Psychological Test and Assessment Modeling, Volume 55, 2013 (1), 3-38

A systematic review of the methodology for person fit research in Item Response Theory: Lessons about generalizability of inferences from the design of simulation studies

André A. Rupp[1]

Abstract

This paper is a systematic review of the methodology for person fit research targeted specifically at methodologists in training. I analyze the ways in which researchers in the area of person fit have conducted simulation studies for parametric and nonparametric unidimensional IRT models since the seminal review paper by Meijer and Sijtsma (2001). I specifically review how researchers have operationalized different types of aberrant responding for particular testing conditions in order to compare these simulation design characteristics with features of the real-life testing situations for which person fit analyses are officially reported. I discuss the alignment between the theoretical and practical work and the implications for future simulation work and guidelines for best practice.

Key words: Person fit, systematic review, aberrant responding, item response theory, simulation study, generalizability, experimental design.

[1] *Correspondence concerning this article should be addressed to:* André A. Rupp, Associate Professor, HDQM Department, EDMS Program, University of Maryland, 1230-A Benjamin Building, College Park, MD 20742, USA; email: ruppandr@umd.edu

This paper is situated in the conceptual space of research on *person fit*, which is one aspect of the comprehensive enterprise of critiquing the alignment of the structure of a particular statistical model with a particular data set using residual-based statistics (Engelhard Jr., 2009). I first analyze the ways in which researchers in the area of person fit have conducted simulation studies in *non-parametric* (e.g., Sijtsma & Molenaar, 2002; van der Aark, Hemker, & Sijtsma, 2002) and *parametric unidimensional item response theory* (IRT) (e.g., DeAyala, 2009; Yen & Fitzpatrick, 2006) since the seminal review paper by Meijer and Sijtsma (2001). I then discuss the alignment between the theoretical and practical work and the implications for future simulation work and guidelines for best practice.

This paper is primarily intended for methodologists in training but should also prove useful for practitioners who are curious about the statistical foundations for proposed guidelines of best practice. The information in this paper may be of less interest for the relatively few specialists who are already conducting advanced simulation studies in this area. However, it should provide some useful insight into the ways these researchers conduct their work for the many other researchers and practitioners who want to be critical consumers of this work.

Simulation studies are designed statistical experiments that can provide reliable scientific evidence about the performance of statistical methods. As noted concisely by Cook and Teo (2011):

> *In evaluating methodologies, simulation studies: (i) provide a cost-effective way to quantify potential performance for a large range of scenarios, spanning different combinations of sample sizes and underlying parameters, (ii) allow average performance to be estimated under repeat Monte Carlo sampling and (iii) facilitate comparison of estimates against the "true" system underlying the simulations, none of which is really achievable via genuine applications, as gratifying as those are. (p. 1)*

In the context of person fit research, simulation studies are most commonly used to quantify the frequency of type-I and type-II errors and associated power rates under a variety of test design and model misspecification conditions.

Researchers who publish in this area clearly make some concerted and thoughtful efforts to summarize findings from simulation studies, especially when they are trying to situate their particular theoretical work within a relevant part of the literature. Thus, I initially started out writing this paper as a more "traditional" review paper that focused on what researchers had learned about person fit in roughly the last 10 years. However, while reviewing the recent body of work it became quickly clear that there is perhaps a more urgent need to discuss the methodology of simulation research with more scrutiny in order to help methodologists in training understand the kinds of generalizations that can and cannot be made based on this work.

Thus, in this paper I specifically focus on how researchers have operationalized different types of aberrant responding for particular testing conditions. I then compare these simulation design characteristics with features of the real-life testing situations for which person fit analyses are officially reported. I decided to focus on published reports of applied person fit work because those sources are accessible to members of the target audience working within a broad range of educational and psychological measurement contexts. Testing companies may perform similar analyses in-house within an operational testing program, of course, but if such work remains publicly undocumented any resulting insights are effectively shared with only a privileged few.

In particular, the research questions that guided my review work were the following:

1. What kinds of real-life aberrant behavior have researchers decided to operationalize in their simulation studies and how did they go about operationalizing them?
2. For what kinds of test designs have they decided to investigate the behavior of person fit statistics?
3. What kinds of assessments and model-fit assessment strategies are used by researchers who utilize person fit statistics in their applied work?
4. What is the alignment between the simulation study designs and results and the data-collection designs and fit assessment strategies that practitioners use? What are the implications of this (mis) alignment for future simulation work and practice?

While answering these research questions I specifically discuss what the implications of the simulation design choices are for the generalizability of inferences that can be drawn from these studies.

Rhetorically, I argue that the application of person fit strategies in real-data contexts could be accelerated if practitioners and applied researchers were given a more accessible pathway into the methodological literature. Despite laudable efforts by some researchers, there are many instances where variations and nuances in the operationalizations of person misfit and associated test design conditions across simulation studies could be conveyed with more clarity. This would allow readers of this work to compare more easily the causal relationship between induced aberrancy effects and the behavior of person fit statistics across simulation studies.

To build this argument on the basis of the research questions above, I have organized this paper into three main sections. In the next section I present a brief review of the key ideas around model-data fit assessment in IRT and the methodology that I used to locate relevant sources. In the following three sections I address each of the first three research questions in turn. I close this paper with a discussion that addresses the fourth research question and associated observations.

Model-data fit assessment

The criticism of a particular statistical model and its subsequent refinement are key steps in any modeling endeavor to ensure trustworthy parameter inference from a sample of data to the population from which the data were sampled (e.g., Levy, 2011; Levy, Mislevy, & Sinharay, 2009; Sinharay, 2005; Sinharay, Johnson, & Stern, 2006). For real-data analyses a comprehensive set of model criticism and refinement strategies that cover a range of potential threats to model-data fit should always be used, especially if resulting interpretations from estimated parameters are relatively high stakes. Conceptually, in IRT this assessment can take place at four different key levels, (1) relative model-data fit, (2) absolute model-data fit, (3) item-level model-data fit, and (4) person-level model-data fit.

Put simply, *relative model-data fit* is about determining which of a subset of candidate models appears to fit the data structure best without fully determining whether that model actually fits the data well in and of itself. This is typically accomplished with variants of likelihood-ratio test statistics for nested models or information criteria for non-nested models (e.g., Baker & Kim, 2004; Li, Cohen, Kim, & Cho, 2009; Li & Rupp, 2011). *Absolute model-data fit* is about determining whether a chosen model fits well overall; both relative and absolute model-data fit assessment can be viewed as *global assessments* of the suitability of a particular statistical model for the data structure at hand.

In contrast, both item- and person-level model-data fit assessment are concerned with *local assessments* of the suitability of a particular statistical model. *Item-level model-data fit* – or *item fit* in short – focuses on whether the data vector/response string for individual test items appears to be consistent with the remaining data structure for a particular statistical model of interest. *Person-level model-data fit* – or *person fit* in short – focuses on whether the data vector/response string for individual respondents/subjects/persons appears to be consistent with the remaining data structure for a particular statistical model of interest. For either type of *local assessment* of model-data (mis)fit unusual items or persons are subject to further scrutiny and potential removal for subsequent re-estimations of the remaining model parameters.

One important practical question that arises in person fit is similar to the issue of scale purification in analyses of differential item functioning (e.g., Ferne & Rupp, 2007), namely what one should do with persons that are flagged as aberrant responders. While items can be revised for subsequent trials, this can of course not be done with persons so that the only realistic actionable option for misfitting persons appears to be to exclude these persons from re-calibration runs. However, researchers seem to find only moderate success with this strategy and only in cases where the sample size is large, the proportion of aberrant responders is large, and the misfit is severe (i.e., when the model-data fit is very poor; see, e.g., Ferrando, 2007; Meijer, 1997).

The global vs. local distinction is sometimes used to differentiate further how some person fit statistics work. Some researchers distinguish between person fit statistics that are designed to detect unusual response patterns across *all items* in a score vector – which they call statistics for "global" person-fit assessment – and person fit statistics that are designed to detect unusual response patterns across *subsets of items* in a score vector – which they call statistics for "local" person fit assessment (e.g., Emons, 2009). Put simply, the different distinctions surrounding global and local fit assessment remind us that each fit statistic is only suitable for critiquing a particular aspect of model-data fit.

Many publications of person fit are situated in the area of educational achievement testing (e.g., Brown & Villareal, 2007; Engelhard Jr., 2009; see Table 2 later in the paper), which is unsurprising given that IRT models are predominantly used in these contexts. However, in the last 10 years some notable applications of person fit analyses have appeared in other assessment areas such as psychological assessment (e.g., Meijer, Egberink, Emons, & Sijtsma, 2008), personality assessment (e.g., Dodeen & Darabi, 2009; Ferrando, 2004, 2009, 2012; Woods, Oltmanns, & Turkheimer, 2008), attitudinal assessment (e.g., Curtis, 2004), and health outcomes assessment (e.g., Custers, Hoijtink, van der Net, & Helders, 2000; Tang et al., 2010).

Thus, it is probably fair to say that person fit research has gained prominence across assessment fields in the last 10 years. Nevertheless, person fit assessment seems to remain somewhat of a distant cousin of other forms of fit assessment, especially item fit assessment. Apart from short illustrative applications in simulation studies, there is a relative paucity of publicly available sources that comprehensively describe how to detect, explain, and rectify person misfit. Moreover, patterns of person misfit are rarely reported in official reports for large-scale educational surveys such as the *Programme for International Student Assessment* (PISA), the *Trends in Mathematics and Science Study* (TIMSS), and the *National Assessment of Educational Progress* (NAEP) despite a few early reports on this issue (e.g., Rudner, Scagg, Bracey, & Getson, 1995).

As noted in the introduction, many research examples in this paper come from nonparametric and parametric IRT because it is a very powerful general latent-variable modeling framework for analyzing discrete response data from various types of assessment instruments. Consequently, IRT-based research on person fit is arguably more advanced relative to other areas even though methodological research on person-fit is slowly becoming more prominent in related latent-modeling frameworks as well. As a result, my systematic review also includes relevant recent sources from the area of *latent class and mixture analysis* (e.g., Emons, Glas, Meijer, & Sijtsma, 2003; von Davier & Molenaar, 2003), *diagnostic classification modeling* (e.g., Cui & Leighton, 2009; Liu, Douglas, & Henson, 2009; see also Tatsuoka, 2009; Rupp, Templin, & Henson, 2010), *multilevel logistic regression analysis* (e.g., Reise, 2000; Woods, 2008), and, especially, *factor analysis/ covariance structure modeling* (e.g., Clark, 2010; Ferrando, 2007, 2009) as this framework is closely related to IRT (see, e.g., McDonald, 1999; Meade & Lautenschlager, 2004; Thissen & Wainer, 2001).

In order to locate relevant simulation studies for this paper I searched references from 2000 to 2010, which seemed like a suitable window given that Meijer and Sijtsma (2001) had done an excellent job at summarizing the research up until that point. I located sources primarily by using *PsychInfo*, dissertation databases, *Google*, and related search environments, as well as by tracing references within the sources located. I also consulted conference programs from meetings of the *National Council on Measurement in Education* (NCME), the *American Educational Research Association* (AERA), the *International Meeting of the Psychometric Society* (IMPS), and the *European Association of Methodology* (EAM). I also reviewed technical reports of core educational surveys such as PISA, TIMSS, and NAEP.

I used keywords such as “person fit”, “fit”, “model-data fit”, “model fit”, “aberrant response”, “cheating”, “guessing”, and related words in conjunction with words like “simulation", "application", "item response theory", "theory", "nonparametric", and so on. I did not include the two dissertations by Deng (2007) and Shin (2007) in my summary tables because both resulted in peer-reviewed publications. I similarly did not include conference papers if they were later augmented and made into peer-reviewed publications that I had already included in my set of sources.

