Article

Theory & Psychology

Theory & Psychology
1–18

DOI: 10.1177/0959354317706272
journals.sagepub.com/home/tap
SAGE

The sirens' call in psychometrics: The invariance of IRT models

Rodrigo A. Asún
University of Chile

Karina Rdz-Navarro
University of Chile

Jesus M. Alvarado
Complutense University of Madrid

Abstract

The idea that item response theory (IRT) models yield invariant parameter estimates is widely accepted among scholars interested in achieving truly scientific measurements in social and behavioral sciences. Starting from a conceptual and mathematical definition of invariance, this article presents a critical examination of the theoretical and empirical support for the property of invariance with regard to populations and samples of items and subjects by means of simulated data. The distinction between internal and external invariance is introduced to clarify the meaning and limitations of invariance in IRT models. Furthermore, the consequences of "giving in to the sirens' call" of achieving invariant measurements in behavioral sciences are also discussed.

Keywords

external invariance, internal invariance, item response theory, measurement, measurement invariance

In his well-known book *Against Method*, Feyerabend (1975) develops a thesis on the force of attraction that beliefs and strong theoretical standpoints exert on scientists, and how this force erodes their capacity to perceive the limits of empirical evidence.

Corresponding author:
Karina Rdz-Navarro, Department of Sociology, Faculty of Social Sciences, University of Chile, Av. Capitán Ignácio Carrera Pinto 1045, Ñuñoa, Santiago, Chile.
Email: rdznavarro@uchile.cl

Feyerabend illustrated his theory by describing how Galileo used arguments unrelated to astronomical evidence to advocate the heliocentric system. This example and other historical cases (Gratzer, 2001) reveal that strong beliefs can lead scientists to attempt to convince others (and indeed themselves) by means of discursive strategies rather than empirical evidence.

This paper will demonstrate that something similar has occurred (and is continuing to do so) in the field of psychometrics, where an important part of the psychometric community that encourages the use of item response theory (IRT), has come to promote the idea that IRT models have the capacity to achieve invariant measurements. Unfortunately, this property (Hambleton, Swaminathan, & Rogers, 1991) has only been weakly demonstrated, and its actual meaning and implications are more limited than those suggested in disseminating literature.

To support our thesis we provide a conceptual and mathematical definition for the property of invariance and develop a critical examination of the theoretical and empirical support for invariance in IRT models. Furthermore, we introduce the distinction between internal and external invariance to clarify the meaning of invariance in IRT, and discuss some of the negative consequences of giving in to the sirens' call of the imperative achievement of invariant measurements in behavioral sciences.

Invariance in psychometrics: Asking the question

In the fields of mathematics and physics, *invariance* is a property of real or formal systems for which particular types of transformations do not alter the relationships between the elements of a system. Thus, if a system is comprised of measurements of several objects using a single instrument, *measurement invariance* shall be defined as observing the same relationships between measurements when a second measurement instrument is used in the assessment.

In certain fields of science, measurement invariance may be somewhat trivial because measuring an object using equally valid and reliable instruments (e.g., thermometers based on different principles) yields equivalent results. Unfortunately, measurement invariance is not guaranteed in social and behavioral sciences, where meaningful differences in scaling are often found when using two equally valid and reliable instruments (e.g., tests or scales) developed to measure the same construct, or where notoriously different properties are found for the same instrument when applied in different groups, samples, or populations.

Measurement invariance has been regarded as the essential attribute for truly scientific measurement in psychometrics (Jones, 1960), and its achievement is considered a "matter of life and death to the science of mental measurement" (Wright, 1968, p. 85). Thus, a lack of invariance is deemed unsatisfactory for scientific measurement (De Ayala, 2009; Embretson, 1999; Hambleton et al., 1991; Wright, 1968), and considerable efforts have been devoted to defining situations in which it is possible to assume the existence of measurement invariance in behavioral sciences.

