement*, *67*(4), 606–619. doi:10.1177/0013164406294779

Kroehne, U., Roelke, H., Kuger, S., Goldhammer, F., & Klieme, E. (2016). *Theoretical framework for log-data in technology-based assessments with empirical applications from PISA*. Paper presented at the NCME 2017 Annual Meeting, San Antonio, TX.

Lee, Y.-H., & Jia, Y. (2014). Using response time to investigate students' test-taking behaviors in a NAEP computer-based study. *Large-Scale Assessments in Education*, *2*(1), 1–24. doi:10.1186/s40536-014-0008-1

Maddox, B. (2017). Talk and gesture as process data. *Measurement: Interdisciplinary Research and Perspectives*, *15* (3&4).

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series*, *2003*(1), i–29. doi:10.1002/j.2333-8504.2003.tb01908.x

Molenaar, D. (2015). The value of response times in item response modeling. *Measurement: Interdisciplinary Research and Perspectives*, *13*(3–4), 177–181. doi:10.1080/15366367.2015.1105073

Oranje, A., Gorin, J., Jia, Y., & Kerr, D. (2017). Collecting, analysing, and interpreting response time, eye tracking and log data. In K. Ercikan, & J. W. Pellegrino (Eds.), *Validation of score meaning for the next generation of assessments* (pp. 39–51). New York, NY Routledge.

Suchman, L., & Jordan, B. (1990). Interactional troubles in face-to-face survey interviews. *Journal of the American Statistical Association*, *85*(409), 232–241. doi:10.1080/01621459.1990.10475331

Traum, D., & Rickel, J. (2002). Embodied agents for multi-party dialogue in immersive virtual worlds. In *Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems: Part 2* (Vol. 2, pp. 766–773) Bologna: ACM.

Van Der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, *73*(3), 365–384. doi:10.1007/s11336-007-9046-8