I did not perform a detailed analysis of simulation studies in the areas of *computerized adaptive testing* (e.g., Hui, 2008; McLeod, Lewis, & Thissen, 2003; van Krimpen-Stoop & Meijer, 2001, 2002; see also Meijer, 2002) and *multi-level regression analysis* (e.g., LaHuis & Copeland, 2007; Reise, 1999; Woods et al., 2008). For both of these areas the simulation methods and analysis approaches differ qualitatively from those in the core sources for this paper and are probably best summarized in separate publications situated within these specific areas.

Overall, a total of 44 peer-reviewed sources fit the search frame that I had defined for the purpose of this paper; all of these sources are included with an asterisk (*) in the reference section at the end of this paper. One notable publication in Chinese (Liu, Cao, & Dai, 2011) was not included due to its unavailability in English.

Due to the diversity of ways in which numerical information was presented in my sources, I chose means of summarizing information that allowed me to maximize the direct comparability of design features and findings across studies. For example, I commonly chose the range statistic for summarizing parameter values used in data generation as some authors provided all explicit values, some provided only distributional specifications, and others only summary statistics such as means and standard deviations. When authors included different ranges in different conditions, I selected the smallest lower bound and the largest upper bound across conditions for reporting.

Research question 1: Generalizability considerations based on the simulation of aberrant responses

As noted before, findings from simulation studies should be used to create guidelines for practitioners. Thus, it is arguably important that the aberrant behaviors that are being simulated in these studies have a reasonably close relationship to what real persons are likely doing in real assessment contexts when they are responding aberrantly. At the same time, the creation of a mechanism for aberrancy in the context of a simulation study will always represent some form of abstraction of the complexity of real-life behavior that can never do it fully justice.

The basis of the following discussions is the information contained in Tables 1 and 2, which shows the ways in which the authors of my sources have operationalized different kinds of aberrant response behavior; note that X^* in Table 1 denotes the replaced response while X denotes the generated response. The 25 sources in Tables 1 and 2 are listed alphabetically by the name of the first author with the year of publication and with separate rows for separate studies within the same paper if necessary.

Labels for aberrant responding

Authors used a rather wide array of labels for the real-life response mechanism that they intended to mimic, which is likely reflective of the fact that one might expect a range of different aberrant response behaviors across testing situations. Labels included, among others, “guessing”, “answer copying”, “cheating”, “special expertise in a subarea of the test”, “fatigued responding”, “careless/inattentive responding”, “creative responding”, “random/haphazard responding”, “speeded responding”, “unfamiliarity with the test format”, “misunderstanding of instructions”, “displaying test anxiety”, “displaying high or low motivation”, "socially desirable responding", "malingering", "aberrant responding due to pathology", “ignoring reverse scoring of options”, and “displaying a tendency to choose extreme response options”.

There is a subtle semantic nuance present in these labels, which is that some labels denote the ways in which *persons* respond aberrantly such as when they “tend to choose extreme response options”, are “responding randomly”, or are “guessing”. Other labels, however, explicitly reflect *postulated causes* for a certain type of aberrant behavior such as when persons respond aberrantly because they are “unfamiliar with the test format”, “misunderstand instructions”, or have a certain “clinical pathology”. This distinction is important for understanding operationalization processes because the labels that suggest causes may lead to the same kind of behavior that is suggested by the labels that reflect ways of responding.

For example, a person who is “unfamiliar with the test format” may make unexpected mistakes early on in the test (e.g., may provide incorrect answers to items they are expected to get correct), which is the same surface-level behavior that is displayed by a person who may display “test anxiety” or “carelessness” in responding. Thus, if a further differentiation between these two types of behavior were desired for a given real-life assessment contexts, simulation studies would have to mimic this situation by inducing differential response propensities conditional on values of covariates for persons during data-generation for example.

Labels can also have domain-specific meanings. For example, as Emons (2009) reminds us, the label “careless responding” can be meaningfully used to characterize the situation where respondents on an instrument for attitudinal or personality assessment fail to recognize that some items are reversely worded and respond to them as if they were not. Similarly, the label “socially desirable responding” is probably only meaningful for attitudinal assessments and related surveys, but is of limited use for speeded intelligence tests or other types of educational achievement tests.

However, despite the relatively large array of labels for aberrant responding, there are really only two types of statistical score effects that are effectively created, which are (1) *spuriously low scores* (i.e., when persons provide a lower score than would be expected based on the chosen model) and (2) *spuriously high scores* (i.e., when persons provide a higher score than would be expected based on the chosen model); if both types of aberrant responding are present for a given assessment this leads to what one might call

Table 1: Overview of operationalizations for aberrant response behavior

Source	Person Characteristics					Aberrant Response	Item Characteristics				Original Label
	Total	# affected	% affected	Ability	Range		Total	# affected	% affected	Kind of item	
Armstrong (2007)	10,000	5,000	50	Low	< -.5	X* = 1	121	18, 24, 36	15, 20, 30	Any	Spuriously high
Armstrong (2007)	10,000	5,000	50	High	> .5	P(X* = 1) = .20	121	18, 24, 36	15, 20, 30	Any	Spuriously low
Armstrong (2009a)	10,000	100, 300	1, 3	Low	< .5	X* = 1	100	8, 10, 12	8, 10, 12	Any	Aberrant
Armstrong (2009a)	10,000	100, 300	1, 3	High	> .5	X* = 0	100	8, 10, 12	8, 10, 12	Any	Aberrant
Armstrong (2009b)	10,000	100, 300	1, 3	Low	< -.5	X* = 1	100	8, 10, 12	8, 10, 12	Any	Spuriously high
Armstrong (2009b)	10,000	100, 300	1, 3	High	> .5	X* = 0	100	8, 10, 12	8, 10, 12	Any	Spuriously low
Armstrong (2009b)	10,000	100, 300	1, 3	Middle	[-.5, .5]	X* = 1	100	8, 10, 12	8, 10, 12	Any	Spuriously mixed
Choi (2008)	1,000	50, 100, 200	5, 10, 20	N/P	N/P	N/P	35	7	20	Any	Guessing
Clark (2010)	1,000	10, 50, 100, 250	1, 5, 10, 25	All	N/A	P(X* = X_{max}) = .80	25	3, 7, 13	10, 30, 50	Any	Cheating
Cui (2009)	4,000	2,000	50	High	N/A	X* = 0	14, 28, 42	N/P	N/P	Easy items	Creative
Cui (2009)	4,000	2,000	50	All	N/A	Variome (model misspecification)	14, 28, 42	14, 28, 42	100	Any	Structural Misspecification
Cui (2009)	4,000	2,000	50	All	N/A	P(X* = 1) = .25	14, 28, 42	14, 28, 42	100	Any	Random
De la Torre (2008)	5,000	5,000	100	All	(-2.5, 2.5)	X* = 1	10, 30, 50	1, 3, 5, 9, 15	10, 30	Difficult items	Cheating
De la Torre (2008)	5,000	5,000	100	All	[-2.5, 2.5]	P(X* = 1) = .25	10, 30, 50	1, 3, 5, 9, 15	10, 30	(Later) difficult items	Speeded
De la Torre (2008)	5,000	5,000	100	All	[-2.5, 2.5]	P(X* = 1) = .25	10, 30, 50	1, 3, 5, 9, 15	10, 30	(Early) easy items	Lack of motivation
Dimitrov (2006)	9,000	2,430	27	Low	< -.61	P(X* = 1) = .25	10, 20, 30	2, 4, 6, 8, 12	20, 40	Difficult items	Guessing
Dimitrov (2006)	9,000	2,430	27	Low	< -.61	P(X* = 1) = .90	10, 20, 30	2, 4, 6, 8, 12	20, 40	Difficult items	Cheating
Emons (2003)	1,000	1,000	100	All	N(0, 1.66)	X* = 1	20, 40	5, 8, 10	12, 20, 25, 50	Difficult items	Cheating
Emons (2003)	1,000	1,000	100	All	N(0, 1.66)	P(X* = 1) = .25	20, 40	5, 8, 10	12, 20, 25, 50	Easy items	Inattentive
Emons (2004)	1,000	N/P	N/P	N/P	N/P	X* = 1	20, 40	5, 8,10	12.5, 25, 40	Difficult items	Answer copying
Emons (2004)	1,000	N/P	N/P	N/P	N/P	P(X* = 1) = .25	20, 40	5, 8,10	12.5, 25, 40	Easy items	Test anxiety
Emons (2008)	6,000	3,000	50	All	N/P	P(X*= x*) = 1/M	12, 24	6, 12, 18, 24	25, 50, 75, 100	Any	Careless
Emons (2008)	6,000	3,000	50	All	N/P	High P for extreme scores	12, 24	6, 12, 18, 24	25, 50, 75, 100	Any	Extreme options
Emons (2008)	6,000	3,000	50	All	N/P	X* = M – X	12, 24	6, 12, 18, 24	25, 50, 75, 100	Any	Reverse wording
Emons (2009)	1000, 3000	50, 100, 150, 300	5, 10	N/P	N/P	X* = M – X	12, 24	6, 12	25, 50	Any	Careless
Emons (2009)	1000, 3000	50, 100, 150, 300	5, 10	N/P	N/P	Recode almost extreme into extreme scores	12, 24	12, 24	100	Any	Extreme options

Emons (2009)	1000, 3000	50, 100, 150, 300	5, 10	N/P	N/P	Replace X with score of a differentially able person	12, 24	4, 6, 8	17, 25, 33, 50, 67	Any	Atypical content
Ferrando (2009)	5,000	300	6	N/P	N/P	N/P	10, 18, 24	2, 4, 6	20, 25	N/P	Random
Ferrando (2010)	500	30	6	N/P	N/P	P(X* = 1) =.50	10, 30	2, 6	20	Any	Random
Glas (2003)	400, 1,000	40, 100	10	N/P	N/P	P(X*= x) > P(X = x)	30, 60	5, 10, 15, 20, 30	17, 33, 50	Any items	Local dependence
Glas (2003)	400, 1,000	40, 100	10	N/P	N/P	P(X* = 1) = .20	30, 60	5, 10, 15, 20, 30	17, 33, 50	Easy items	Guessing
Glas (2007)	400, 1,000	40, 100	10	N/P	N/P	P(X* = 1) > P(X = 1)	40, 60	10, 15, 20, 30	25, 50	Any	Ability increase
Glas (2007)	400, 1,000	40, 100	10	N/P	N/P	P(X* = 1) =.20	40, 60	10, 15, 20, 30	25, 50	Any	Guessing
Hendrawan (2005)	400, 1000	40, 100	10	N/P	N/P	P(X* = 1) = .80	30, 60	5, 10, 15, 20, 30	17, 33, 50	Difficult items	Item disclosure
Hendrawan (2005)	400, 1000	40, 100	10	N/P	N/P	P(X* = 1) = .20	30, 60	5, 10, 15, 20, 30	17, 33, 50	Easy items	Guessing
Karabatsos (2003)	500	25, 50, 125, 250	5, 10, 25, 50	Low	[-2, -.5]	X* = 1	17, 33, 65	3, 6, 12	18	Difficult items	Cheating
Karabatsos (2003)	500	25, 50, 125, 250	5, 10, 25, 50	Low	[-2, -.5]	P(X* = 1) = .25	17, 33, 65	7, 14, 27	41	Difficult items	Guessing
Karabatsos (2003)	500	25, 50, 125, 250	5, 10, 25, 50	High	[.5, 2]	P(X* = 1) = .50	17, 33, 65	7, 14, 27	41	Easy items	Careless
Karabatsos (2003)	500	25, 50, 125, 250	5, 10, 25, 50	High	[.5, 2]	X* = 0	17, 33, 65	3, 6, 12	18	Easy items	Creative
Karabatsos (2003)	500	25, 50, 125, 250	5, 10, 25, 50	All	[-2, 2]	P(X* = 1) =.25	17, 33, 65	17, 33, 65	100	Any	Random
Liu (2009)	2,200	400	18	N/P	N/A	P(X* = 1) > P(X = 1)	60, 90	60, 90	100	Any	Spuriously high
Liu (2009)	1,000	1,000	100	All	N/A	P(X* = 1) > P(X = 1)	60, 90	60, 90	100	Any	Strategy switching
Liu (2009)	1,000	N/P	N/P	N/P	N/P	P(X*= 1) < P(X = 1)	90	N/P	N/P	N/P	Spuriously low
Raîche (2003)	1,000	1,000	100	All	[-2, 2]	P(X*=1)=1–P(X =1)	N/A	N/A	10, 20	Any	Incorrect
Raîche (2003)	1,000	1,000	100	All	[-2, 2]	P(X* = 1) = 1/item#	Various	Various	10, 20	Any	Random
Sijtsma (2001)	3,000	3,000	100	All	[-2, 2]	P(X* = 1) = .25	40, 80	8, 10	12.5	Easy items	Careless
St-Onge (2009)	100, 1,000	5, 50	5	N/P	N/P	X* = 1	40	8	20	Any	Spuriously high
St-Onge (2011)	1,000	50	5	Low	< 0	X* = 1	20, 40, 60, 80	2-48	10-60	All [Random]	Spuriously high
St-Onge (2011)	1,000	50	5	High	> 0	P(X* = 1) = .25	20, 40, 60, 80	2-48	10-60	All [Random]	Spuriously low
Wang (2008)	6,000	1,080	18	Low	[-2, -.5]	P(X*= x) > P(X = x)	60	12	20	Difficult items	Cheating
Zhang (2008)	1,000	100	10	Low	[-2, -.5]	X* = 1	10, 20, 40	2, 4, 8	20	Difficult items	Cheating

Note. X = generated response, X*= replaced response. P(X*= x) > P(X = x) indicates that a complex mechanism was used that resulted in the new response probabilities being larger than the original response probabilities. N/A = not applicable, N/P = not provided.

spuriously mixed scores. Indeed, some authors (e.g., Liu, Douglas, & Henson, 2009) use the first two labels exclusively when designing and discussing their simulation study results to avoid any confusion about what real-life contexts statistical mechanisms might align with.

Five questions for deconstructing the operationalization process for aberrant responding

Almost all of the authors of simulation studies used a four-step approach whereby (1) data were generated with a particular statistical model, (2) aberrant responses were created by manipulating the generated response vectors, (3) one or different statistical models were fit to the manipulated data, and (4) person fit statistics were computed.

I note that there is technically an alternative, and more indirect, way of inducing person misfit, which consists of simply generating data with a more flexible statistical model and fitting a more constrained – or structurally very different – statistical model to the generated data without ever manipulating the individual response vectors in a separate step. One can think about induced effects in similar ways but the ways in which these are controlled by the researcher in the two general approaches are somewhat different.

In order to keep different operational steps for the creation of aberrant responses conceptually and practically separate I have found it meaningful to consider the following five questions, which are reflected in the organization of the columns in Table 1:

1. How many persons respond aberrantly?
2. What kinds of persons respond aberrantly?
3. How do they respond aberrantly to selected items?
4. To how many selected items do they respond aberrantly?
5. To what kinds of items do they respond aberrantly?

For example, "random guessing" behavior is often created by having (1) a small number of (2) low-ability persons provide (3) correct answers to (4) a small number of (5) high-difficulty items for which they were expected to get an incorrect response in the first place; this creates predominantly "spuriously high" responding. Similarly, "fatigued responding" is often created by having (1) a small number of (2) high-ability persons provide (3) incorrect answers to (4) a small number of (5) very easy items for which they were expected to get a correct answer in the first place; this creates predominantly "spuriously low" responding.

As in other designed experiments, the primary objective of a simulation study is to induce effects that are predictable, but not yet fully quantified, in terms of direction and/or magnitude. Arguably, for the results of the simulation study to be of practical use, authors should also choose design factors and associated levels that match, at least in part, the kinds of real-life application contexts that practitioners operate in. In the next five subsections I now discuss a few implications of the design choices that authors of my sources have made for the generalizability of the interpretations that can be drawn from these studies.

How many persons respond aberrantly? The number of aberrantly responding persons can have an impact on the performance of person fit statistics even though only five of the simulation papers I reviewed explicitly manipulated this design factor (Armstrong & Shi, 2009a, 2009b; Choi & Cohen, 2008; Emons, 2009; Karabatsos, 2003). When these authors discuss the effect of the number of aberrantly responding persons, they typically report that detection rates decrease with an increase in the number of aberrantly responding persons because typical and atypical persons become harder to distinguish. For example, Karabatsos (2003) included a level of 50% of persons responding aberrantly, which showed notably weaker power across all 36 investigated person fit statistics while the remaining levels of 5%, 10%, and even 25% showed similarly high detection rates, albeit with the expected decrease in power for larger percentages.

What kinds of persons respond aberrantly? The kinds of persons that are simulated to respond aberrantly depend, to a large degree, on the kind of aberrant behavior that is being induced in a particular simulation condition. For example, within the context of an achievement test "guessing" and "cheating" persons are typically those of "low-ability" while "fatigued" and "careless" persons are typically those of "high-ability".

Therefore, there are some apparent consistent conventions for selecting parameter ranges for certain kinds of aberrant behavior. For example, Karabatsos (2003), Wang, Pan, & Bai (2008), and Zhang and Walker (2008) all defined "guessing" persons as those persons with latent variable values between of -2 and -.5, which is similar to Armstrong et al. (2007) and Armstrong and Shi (2009b) who selected "guessing persons from the lowest estimated value to -.5.

At the same time, some researchers induce aberrant responses for all persons in their sample for some of their design conditions (e.g., Emons et al. 2003) or all of their design conditions (e.g., de la Torre & Deng, 2008; Sijtsma & Meijer, 2001) even when the same kind of aberrant behavior is generated. This kind of setup has methodological value in that it allows those researchers to describe the relationship between values of the latent trait variable and observed power rates for the entire range of latent trait values under a given latent trait distribution. Nevertheless, as a critical reader of this work one has to be very careful in reviewing how aberrant responding is operationalized as the induced effects can be different for the same kinds of persons across studies.

How do they respond aberrantly to the selected items? All of the studies I reviewed investigated one particular aberrant response behavior for each simulated person in a particular cell of the simulation design; none the authors investigated mixed types of aberrant responding for individual persons to different subsets of items.

As noted previously, aberrant responses were generally induced by replacing conforming responses in the generated data sets by non-conforming ones. Depending on whether the replacement is deterministic (e.g., when the '0' responses of low-ability persons that are "cheating" are replaced by '1's as in de la Torre and Deng, 2008, Emons, 2003, and others) or probabilistic (e.g., when such responses are replaced by '1's with a probability of .80 as in Clark, 2010, or .90 as in Dimitrov and Smith, 2006, and others) spuriously low or high responding can be in effect for all of the targeted items for an aberrantly responding person or only some of the items. In other words, the direction and magnitude of the induced effect can vary for specific items for the same kinds of persons across studies.

Looking across studies, some labels are relatively consistently operationalized with either a probabilistic or deterministic mechanism. For example, "guessing" behavior is a natural candidate for a probabilistic replacement whereas "cheating" behavior is a natural candidate for a deterministic behavior. It may be argued, though, that cheating in real life, when driven by copying answers from a presumable smarter test-taker, may also include the occasional copying of an incorrect response thus justifying a probabilistic replacement of correct responses.

Importantly, in probabilistic replacements, it is outside of the control of the researcher which responses are changed and which ones remain unchanged resulting in less localized control over the precise size of the induced effect for particular items. This variation in the induced effect, even within a particular aberrant responding category such as "cheating", then affects the recorded power of the statistic of interest. Consequently, it affects the strength of the inferred causal relationship between the induced effects and the performance of person fit statistics across studies. As I argue in the discussion, I believe it is the responsibility of methodologists who conduct such simulation studies to discuss comprehensively how their induced effects relate to their observed outcomes, which is unfortunately less often the case than I would have expected.

How many items do they respond aberrantly to? Generally speaking, as the number of items with aberrant responses in a score vector for a person increases locally or globally, detection rates increase for statistics that are sensitive to either local or global aberrancies, respectively, holding all other factors constant (e.g., Emons, 2009; Meijer, 2003). As a result, most authors vary the number of affected items; the ranges vary notably across simulation studies with small values of 8% (Armstrong, 2009a, 2009b) or only 1 or 2 items (de la Torre & Deng, 2008; Dimitrov & Smith, 2006) in some studies all the way to high values of 67%, 75%, or even 100% in other studies (Emons, 2008, 2009).

What kinds of items do they respond aberrantly to? Researchers often use labels such as "easy" or "difficult" items to describe the subset of items with aberrant responses even though they frequently do not provide ranges of item parameter values to identify these items exactly. This is somewhat surprising given that aberrant responses to different *item types* (e.g., high difficulty and high discrimination items versus low difficulty and low discrimination items) will lead to differentially strong induced effects.

Similarly, researchers are not always clear about the way they combined the simulated physical item location and the item type on a simulated test. One example of how to do this well is de la Torre and Deng (2008) who simulated persons who "speed" toward the end of the test as responding at chance levels to the most difficult items. These authors then provided a table with the item parameters used for data generation, which shows that the later items on their test were, indeed, the most difficult ones.

Such a setup is, of course, not the only possible one and "difficult" items could be simulated to be dispersed across different positions on a simulated test. In that case, "later" items that persons may respond aberrantly to may represent a mix of different item types, which again could lead to induced effects of different magnitudes across studies.

Table 2: Summary of test design specifications for simulation studies of person fit

Source	Generating Models	Latent Trait Distribution	Difficulty	Discrimination	Guessing	Fit Statistics
Armstrong (2007)	3P	Uniform	[-2.7, 2.7]	~ [.7, 1.3]	~ [0, .5]	l_z
Armstrong (2009a)	3P	Normal	N/P (M = .07)	M = .77	N/P (M = .17)	CUSUM$_{LR}$, CUSUM$_{LV}$, l_z, U, W, UB
Armstrong (2009b)	3P	Normal	N/P (M = .09)	M = .73	N/P (M = .18)	CUSUM$_{LR}$, CUSUM$_{LV}$, CUSUM$_{IRT}$, U3,C*
Choi (2008)	3P-T	Normal	[-2.43, 2.50]	[.71,2.73]	[.06, .55]	l, U, W
Clark (2010)	GRM	Normal	~[-3, 3][c]	[.5, 2]	N/A	l_{c0}, l_{cz}, M-l_{c0}, M-l_{cz}
Cui (2009)	AHM	Uniform	N/A	[.4, .8][a]	[.1, .2]	HCI
de la Torre (2008)	3P	N/P	[-1.51, 1.06]	[.57, 1.4]	[0, .27]	l_z
Dimitrov (2006)	1P	Normal	[-2.75, 2.75]	N/A	N/A	t, t*, Z_3, Z_3*, H^T
Emons (2002)	MMI, 1P, 3P, MHM	Normal	[-2, 2]	M = (.5, 1, 2)	[0, .2]	U3
Emons (2003)	4P	Normal, Uniform	[-2, 2]	[.6, 1.4]	[.01, .4]	G*, U3, l_0, ζ
Emons (2004)	4P	Normal	N/P	[1, 2]	N/P	LR-β, ZU3, G^2_{β}, $G^2_{\gamma,SL}$, $G^2_{\gamma,SH}$
Emons (2008)	2P, GRM	N/P	[-4.43, 3.96]	[.65, 2.09]	N/A	$U3^p$, G^p_N, l^p_z
Emons (2009)	GRM	Normal	[-3.80, 3.96]	[.65, 2.09]	N/A	l^p_z, $p_{x_{v+}}$
Ferrando (2009)	LFA	Normal	N/A	λ = {.2, .6}	N/A	M-l_{c0}, M-l_{cz}
Ferrando (2010)	CRM	Normal	N/A	[.75, 1.33]	N/A	l_{c0}, l_{cz}
Glas (2003)	3P	Normal	[-2, 1.60]	[.5, 1.5]	.20	l, W, UB, T_1, T_2, T_{lag}, ζ_1, ζ_2
Glas (2007)	2P, GRM, SM, GPCM	Normal	[-2, 1.80]	N/P	N/A	LM test
Hendrawan (2005)	3P	Normal	[-2, 2]	[.5, 1.5]	.20	l, UB, W, ζ_1, ζ_2
Karabatsos (2003)	1P	Uniform	[-2, 2]	N/A	N/A	36 different ones
Liu (2009)	DINA	N/P	N/A	[.40, 1][a]	[0, .3]	LR test (T_1, T_2)
Raîche (2003)	1P	N/P	N/P	N/P	N/P	l_z, W, ζ, I_{ran}, I_{inv}, I_{comb}
Sijtsma (2001)	1P, 4P	Uniform	[-2, 2]	[1, 1]	[.01, .4]	P, ZU3
St-Onge (2009)	1P, 2P, 3P	N/P	[-2.5, 2.5]	[.5, 1.5]	[0, .25]	ECI2$_z$, ECI4$_z$, l_z
St-Onge (2011)	2P	Normal	[-2.5, 2.5]	[.5, 1.5]	N/A	l_z, U3, ECI2, H^T
von Davier (2003)	MRM	Normal	[-2.75, 2.75][b]	N/A	N/A	Z^*_p
Wang (2008)	1P	Normal	[-2, 2]	N/A	N/A	l_z, ECI4$_z$, X^2_D
Zhang (2008)	2P	Normal, Uniform	[-2, 2]	[.4, 2]	N/A	H^T, D(θ)

Note. N/A = not applicable, N/P = not provided, GRM = graded response model, GPCM = generalized partial credit model, SM = sequential model, CRM = continuous response model, MMI = model of marginal independence, MHM = monotone homogeneity model, MRM = mixture Rasch model, ORLCM = order-restricted latent class model, DINA = deterministic inputs noisy and-gate, AHM = attribute hierarchy method. [a] Computed as the difference in slipping and guessing parameters. [b] Visually guesstimated from Figure 1 in paper. [c] Based on standard normal distribution. For the definition of the fit statistics used please see Meijer and Sijtsma (2001) or the sources I analyzed.

Research question 2: Generalizability considerations based on simulated test design conditions

In this section, I critically discuss a few generalizability considerations based on how researchers have decided to set up the design factors in a simulation study other than the aberrant response mechanisms. For organizational purposes I distinguish between primary design factors, which concern the testing conditions that are simulated, and secondary design factors, which concern the number of replications and the type-I error rates investigated.

Primary design factors

The primary design factors consist of (a) the statistical model used for data-generation, (b) the number of simulated persons, (c) the postulated distribution for the latent trait variable, (d) the number of simulated items, and (e) the values of the item parameters for the chosen model. The information for (b) and (d) is embedded in Table 1 whereas Table 2 shows the information for (a), (c), (e), and (f).

Data-generation model. The majority of studies utilized parametric unidimensional IRT models for dichotomous data. The most frequently used models were the Rasch and 3PL models with an intermittent number of studies using the 2PL or even the 4PL model for data generation, the latter being rarely used in practice for real-data analyses.

Notably, only a few researchers have systematically investigated person fit for polytomous IRT models (Emons, 2008, 2009; Glas & Dagohoy, 2007) or continuous-response IRT models (Ferrando, 2010). Similarly, I found only one simulation study that generated data according to non-parametric model assumptions (Emons, Meijer, & Sijtsma, 2002).

Number of simulated persons. A glance at the person sample sizes that were simulated reveals that 1,000 persons is, by far, the most commonly simulated sample size that about half of the authors used. This owes, in large part, to the parametric complexity of the statistical models that were employed. Larger sample sizes increase the estimation accuracy for item parameters and, thus, the accuracy of comparisons between observed and predicted response vectors for individual persons. Only two studies by Armstrong and Shi (2009a, 2009b) used 10,000 simulated persons for the three-parameter IRT model. In general, only about a third of the authors investigated different sample sizes in their work.

Distributional assumption for latent trait variable. When conducting simulation studies most researchers make the common standard normal distribution assumption for the latent trait variable even though about a third of the studies also use non-normal distributions – in particular uniform distributions. However, only two authors used both normal and non-normal distributions in the same study (Emons et al., 2003; Zhang & Walker, 2008).

Differences in latent-variable distributions can be important to investigate, however. As Sass, Schmitt, and Walker (2008) have demonstrated, differences between true and estimated latent trait values under a given model increase when the distributional assumptions for the latent trait variable are violated even though item parameter estimation remains essentially unaffected. Consequently, person-fit statistics that include person parameter estimates in their computation would be affected by an incorrect distributional assumption. Thus, it would be important to include this design factor in more simulation studies on person fit to understand its interaction with the remaining design factors.

Number of items. Researchers are typically using test lengths between about 20 and 60 items with the exception of a few researchers such as Armstrong and Shi (2009a, 2009b) who investigated tests with 100 items. There are two different methodological sides to this issue that are worth pointing out.

On the one hand, simulated longer tests are meant to reflect situations where the persons who take a test are either cognitively mature (e.g., students in higher grades who take a district-wide achievement test or an educational survey) and / or the items are not as demanding (e.g., as in certain attitudinal surveys or personality inventories). However, this excludes any contexts where shorter tests might be used such as when cohorts of children in elementary school or certain special populations are assessed.

On the other hand, there is a methodological rationale for using longer tests in the context of person fit. The chances of detecting person misfit generally increase with the length of the response vector as more statistical information about typical and aberrant responses is available for persons who respond to longer tests. In other words, detecting person misfit for shorter tests is statistically almost impossible independent of which statistical model is used (but see Kim, Finkelman, & Nering, 2008, for a potential approach for shorter tests).

Values of item parameters. The parameters for the items were generally sampled from ranges that are common in large-scale assessment practice even though there are some differences in these ranges. Difficulty parameters for all parametric unidimensional IRT models are typically sampled from a range of about -2 to 2 (e.g., Emons, Sijtsma, & Meijer, 2002; Karabatsos, 2003; Zhang & Walker, 2008) while in a few cases they are sampled from wider ranges such as about -4 to 4 (Emons, 2008, 2009). In the latter case some items at the extreme ends of the scales may have very little score variation, which might have affected how strong some of the induced effects for aberrant responding were for those items.

Ranges for discrimination parameters for the two-, three-, or four-parameter IRT model have lower bounds at about .4 or .5 in some studies (e.g., Glas & Meijer, 2003; Hendrawan, Glas, & Meijer, 2003; St.-Onge et al., 2011), which is quite a bit lower than what most practitioners would consider acceptable for operational purposes. Nevertheless, most authors set them to more realistic ranges of about .75 to 2. Guessing parameters for the three- and four-parameter IRT model typically range between 0 and about .25 to .30 even though some authors generate them from ranges that include .40 (Emons et al., 2003; Sijtsma & Meijer, 2001) or even .55 (Choi & Cohen, 2008).

Inspecting the ranges of individual item parameters is only part of the picture however. It is even more critical to know what the test composition looks like in terms of item types; that is, how many difficult, well-discriminating, and hard-to-guess items versus how many easy, poorly discriminating, and easy-to-guess items there are. Most authors break down results by at least one item characteristic (e.g., Emons, Sijtsma, & Meijer, 2003; St.-Onge et al., 2009), some break them down by multiple item characteristics (e.g., St.-Onge et al., 2011) while others do not break them down by item types at all (e.g., Karabatsos, 2003). A breakdown often seems statistically necessary given the variation of type-I and power rates across item types. If a breakdown is not done, a clear rationale for its absence should be provided. Independent of the result breakdown, it would be helpful if authors discussed item types earlier on in their simulation study design section more often.

Secondary design factors

The secondary design factors that can affect the generalizability of findings in simulation studies are (a) the number of replications used, (b) the nominal type-I error rates investigated, and (c) the fit statistics investigated. The information for (a) and (b) is not shown in either table for space considerations whereas the information for (c) is included in Table 2.

Number of replications. Generally speaking, as the article by Koehler, Brown, and Haneuse (2009) reminds us, more replications are certainly better but only up to a point of diminishing return. The numbers of replications for simulation studies in statistics often seem to be chosen rather arbitrarily, either because the number appears "appealingly simple" (e.g., 100, 250, 500, 1000), or because time constraints prevent authors from running more replications. A more meaningful way to determine the number of needed replications for a simulation study is to run a small-scale pre-simulation study for a few conditions to determine the increase in precision for certain statistics that is achieved by increasing the number of replications.

There appears to be no particular consensus in the person fit research community on how many replications are desirable, which may stem from the fact that there are, in fact, two interesting scenarios to distinguish when it comes to computing type-I error and power rates using replicated data sets. In the first scenario, researchers do not use actual replications but rather simulate a certain number of normally and aberrantly responding persons and then compute type-I error and power rates once as the percentages of persons who are flagged (e.g., Armstrong & Shi, 2009a, 2009b; de la Torre & Deng, 2008; Emons, Sijtsma, & Meijer, 2003; von Davier & Molenaar, 2003).

In the second situation, researchers again simulate normally and aberrantly responding persons but also use replications such that the counts for type-I error and power rates are computed jointly over replications and persons, which accounts for additional uncertainty and makes results more generalizable (e.g., Clark, 2010; Dimitrov & Smith, 2006; St.-Onge et al., 2009, 2011; see also Emons et al., 2003; Zhang & Walker, 2008).

Finally, several researchers simulate persons once across the entire range of the specified latent variable distribution (e.g., Armstrong & Shi, 2009a; Choi & Cohen, 2008; Emons, 2009) while others specifically simulate a targeted number of persons at all intervals of the latent trait distribution (e.g., Armstrong & Shi, 2009b; de la Torre & Deng, 2008; Emons, Sijtsma, & Meijer, 2003; Ro, 2001; Sijtsma & Meijer, 2001).

This has implications for the reporting of type-I error and power rates. If researchers simulate the same number of persons at each latent trait interval they are ensured to have the same precision for the computation of these rates across the scale. In contrast, a post-hoc breakdown from an unequal distribution of persons across the latent trait scale for reporting purposes introduces some chance fluctuations in precision. This can be counteracted somewhat by choosing intervals so that they contain similar or identical numbers of persons, of course, but the resulting breakdown may then seem somewhat odd in terms of interval widths.

In the context of estimation the distinction between *frequentist* and *Bayesian* approaches to model estimation become important; for details on Bayesian statistics more generally I refer the reader to sources like Lynch (2007), Patz and Junker (1999a, 1999b), or Rupp, Dey, and Zumbo (2004). The statistical analogue of replications in Bayesian estimation is the number of approximately independent draws from the *posterior predictive distribution* after estimation has stabilized (i.e., after a so-called “burn-in” period). A larger number of draws provides a more accurate histogram of the shape of this distribution with obvious diminishing returns at some point.

The number of draws after burn-in varied notably across studies with generally few explanations given for why those numbers were selected. For example, Glas and Meijer (2003) used 1,000 iterations as burn-in and retained every fifth draw of the remaining 3,000 iterations for a total of 600 draws from the posterior predictive distributions. Choi and Cohen (2008) used 4,000 iterations as burn-in and retained every draw of the remaining 6,000 iterations for a total of 6,000 draws from the posterior predictive distributions. Emons et al. (2003) used 2,000 iterations as burn-in and retained every 15th draw from the remaining 11,250 iterations for a total of 750 draws from the posterior predictive distributions.

Type-I error rate/empirical sampling distribution. This second factor is really a consideration about whether the shapes of the empirical sampling distributions for person fit statistics conform with the shapes of theoretical sampling distributions, if those have been postulated. Not surprisingly, the most commonly used type-I error rate that was investigated was .05 – it was used in all but one study – which was followed by .10 and .01, which are common alternative cut-offs in practice.

Methodologically, there are four core ways in which researchers working within a frequentist estimation framework can proceed when it comes to computing type-I error and power rates:

1. They can compute the empirical sampling distributions and always use the appropriate empirically-derived cut-off values that ensure nominal type-I error rates for computing power rates (*best method with highest precision*).

2. They can compute the empirical sampling distributions and compare the appropriate empirically-derived cut-off values to the theoretical cut-off values under the theoretical sampling distributions. If differences between the values are "small", the theoretical cut-off values are used to compute power rates, perhaps due to computational simplicity (*defensible, but a bit less precise than using the accurate empirical cut-offs*).
3. They can compute the empirical sampling distributions and compare the appropriate empirically-derived cut-off values to the theoretical cut-off values under the theoretical sampling distributions. If differences are "notable"/"large", they are simply noted as inflated type-I error rates but the theoretical cut-off values under the theoretical sampling distribution are used for the computation of power rates (*not advisable due to implicit over- or under-estimation of power rates*).
4. They do not compute the empirical sampling distributions and always use the theoretical cut-off values under the theoretical sampling distributions (*potentially defensible but not advised if research has not investigated the test design conditions under consideration due to potentially over- or under-estimated power rates*).

While empirical sampling distributions are not always investigated they are sometimes the primary, or even sole, focus of a person fit simulation study (e.g., Emons, Meijer, & Sijtsma, 2002; Ferrando, 2007, 2009; Ro, 2001). My review has shown that most authors are ambiguous about how they computed power rates exactly even though I would assume that most use approaches (1) or (2).

Comparing sampling distributions can be done either for the entire distribution or only in the tails, which can matter when different type-I error rates are used. It could be the case that the density of an empirical sampling distribution is essentially identical to the density of a theoretical sampling distribution at the .01 point but somewhat divergent at the .10 point so that empirical power computations would have to be adjusted for the latter but not the former cut-off. To communicate a more complete picture of how empirical and theoretical sampling distributions are aligned some authors supply graphs of cumulative distribution functions (see, e.g., Liu, Douglas, & Henson, 2009), which is advisable.

Practically, if the empirical sampling distribution of a person fit statistic is being investigated, having a larger number of normally responding persons under the null condition ensures more precision for the determination of appropriate empirically-derived cut-offs that ensure approximately nominal type-I error rates. This is beneficial in that the resulting power computations for the smaller number of aberrant respondents are more accurate – of course, having larger numbers of replications for those conditions would also increase the precision of those rates. In general, allowing for variations in the type-I error rate and using a larger number of normally responding persons and/or replications allows for broader generalizability of the findings.

Range of person fit statistics investigated. The third factor speaks to the ease with which comparisons of performance across fit statistics can be made, because a comparison of many fit statistics with the same simulation design is generally easier than exactly synthesizing findings across simulation studies that use slightly different designs and only partially overlapping sets of statistics.

As noted earlier in the paper, Meijer and Sijtsma (2001) provided a comprehensive review of the statistical formulae, structural similarities, as well as relative performance of 30 person fit statistics based on research conducted up to that point. As shown in Table 2, I found that most simulation studies were conducted for comparative purposes and investigated the relative performance of about two to three person fit statistics, on average. The one major exception to this rule was the study by Karabatsos (2003) that compared a total of 11 non-parametric and 25 parametric person fit statistics.

Among the person fit indices that I came across in my review of the literature, variants of the very flexible likelihood-based l_0 and l_z statistics were most frequently investigated followed by the ZU3 statistic as well as the UB and W statistics for the Rasch model; formulas for these statistics can be found in Meijer and Sijtsma (2001) and the appendix of Karabatsos (2003). The operationalizations of aberrant response behavior that I noted in the previous section can make synthesizing findings across simulation studies in detail challenging even if comparable effect sizes such as power rates or areas under receiver operating curves are used (see specifically Karabatsos, 2003, St.-Onge et al., 2009, and Zhang and Walker, 2008, for examples of those).

What makes syntheses further challenging is that an understanding of the exact power of a particular statistic requires reasoning through a set of interactions between the mechanism for the creation of the aberrant response behavior and several test design conditions, including the way the latent trait variable is estimated. Thus, general statements about the behavior of a statistic can often be made, but precise numerical aggregates of power rates across studies are very hard to compute. Authors often do not present full breakdowns of these rates across all design conditions, probably due to space limitations.

Research question 3: Practical implementations of person fit analyses

In this section I take a brief look at real-data analyses that have been published in the same time frame as the simulation studies that I reviewed to discern what kinds of tools more practically inclined researchers are using when they investigate person fit. This is critical because it speaks to how successful the research literature has been in providing useful tools for practitioners. It also helps to judge whether the investigated simulation conditions mimic those settings in which person fit is actually seriously investigated in practice.

I was able to locate a total of 29 sources that provided details on person fit analyses from the areas of educational and psychological, attitudinal, personality, and health assessment; Table 3 summarizes the main information about the data sets and person fit analyses for them. The sample sizes used in these studies ranged from 375 for a study that used a Rasch model up to 10,000 for a study that used the three-parameter model. Overall, most of the examples were based on assessments that used dichotomously scored items.

Table 3: Summary of key features of published applied examples of person fit analyses

Source	Type of Study	Domain	Construct	Scale	# Items	# Cats	n	Models	Software	Fit Statistics	A Priori Postulated Aberrant Behavior	Steps Taken
Armstrong (2009a)	SS + RD	N/P	N/P	N/P	100	2	10,000	3P	Specialized	$\mathrm{CUSUM}_{\mathrm{LR}}$, $\mathrm{CUSUM}_{\mathrm{LV}}$, l_z, U, W, UB	Fatigue, distraction, cheating, special knowledge	IRV
Armstrong (2009b)	SS + RD	N/P	N/P	N/P	100	2	10,000	3P	Specialized	$\mathrm{CUSUM}_{\mathrm{LR}}$, $\mathrm{CUSUM}_{\mathrm{LV}}$, U3, ZU3, C*	N/P	IRV
Baek (2009)	RD	Educational Achievement	Mathematics Ability	Special	20	> 2	321	PCM	WINSTEPS	Infit-person, outfit-person	N/P	CVA-D
Brown (2007)	SS + RD	Educational Achievement	Mathematics & Reading Ability	N/P	N/P	2	80,000; 82,000	2P	N/P	l_z	N/P	CVA-G
Choi (2008)	SS + RD	Educational Achievement	Reading Ability	FCAT	46, 43	N/P	1,000	3P-T, M-3P	Winbugs	L, U, W	N/P	CVA-G
Conijn (2011)	SS + RD	Personality Assessment	State Anxiety	STAI	N/P	> 2	868	GRM	N/P	l_z	N/P	CVA-D
Custers (2000)	RD	Health Assessment	Child disabilities	PEDI	59, 65, 73	2, 6	412	Rasch	Specialized	“Fit scores”	Cultural differences	IRV
Dodeen (2009)	RD	Educational Achievement	Mathematics Ability	N/P	60	2	1,075	3P	BILOG, WPerfit	l_z	N/P	CVA-D,
Engelhard Jr. (2009)	RD	Educational Achievement	Mathematics Ability	CRCT	15	2	997	Fmy facet models	FACETS	outfit-person	N/P	IRV
Emons (2005)	RD	Educational Assessment	Perceptional Reasoning	RAKIT	45, 50, 50, 60	2	1,646	DMM	MSP	U3, G	N/P	CVA-D
Emons (2008)	SS + RD	Work Psychology	Coping with Malodor	N/P	17	4	828	DMM	MSP5	$G^p, G_N^p, U3^p$	Extreme responding	IRV
Emons (2009)	SS + RD	Personality Assessment	Neuroticism	NEO-PI-R	240 (total)	5	1,046	GRM	MULTILOG	$l_z^p, p_{x_{v+}}$	Negative wording	IRV

Ferrando (2001)	RD	Personality Assessment	Extraversion, Neuroticism, Psychoticism	Modified Eysenck Scale	N/P	N/P	489, 140	2P	NOHARM, BILOG, WPERFIT	l_z	Faking good self-representations	CVA-D
Ferrando (2004)	RD	Personality Assessment	Neuroticism	Modified Eysenck Scale	60	2	436	1P, 2P	NOHARM, BILOG, MATLAB, WPERFIT	X^2, l_0, l_z, PRF	N/P	N/P
Ferrando (2007)	SS + RD	Personality Assessment	Worry	Special	14	5	637	LFA, 2P, GRM	LISREL, MULTILOG, BILOG, WPERFIT	l_{co}, l_{cz}, l_0, l_z	N/P	IRV, CVA-D
Ferrando (2009)	SS + RD	Personality Assessment	Worry, Emotionality	CAR	18	5	734	LFA	Mplus	M- l_{co}, M- l_{cz}	N/P	IRV
Ferrando (2010)	SS + RD	Personality Assessment	Extraversion	Modified Eysenck Scale	35	5	426	LFA, CRM	LISREL	l_{co}, l_{cz}	N/P	IRV
Ferrando (2012)	RD	Personality Assessment	Extraversion, Neuroticism	Modified Eysenck Scale	55, 60	5	531, 436	2P	NOHARM, BILOG, WPERFIT	l_z	N/P	IRV, CVA-D
Glas (2007)	SS + RD	Personality Assessment	Neuroticism	NEO-PI	16, 16, 16	5	N/P	GPCM, GRM, SM [uni + multi]	MULTILOG, Specialized	LM Test	Content relation for subsets of test	N/P
Goegebeur (2010)	RD	Educational Achievement	Mathematics Ability	SIMCE	48	2	3,000	3P-Speed	SAS NLMIXED	LR Test, slope parameters	Speededness	
LaHuis (2009)	SS + RD	Personality Assessment	Big Five	Big Five	N/P	5	387	GRM	MULTILOG	LR Test, slope parameters	Faking	CVA-D
Liu (2009)	SS + RD	Educational Achievement	Grammatical Knowledge	ECPE	30	2	2,922	DINA, Reduced RUM	Specialized	LR Test	Spuriously low, spuriously high	N/P
Liu (2011)	RD	Educational Achievement	Library Sciences	Test for library tutorial	40	2	375	Rasch	BILOG	Infit-person, Outfit-person	N/P	IRV
Magis (2012)	RD	Language Assessment	English-language Aptitude	TCALS-II	85	2	1,373	2P	*ltm* in R	l_z, l_z*	N/P	IRV

Meijer (2002)	RD	Certification	N/P	N/P	70-140	N/P	1,392	N/P	N/P	CUSUM	Fatigue, guessing, loss of concentration	IRV
Meijer (2003)	RD	Personnel Selection	Nonverbal abstract reasoning	TNVA	40	2	992	Rasch, MHM, DMM	RSP, MSP5	l_z, X^2	Misunderstanding of instruction, item disclosure, random responding	IRV
Meijer (2008)	RD	Personality Assessment	Functioning, Self-worth	SPPC	N/P	4	702	MHM	MSP	G_z	Lack of cognitive ability	RV, D
Tang (2008)	RD	Health Assessment	Work instability	RA-WIS	23	2	130	Rasch	RUMM	Infit-person, Outfit-person	N/P	N/P
Wirtz (2011)	RD	Clinical Psychology	Consultation and Relational Empathy	CARE	10	5	326	Rasch	WINMIRA	Infit-person	N/P	CVA-D, CVA-G
Woods (2008)	RD	Personality Assessment	Various	SNAP	15-35	2	2,026	2-level LR	RGLOG, Mplus	Slope parameters	Malingering, socially desirable responding, carelessness, haphazard responding, uncooperativeness, pathology	CVA-D, CVA-G

Notes. SS + RD = simulation study with real-data analysis, RD = real-data analysis only, IRV = inspect response vectors only, CVA-D = covariate analysis with domain-specific covariates, CVA-G = covariate analysis with general demographics. For meanings of the test acronyms please see the source articles.

Types of suspected aberrant behaviors

Overall, the suspected behaviors in the studies I reviewed included aberrant responding due to more than 15 different potential causes, including “choosing extreme response options” (Emons et al., 2008), “negatively worded items” (Emons, 2009), “misunderstanding of instructions” (Meyer, 2003), “item disclosure” (Meyer, 2003), “random responding” (Meyer, 2003), “fatigued responding” (Armstrong & Shi, 2009a), “cheating” (Armstrong & Shi, 2009a), “special knowledge” (Armstrong & Shi, 2009a), “cultural differences” (Custers et al., 2000), “content relations across subparts of the assessment” (Glas & Dagohoy, 2007), “malingering” (Woods et al., 2008), “socially desirable responding” (Woods et al., 2008), “carelessness” (Woods et al., 2008), “haphazard responding” (Woods et al., 2008), “uncooperativeness” (Woods et al., 2008), and “pathology” (Woods et al., 2008).

As these labels suggest, some authors attribute aberrant responding to testing effects such as “content relations across subparts of the assessment”, “negatively worded items”, “cultural differences”), some attribute them to specific – often negative – response intentions of the persons (e.g., “uncooperative responding”, “malingering”, “cheating”), while others suggest more harmless intentions (e.g., “special knowledge”, “misunderstanding of instructions”).

The induced effects include both spuriously low and spuriously high responses. As one might expect, researchers in high-stakes achievement testing situations are mostly concerned with spuriously high responding whereas researchers in personality assessment, attitudinal assessment, and health-outcomes assessment are concerned with both types of spurious responding.

Exploratory and confirmatory approaches for person fit

The assessment of person fit in real-data analyses either takes the form of an exploratory approach, in which no specific hypotheses about aberrant responding are either tested or specifically pursued, or a more confirmatory approach in which specific types of aberrant behavior is suspected for a given assessment. Interestingly, fewer authors reported explicit a priori hypotheses about why persons may be responding aberrantly on their assessment.

Similarly, there were two dominant analytic strategies that researchers reported in order to explain patterns of person fit. The first one was to either report no follow-up strategy or to simply inspect score vectors of persons who were identified as responding aberrantly and speculate about the reasons for the aberrant responses. The second one was to use covariates, often socio-demographic ones such as gender and first language, but also scores from ancillary assessment instruments, to explore correlations between values of person fit statistics and these explanatory variables. Both strategies were about equally common.

As an example of a more exploratory study with domain-specific covariates, consider the study by Doreen and Darabi (2009). The authors administered four personality questionnaires related to attitudes towards math, motivation in math, anxiety towards math, and anxiety towards testing along with a mathematics achievement test to a sample of 1,075 students in 10th grade. They analyzed the achievement test data with the three-parameter IRT model for person fit and the values of the l_z person fit statistics across post-hoc score groups were then correlated with the scores on the personality scales. Correlations were rather weak with the exceptions of those associated with the motivation scale, which ranged from -.31 to -.52, allowing only for some cautious interpretations about the reasons for person misfit; the authors did not give any recommendations for what to do next with the flagged persons.

As an example of a more confirmatory study with both demographics and domain-specific covariates consider the study by Woods et al. (2008). The authors administered a total of 15 personality scales to a sample of 2,026 air force military recruits along with three so-called validity scales measuring rare virtues, deviance, and variable response inconsistency and two pathology scales for compulsive personality disorder and borderline personality disorder; the scores from these five scales were used as covariates, along with basic demographics, for person-fit analyses of the 15 personality scales using a two-level logistic regression model. They found different predictive patterns across five of the 15 personality scales and were able to partially interpret them for this sample even though they did not make any recommendations for how to follow up with the flagged persons.

Software programs and person fit statistics

Most authors used a broad array of software programs that ranged from commercially available ones to programs that were specifically developed for the data-analytic needs of the research teams, including specialized programs just for computing person fit statistics. Even though powerful commercial packages exist that estimate a wide range of parametric IRT models (e.g., AcerConquest, BILOG-MG, IRTPRO, NOHARM, MULTILOG, PARSCALE, TESTFACT), practitioners in the area of IRT have long yearned for more comprehensive data-analytic suites that integrate advanced graphical tools, a wide variety of relative, absolute, item-, and person-fit statistics, as well as routines for estimating uni- and multidimensional IRT models.

Having multiple statistics available in a single program for IRT analysis would certainly be desirable. As Meijer (2003) and Emons (2009) discuss, there is often a relatively strong discrepancy between the persons that get classified as aberrantly responding with different statistics. This can typically be explained by the different sensitivities that person fit statistics have toward different kinds of aberrant responding, which underscores that an analyst’s toolbox should contain person fit statistics that are powerful for detecting a wide range of likely unusual response patterns.

Unfortunately, practitioners often have few statistics available within a given commercial program, which is why researchers typically use a commercial program to estimate item and person parameter estimates and then secondary self-created codes to parse the output and compute the relevant statistics (e.g., Karabatsos, 2003; Zhang & Walker, 2008).

Some researchers have further written specialized secondary programs that compute some popular person fit statistics (see, e.g., the PERSON$_z$ program for the l_z statistic by Choi, 2011, or the WPERFIT program for the l_z, the ECI4$_z$, and a X^2 statistic by Ferrando and Lorenzo, 2000).

This has led to a commonly observed phenomenon where applied researchers are forced to use multiple software programs to accomplish their analytical goals. To illustrate this point, consider the study by Meijer (2003). He first used the program MSP5 to investigate the scalability of his data set from a non-parametric perspective. He then used the Rasch-scaling program RSP to estimate the parametric Rasch model and compute associated fit statistics for monotonicity and local independence. He then used a specialized program written by another researcher to compute two X^2-based person fit statistics. He finally used another specialized code written in S-PLUS by yet another researcher to compute the person fit statistic ρ. Similarly, Ferrando (2004) used NOHARM and BILOG to estimate models, a newly written specialized MATLAB code for specific latent trait estimates and person response curves estimates, and the program WPERFIT to compute the l_z statistic. If person fit analyses are to become more widely used by practitioners, more integrated software suites are indeed desirable to make this process easier.

A brief note on multilevel logistic regression approaches

Of particular note is a recent line of work that situates person fit analysis within a two-step approach using multi-level regression analysis (e.g., LaHuis & Copeland, 2007; Reise, 1999; Woods et al., 2008). In this work, a parametric IRT model is used first to estimate item and person parameters and a parametric multi-level logistic regression model is then used to detect variation in person slopes. This work presents a different variation on person fit analyses in that the statistics are an explicit part of the parametric model, rather than separate quantities that are computed after item and person parameters are estimated. Some researchers further supplement numerical analyses with graphical approaches, in particular the person response function approach (e.g., Emons, Sijtsma, & Meijer, 2004, 2005; Sijtsma & Meijer, 2001). Importantly, the original approach has recently been criticized as violating key statistical assumptions of the multi-level model and has been appropriately modified (Conijn et al., 2011).

Discussion

In this discussion section I want to offer a few concluding observations from my methodological analysis of key recent simulation studies that, I believe, might help methodologists in training and interested practitioners understand the nature of findings in the literature, their implications for practice, and the associated lessons for future research more clearly. I specifically discuss the need for (a) a consistent use of terminology for describing aberrant response patterns, (b) a careful articulation of real-life mechanisms for aberrant responding across disciplines, (c) a clear and comprehensive documentation of resulting design choices, and (d) a clear and comprehensive documentation of applied analyses.

Consistent use of terminology for describing aberrant response patterns

First and foremost, a consistent use of the terms “spuriously low responding/scores”, “spuriously high responding/scores”, and, perhaps, “spuriously mixed responding/ scores” alongside the many real-life labels would be important for the literature on person fit. These three labels are driven solely by statistical implications of aberrant response patterns rather than presumed real-world contexts for their appearance.

Some researchers are already using these terms either exclusively or in clean alignment with, and conceptual separation from more substantively motivated labels. Yet there were still several examples of research that either mixed statistical and substantive terms or only used substantive terms. Overall, I would say that there are too many substantive terms currently floating around in the literature that have very similar statistical operationalizations, which does not facilitate an accessible and comprehensive understanding of how simulation study design choices affect observed effect size patterns.

Careful articulation of real-life mechanisms for aberrant responding across disciplines

Clearly, a more careful articulation of the real-life cognitive mechanisms that drive aberrant responding across different disciplines would yield more nuanced and differentiated mechanisms for inducing appropriate changes to response vectors that are different from those that are currently in frequent use. Theory development would certainly be enriched if more explanatory real-data analyses were conducted with person covariates and if results of these analyses were published more frequently. An excellent example of such analyses is the applied article by Meijer et al. (2008) in which the authors asked teachers to provide short profiles of children whose response pattern had been flagged as aberrant using a particular person fit statistic and then juxtaposed these qualitative explanations with covariates such as children’s age and their gender.

Consider also the work by Emons (2008, 2009) on person’s “tendency to choose extreme response options”, which was motivated by observed patterns in health, attitudinal, and personality assessment. There are certainly a variety of choices that could be made for how one could create such a response tendency for different types of instruments. This may include responding in such way to only certain items, fixed bundles of items, items administered in bundles due to adaptive administrations, items at certain positions of a test, and so on. Moreover, with an increasing number of response options the options for manipulating response vectors increase as well; for example, one could collapse the extreme scores together with the scores in the categories that are adjacent to them, one could transform all non-extreme scores into extreme scores, one could do this probabilistically or deterministically, and one could do this for persons with different levels on the latent-variable continuum.

Clear and comprehensive documentation of design choices and breakdown of outcomes

Not surprisingly perhaps, I found that extracting the relevant pieces of information from sources to create Tables 1 and 2 was sometimes rather challenging. While most authors describe the design, implementation, and analytic strategy of the simulation study in a methodology section, information was sometimes incomplete or awkwardly worded. For example, some authors did not list the distributional characteristics of item parameters and item types or only casually mentioned the type-I error rate that they used in the results section even though it was a key methodological consideration. It was also not always clear how concepts such as "low ability" for respondents or "high difficulty" for items were operationalized (i.e., what range of parameter values were chosen to define these sets). There was also ambiguity around notions of percentages. For example, some authors would state that a certain percentage of items was affected but did not state how many of the overall items this translated into. Sometimes it was also unclear whether a percentage referred to all items on a simulated test, only items in a particular simulated section of the test, or all items in a particular range of the distribution.

As I noted earlier in the paper, even though it may be tempting to describe the relationship between the manipulated simulation study design factors and the observed variation in type-I error and power rates in relatively simple terms (e.g., via main effect descriptions for some key design features; see, e.g., the key figures in Karabatsos, 2003) recent research has continued to underscore the interaction effects between these design factors. For example, research on the popular l_z fit statistic has further expanded our understanding of how the estimation method for the latent trait variable interacts with design factors such as sample size, the nominal type-I error rate, and the mechanism used for inducing aberrant responding to affect the performance of this statistic (e.g., de la Torre & Deng, 2008; Snijders, 2001).

As a result, an overly simplistic – albeit convenient and space-saving – presentation of findings may lead some readers to erroneously believe that certain kinds of aberrant behavior are "generally" easier or harder to detect than others because of something "inherently natural" about these response processes. For example, consider Figure 1 in Karabatsos (2003) that shows the power of 36 person-fit statistics for detecting five different kinds of aberrant responding: "cheating", "creative responding", "lucky guessing", "careless responding", and "random responding". The author found essentially that "random responding" was easiest to detect across all statistics while "creative responding" was hardest to detect across all statistics.

While that display is not incorrect, I caution against the seeming conclusion that "creative responding" is *always* harder to detect than "random responding" on the basis of this study. The reason for this caution is that "creative responding" in Karabatsos (2003) was created by changing the generated '1' responses to 18% of the hardest items (3, 6, or 12 items, depending on the test design) to '0' responses only for high-ability persons. In contrast, "random responding" was created by changing generated responses to all items with a probability of .25 to '1' responses for aberrantly responding persons of all ability levels.

The first scenario thus results in a strong "local" discrepancy for a small number of items, relative to the overall test length, for a specific subset of persons in the range of the population distribution. The second scenario results in a weaker "global" discrepancy for a much larger number of items for persons from all ranges in the population distribution. Thus, I would expect the latter effect to be easier to detect, on average, relative to the former effect, especially for statistics that are sensitive to global discrepancies in patterns, which is what the power summary shows. Yet, if "random responding" in the latter scenario had been restricted to the same kinds of items and persons as in the former scenario the effect would likely have been much smaller and might have led one to conclude that "random responding" might be harder to detect than "creative responding".

In general, it is always important to remember that different statistics are differentially sensitive to different types of aberrant responding. For example, Meijer (2003) summarized the differential sensitivity of four person fit statistics to three different types of aberrant responding in a table (his Table 2). But he also carefully discusses how a statistic like the l_z/M statistic is able to detect more "global" aberrancies across an entire response vector while statistics like the X^2_{ord}or ρ statistic are able to detect more "local" aberrancies within a response vector. He thus cautions against overly simplistic and general interpretations of his tabular summary, which is a useful general lesson about generalizability that I fully support based on this systematic review.

I do not argue against comprehensive simulation studies such as the one by Karabatsos (2003), of course, which are exactly what is needed to create a more coherent understanding of the behavior of person fit statistics across a wide range of test design and aberrancy conditions. The fact that it had been cited 30 times in peer-reviewed publications in *PsychInfo* at the time of this writing and over 60 times according to *Google Scholar* shows that there is a clear thirst for easily digestible, albeit also easily misunderstood, information.

But I do argue that it is important that researchers break down results about type-I error rates / empirical sampling distributions and type-II error / power rates as finely as is needed to make readers understand the complex interaction of design factors. This is clearly relatively easy when the simulation design contains a relatively small number of conditions but can become quickly prohibitive for more complex designs due to space considerations of journal articles. Thus, having links to additional websites with additional results would be desirable. Moreover, as is common in other simulation contexts, an analysis of variance or similar regression-based analysis can be used to determine which factors are the most critical for reporting. Results could then be explicitly broken down only for the factors that showed notable interactions (for a good example of this see, e.g., Gushta, 2012), preferably using figures rather than tables (Cook & Teo, 2011) similar to the idea embodied by Karabatsos (2003).

Clear and comprehensive documentation of applied analyses with multiple tools

Following Emons, Sijtsma, and Meijer (2005), Meijer (2003), and Meijer et al. (2008), who worked from the framework of nonparametric IRT, a comprehensive approach to person-fit assessment should be a multi-step approach. Specifically, it should consist of (1) a *statistical detection step* using at least one powerful person fit statistic for general types of aberrancy (i.e., a “global” person fit statistic) and at least one powerful person fit statistic for more specific types of aberrancy (i.e., a “local” person fit statistic), either within a parametric or non-parametric framework, (2) a *numerical tabulation step* that displays the response vectors of persons identified as aberrant responders and helps analyze them, (3) a *graphical exploration step* using person response functions, perhaps coupled with kernel-smoothing approaches, (4) a *quantitative explanation step* that uses covariates to predict variation in the person fit statistics used in (1), and (5) a *qualitative explanation step* that uses think-aloud protocols, interviews, or other means of linking aberrant responding to some cognitive theory of responding.

However, few of the applied sources that I reviewed have provided a comprehensive, explanatorily rich, and accessible documentation of person fit analyses, which is partly a result of the fact that few applied papers are published where person fit analyses are front and center. Even though person fit analysis is just one part of the larger enterprise of model-data fit assessment it would clearly be desirable to have more comprehensive applied papers and reports published so that methodologists understand more completely the complexity of the meaning-making processes that are involved.

Concluding remarks

I want to close this paper with a few practical considerations. Generally speaking, person misfit is an indication of heterogeneity in the population. That is, either different item parameters for the same measurement model hold for different subpopulations or different measurement models are required for the different subpopulations. Consequently, an alternative to the “detection” approach embodied by person fit statistics is a “modeling” approach wherein one models the influence of nuisance factors directly.

Sometimes, this modeling can take the form of additional dimensions that account for the lack in model complexity that led to the person misfit in the original model (e.g., Clark, 2010). For example, in order to account for “speeded responding” in computer-administered testing sessions one can record response times. The item responses and associated response times can then be modeled jointly in a multidimensional measurement model with appropriate distributional assumptions (e.g., van der Linden, Klein, Entink, & Fox, 2010) to improve model-data fit. In this augmented model, “speeded responding” is now an explicit feature of the model and not anymore a hidden nuisance dimension of a simpler model, albeit at the cost of increased model complexity and associated sample size demands.

Similarly, mixture models are potentially able to capture aberrant responding by statistically sorting persons into different previously unobserved groups whose item parameters might, for example, give indications about what kind of cognitive mechanism might have led them to respond aberrantly. Consequently, it would be of interest to combine studies on person fit detection and mixture modeling in order to investigate, for example, the degree of correspondence between the performance of person fit indices and class membership probabilities for different latent classes (see, e.g., von Davier and Molenaar, 2003, and von Davier and Carstensen, 2010, for some theoretical work in this area in IRT).

A critical extended discussion of these and other modeling approaches is beyond the scope of this paper, which was concerned with understanding the limits of generalizability of person fit research based on the design of the underlying simulation studies. I sincerely hope that this paper served as a useful consciousness-raising device for methodologists in training and practitioners who are interested in learning how to become a critical consumer of the methodological literature on person fit. I firmly believe that with a concerted communication, design, and application effort across broader communities of researchers, research in this area can continue to flourish and extend beyond the important efforts of a relatively small number of dedicated individuals who are currently investing their time and resources.

Acknowledgements

I want to thank Ms. Ying-Fang Chen and Ms. Shauna J. Sweet for critical discussions, notes, and contributions to earlier versions of this paper. This work was made possible, in part, by two grants from the National Science Foundation awarded to the University of Wisconsin at Madison (DRL-0918409 and DRL-0946372). The opinions, findings, and conclusions or recommendations in this paper are those of the authors and do not necessarily reflect the views of the funding agencies, cooperating institutions, or other individuals.

References

Information from articles with an * is explicitly included in tables for simulation and/or real-data studies.

*Armstrong, R. D., & Shi, M. (2009a). A parametric cumulative sum statistic for person fit. *Applied Psychological Measurement, 33*, 391-410.

*Armstrong, R. D., & Shi, M. (2009b). Model-free CUSUM methods for person fit. *Journal of Educational Measurement, 46*, 408-428.

*Armstrong, R. D., Stoumbos, Z. G., Kung, M. T., & Shi, M. (2007). On the performance of the l_z person fit statistic. *Practical Assessment, Research, and Evaluation, 12*. Retrieved from http://pareonline.net/pdf/v12n16.pdf

*Baek, S.-G., & Kim, H.-S. (2009). An empirical study on the relationship between teachers' judgments and fit statistics of the partial credit model. *Journal of Applied Measurement, 10*, 87-96.

Baker, F. B., & S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York, NY: Marcel Dekker, Inc.

*Brown, R. S., & Villareal, J. C. (2007). Correcting for person misfit in aggregated score reporting. *International Journal of Testing, 7*, 1-25.

Choi, S. W. (2011). PERSON$_z$: Person misfit detection using the l_z statistic and Monte Carlo simulations. *Applied Psychological Measurement, 34*, 457-458.

*Choi, H.-J., & Cohen, A. S. (2008, March). *A Bayesian approach to the estimation of person-fit in the testlet model*. Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), New York, NY.

Clark, J. M. III. (2010). *Aberrant response patterns as a multidimensional phenomenon: Using factor-analytic model comparison to detect cheating* (Unpublished doctoral dissertation). Lawrence, KA; University of Kansas.

Conijn, J. M., Emons, W. H. M, van Assen, M. A. L. M., & Sijtsma, K. (2008). On the usefulness of a multilevel logistic regression approach to person-fit analysis. *Multivariate Behavioral Research, 46*, 365-388.

Conijn, J. M., Emons, W. H. M., van Assen, M. A. L. M., & Sijtsma, K. (2011). On the usefulness of a multilevel logistic regression approach to person-fit analysis. *Multivariate Behavioral Research, 46*, 365-388.

Cook, A. R., & Teo, S. W. L. (2011). The communicability of graphical alternatives to tabular displays of statistical simulation studies. *PLoS ONE, 6*, 1-7.

*Cui, Y., & Leighton, J. P. (2009). The Hierarchy Consistency Index: Evaluating person fit for cognitive diagnostic assessment. *Journal of Educational Measurement, 46*, 429–449.

Curtis, D. D. (2004). Person misfit in attitude surveys: Influences, impacts and implications. *International Education Journal, 5*, 125-143.

Custers, J. W. H., Hoijtink, H., van der Net, J., & Helders, P. J. M. (2000). Cultural differences in functional statme measurement: Analyses of person fit according to the Rasch model. *Quality of Life Research, 9*, 571–578.

de Ayala, R. J. (2009). *The theory and practice of item response theory.* New York, NY: Guilford Press.

*de la Torre, J., & Deng, W. (2008). Improving person fit assessment by correcting the ability estimate and its reference distribution. *Journal of Educational Measurement, 45*, 159–177.

Deng, W. (2007). *An innovative use of the standardized log-likelihood statistic to evaluate person fit*. (Unpublished doctoral dissertation). Rutgers, The State University of New Jersey, New Brunswick, NJ.

*Dimitrov, D. M., & Smith, R. M. (2006). Adjusted Rasch person-fit statistics. *Journal of Applied Measurement,7*, 170-183.

Dodeen, H., & Darabi, M. (2009). Person-fit: relationship with fmy personality tests in mathematics. *Research Papers in Education, 24*, 115–126.

*Emons, W. H. M. (2008). Nonparametric person fit analysis of polytomous item scores. *Applied Psychological Measurement, 32*, 224-247.

*Emons, W. H. M. (2009). Detection and diagnosis of person misfit from patterns of summed polytomous item. *Applied Psychological Measurement, 33*, 599-619.

*Emons, W. H. M., Meijer, R. R., & Sijtsma, K. (2002). Comparing simulated and theoretical sampling distributions of the U3 person fit statistic. *Applied Psychological Measurement, 26*, 88-108.

*Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2004). Testing hypotheses about the person-response function in person-fit analysis. *Multivariate Behavioral Research, 39*, 1-35.

*Emons, W. H. M., Sijtsma, K., Meijer, R. R. (2005). Global, local, and graphical person-fit analysis using person-response functions. *Psychological Methods, 10*, 101–119.

*Emons, W. H. M., Glas, C. A. W., Meijer, R. R., & Sijtsma, K. (2003). Person fit in order-restricted latent class models. *Applied Psychological Measurement, 27*, 459-478.

Engelhard Jr., G. (2009). Using item response theory and model-data fit to conceptualize differential item and person functioning for students with disabilities. *Educational and Psychological Measurement, 69*, 585-602.

Ferne, T., & Rupp, A. A. (2007). A synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges, and recommendations. *Language Assessment Quarterly*, *4,* 113-148.

*Ferrando, P. F., & Chico, E. (2001). Detecting dissimulation in personality test scores: A comparison between person-fit indices and detection scales. *Educational and Psychological Measurement, 61*, 997-1012.

*Ferrando, P. F. (2004). Person reliability in personality measurement: An item response theory analysis. *Applied Psychological Measurement, 28*, 126-140.

*Ferrando, P. J. (2007). Factor-analytic procedures for assessing response pattern scalability. *Multivariate Behavioral Research, 42*, 481–507.

*Ferrando, P. J. (2009). Multidimensional factor-analysis-based procedures for assessing scalability in personality measurement. *Structural Equation Modeling: A Multidisciplinary Journal, 16*, 109-133.

*Ferrando, P. J. (2010). Some statistics for assessing person fit based on continuous-response models. *Applied Psychological Measurement, 34*, 219-237.

*Ferrando, P. J. (2012). Assessing inconsistent responding in E and N measures: An application of person-fit analysis in personality. *Personality and Individual Differences, 52*, 718-722.

Ferrando, P. J., & Lorenzo, U. (2000). WPerfit: A program for computing parametric person-fit statistics and plotting person response curves. *Educational and Psychological Measurement, 60*, 479-487.

*Glas, A. S. W., & Meijer, R. R. (2003). A Bayesian approach to person fit analysis in item response theory models. *Applied Psychological Measurement, 27*, 217-233.

*Glas, C. A. W., & Dagohoy, A. V. T. (2007). A person fit test for IRT models for polytomous items. *Psychometrika, 72*, 159-180.

*Goegebeur, Y., De Boeck, P., & Molenberghs, G. (2010). Person fit for test speededness normal curvatures, likelihood ratio tests and empirical bayes estimates. Methodology, 6, 3–16.

Gushta, M. (2012). *A unified evaluation of global and local fit performance under differing test construction conditions and model misspecifications* (Unpublished doctoral dissertation). College Park, MD: University of Maryland.

*Hendrawan, I., Glas, C. A. W., & Meijer, R. R. (2005). The effect of person misfit on classification decisions. *Applied Psychological Measurement, 29*, 26-44.

Hui, H.-f. (2008). *Stability and sensitivity of a model-based person-fit index in detecting item pre-knowledge in computerized adaptive test* (Unpublished dissertation). Hong Kong: The Chinese University of Hong Kong.

*Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education, 16*, 277-298.

Kim, W., Finkelman, M., & Nering, M. (2008, June). *A new person fit procedure for short tests using a minimax algorithm*. Paper presented at the annual International Meeting of the Psychometrics Society (IMPS), Durham, NH.

Koehler, E., Brown, E., & Haneuse, S. J.-P. A. (2009). On the assessment of Monte Carlo error in simulation-based statistical analyses. *The American Statistician, 63*, 155-162.

*LaHuis, D. M., & Copeland, D. (2009). Investigating faking using a multilevel logistic regression approach to measuring person fit. *Organizational Research Methods, 12*, 296-319.

Liu, T., Cao, Y.-W., & Dai, X.-y. (2011). Influence of person misfit on IRT item parameter estimation and data purification. *Chinese Journal of Clinical Psychology, 19*, 622-624.

Levy, R. (2011). Posterior predictive model checking for conjunctive multidimensionality in item response theory. *Journal of Educational and Behavioral Statistics, 36*, 672-694.

Levy, R., Mislevy, R. J., & Sinharay, S. (2009). Posterior predictive model checking for multidimensionality in item response theory. *Applied Psychological Measurement, 33*, 519-537.

Li, F., Cohen, A. S., Kim, S.-H., & Cho, S.-J. (2009). Model selection methods for mixture dichotomous IRT models. *Applied Psychological Measurement, 33*, 353-373.

*Liu, Y., Douglas, J. A., & Henson, R. A. (2009). Testing person fit in cognitive diagnosis. *Applied Psychological Measurement, 33*, 579-598.

*Liu, M.-T., & Yu, P.-T. (2011). Aberrant learning achievement detection based on person-fit statistics in personalized e-learning systems learning systems. *Educational Technology & Society, 14*, 107–120.

Li, Y., & Rupp, A. A. (2011). Performance of the S-X^2 statistics for full-information bi-factor models. *Educational and Psychological Measurement, 71*, 986-1005.

Lynch, S. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists.* New York, NY: Springer.

*Magis, D., Raîche, G., & Béland, S. (2012). A didactic presentation of Snijder's l_z^* index of person fit with emphasis on response model selection and ability estimation. *Journal of Educational and Behavioral Statistics, 37*, 57-81.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

McLeod, L., Lewis, C., & Thissen, D. (2003). A Bayesian method for the detection of item preknowledge in computerized adaptive testing. *Applied Psychological Measurement, 27*, 121-137.

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/ invariance. *Organizational Research Methods, 7*, 361-388.

Meijer, R. R. (1997). Person fit and criterion-related validity: An extension of the Schmitt, Cortina, and Whitney study. *Applied Psychological Measurement, 21*, 99–113.

*Meijer, R. R. (2002). Outlier detection in high-stakes certification testing. *Journal of Educational Measurement, 39*, 219-233.

*Meijer, R. R. (2003). Diagnosing item score patterns on a test using item response theory-based person-fit statistics. *Psychological Methods, 8*, 72–87.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25*, 107-135.

*Meijer, R. R., Egberink, I. J. L., Emons, W. H. M., & Sijtsma, K. (2008). Detection and validation of unscalable item score patterns using item response theory: An illustration with Harter's Self-perception Profile for Children. *Journal of Personality Assessment, 90*, 227-238.

Patz, R. J.,&Junker, B.W. (1999a). Applications and extensions ofMCMCin IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics, 24,* 342–366.

Patz, R. J., & Junker, B.W. (1999b). A straightforward approach to Markov Chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24,* 146–178.

*Raîche, G., & Blais, J.-G. (2003). *The distribution of person fit indices conditional on the estimated proficiency level and the detection of underachievement on a placement test*. Presented at the 68th International Meeting of the Psychometric Society, Cagliari, Italy.

Reise, S. P. (2000). Using multilevel logistic regression to evaluate person-fit in IRT models. *Multivariate Behavioral Research, 35*, 543–568.

Reise, S. (2009). *The theory and practice of item response theory*. New York, NY: Guilford.

*Ro, S. (2001). *Characteristics of a likelihood-based person-fit index under the graded response model* (Unpublished doctoral dissertation). Minneapolis, MN: University of Minnesota.

Rudner, L. M., Skagg, G., Bracey, G., & Getson, P. R. (1995). *Use of person-fit statistics in reporting and analysing National Assessment of Educational Progress results* (Report NCES 95-713). Washington, D.C.: National Center for Education Statistics.

Rupp, A. A., Dey, D. K., & Zumbo, B. D. (2004). To Bayes or not to Bayes, from whether to when: Applications of Bayesian methodology to modelling. *Structural Equation Modeling, 11,* 424-521.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications.* New York, NY: The Guilford Press.

Sass, D. A., Schmitt, T. A., & Walker, C. M. (2008). Estimating non-normal latent trait distributions within item response theory using true and estimated item parameters. *Applied Measurement in Education, 21*, 65-88.

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement, 30*, 298-321.

Shin, M. (2007). *Detection of aberrant response patterns in testing using cumulative sum control schemes*. (Unpublished doctoral dissertation). Rutgers, The State University of New Jersey, Newark, NJ.

*Sijtsma, K., & Meijer, R. R. (2001). The person response function as a tool in person fit research. *Psychometrika, 66*, 191-208.

Sijtsma, K., & Molenaar, I W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage.

Sinharay, S. (2005). Assessing fit of unidimensional item response theory models using a Bayesian approach. *Journal of Educational Measurement, 42*, 375-394.

*St-Onge, C., Valois, P., Abdous, B., & Germain, S. (2009). A Monte Carlo study of the effect of item characteristic curve estimation on the accuracy of three person fit statistics. *Applied Psychological Measurement, 33*, 307-324.

*St-Onge, C., Valois, P., Abdous, B., & Germain, S. (2011). Accuracy of person-fit statistics: A Monte Carlo study of the influence of aberrance rates. *Applied Psychological Measurement, 35*, 419-432.

*Tang, K., Beaton, D. E., Lacaille, D., Gignac, M. A. M., Zhang, W., Anis, A. H., & Bombardier, C. (2010). The Work Instability Scale for Rheumatoid Arthritis (RA-WIS): Does it work in osteoarthritis? *Quality of Life Research: An International of Quality of Life Aspects of Treatment, Care, and Rehabilitation, 19*, 1057-1068.

Tatsuoka, K. K. (2009). *Cognitive assessment: An introduction to the rule space method.* New York, NY: Routledge.

Thissen, D., & Wainer, H. (Eds.) (2001). *Test scoring*. Mahwah, NJ: Erlbaum.

van der Ark, L. A., Hemker, B. T., & Sijtsma, K. (2002). Hierarchically related nonparametric IRT models, and practical data analysis methods. In G. A. Marcoulides & I. Moustaki (Eds.), *Latent variable and latent structure models* (pp. 41-62). Mahwah, NJ: Erlbaum.

van der Linden, W., Klein Entink, R. H., & Fox, J.-P. (2010). IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement, 34*, 327-347.

van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (2001). CUSUM-based person-fit statistics for adaptive testing. *Journal of Educational and Behavioral Statistics, 26*, 199-218.

van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (2002). Detection of person misfit in computerized adaptive tests with polytomous items. *Applied Psychological Measurement, 26*, 164-180.

*von Davier, M., & Molenaar, I. W. (2003). A person fit index for polytomous Rasch models, latent class models, and their mixture generalizations. *Psychometrika, 68*, 213-228.

von Davier, M., & Carstensen, C. H. (Eds.). (2010). *Multivariate and mixture distribution Rasch models: Extensions and applications*. New York:, NY Springer.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item Response Theory. In R. L. Brennan (Ed.), *Educational Measurement* (4th Ed.). Westport, CT: American Council on Education and Praeger Publishers.

*Wang, L., Pan, W., & Bai, H. (2008, June). *Detection power of multilevel latent-trait differential person functioning: A Monte Carlo comparison with conventional person misfit statistics.* Paper presented at the International Meeting of the Psychometric Society (IMPS), Durham, NH.

*Wirtz, M., Boecker, M., Forkmann, T., & Neumann, M. (2011). Evaluation of the "Consultation and Relational Empathy" (CARE) measure by means of Rasch analysis at the example of cancer patients. *Patient Education and Counseling, 82*, 298-306.

Woods, C. M. (2008). Monte Carlo evaluation of two-level logistic regression for assessing person fit. *Multivariate Behavioral Research, 43*, 50-76.

*Woods, C. M., Oltmanns, T. F., & Turkheimer, E. (2008). Detection of aberrant responding on a personality scale in a military sample: An application of evaluating person fit with two-level logistic regression. *Psychological Assessment, 20*, 159–168.

*Zhang, B., & Walker, C. M. (2008). Impact of missing data on person-model fit and person trait estimation. *Applied Psychological Measurement, 32*, 466-479.