Following Meredith (1993) and Millsap (2008), in the field of psychometrics measurement invariance is generally defined as the equivalence between (a) the probability that a subject *j* drawn for a population (Q_k) endorse an item X_i given the subject's ability

(θ_j) and (b) the probability of endorsing the item solely on subject's ability. This equivalence is formally expressed in Equation 1. Thus, if item parameter estimates are equal (within error levels) across different groups of participants, regardless of the population they belong to, items are regarded as invariant.

$$P\left(X_i \middle| \theta_j, Q_k\right) = P\left(X_i \middle| \theta_j\right) \quad (1)$$

Given that participants might belong to an infinite number of populations, each with different characteristics, the necessary and sufficient condition for achieving invariant measurements is that no population characteristic is associated with the probabilities of endorsing the items of the test or scale conditional on the latent trait θ (McDonald, 1982). This means that the probability of endorsing any item of the test is solely a function of the latent trait and, if all items are invariant, all participants with the same level of ability (i.e., the same value in the latent trait) will exhibit the same estimated ability score (within error levels), regardless of the population they belong to. Therefore, invariance should be regarded as a conditional property, which is only relevant in the context of multiple populations (Rupp & Zumbo, 2006), or where one can assume the existence of at least two populations with different characteristics that may interfere with the probabilities of endorsing the items of the test given θ.

Bearing this in mind, the reader might speculate as to whether it is possible to achieve invariance in the context of social and behavioral sciences, and if so, what kind of methodological and statistical tools are available to ensure invariant results.

Invariance in IRT: Stating the answer

Since the earliest developments in the field of psychometrics, numerous psychometricians have struggled to develop methodological procedures to achieve invariant measurements. For example, Thorndike (1922), Thurstone (1927), and Guttman (1950) developed different techniques (i.e., transmuting scores, absolute scaling, and scalogram analysis, respectively; cf. Engelhard, 1984, 2008) with this purpose, in order to be able to compare test scores across different groups. During the mid-20th century, efforts devoted to seeking invariance in the field of psychometrics were transferred to IRT. In this context, Lord (1952) argued that measurement invariance was achievable within latent variable models because it is "possible under certain conditions to define a metric for the ability such that the frequency distribution of ability in the group tested will remain the same even though the composition of the test is changed" (pp. 1–2).

In IRT models, measurement invariance occurs when items exhibit the same item characteristic curves (ICC) across groups of participants or, equivalently, items exhibit the same parameter estimates across groups (Embretson & Reise, 2000). Therefore, measurement invariance should be a matter of empirical research, as addressed by the differential item functioning (DIF) research tradition within IRT models (cf. Camilli & Shepard, 1994; Holland & Wainer, 1993), however, an influential group of scholars have promoted the belief that measurement invariance is an intrinsic property of the IRT framework and, as such, is achievable by fitting IRT models.

One of the first psychometricians arguing that measurement invariance is an intrinsic property of an IRT model was Rasch (1960/1980), who developed the IRT model that bears his name. The Rasch model has the property of *specific objectivity*, which would allow for comparing participants regardless of the specific set of items or instruments used in the measurement process. Consequently, if participant responses fit to a Rasch model, the specific objectivity ensures that (a) differences in the logarithms of the odds of endorsing an item will be equal among any pair of participants, regardless of the item used in the comparison and (b) differences in difficulty parameters between any pair of items will be equal across groups of participants participating in the estimation. Thus, it would allow the achievement of invariant estimates across items and participant parameters.

Given that specific objectivity was originally claimed as a property of Rasch-type models, certain authors argue that invariance is a property restricted to these types of models (Fischer & Molenaar, 1995; Wright, 1999), whereas other authors (e.g., De Ayala, 2009; Embretson & Reise, 2000; Hambleton et al., 1991; Reise & Haviland, 2005) generalize this property to all IRT models. These latter authors argue that all IRT models are invariant because they: (a) estimate participants' parameters taking into account the properties of the items and estimate item parameters taking into account participant abilities (De Ayala, 2009; Embretson, 1996); (b) ensure that estimated probabilities of endorsing an item solely depend on the ICC and not on participants' abilities (Hambleton & Swaminathan, 1985); and (c) take the form of a regression (albeit nonlinear), and the estimation of a regression is invariant as it does not depend on the distribution of the abilities of the group assessed (Lord, 1980).

Thus, in most handbooks and articles promoting the IRT framework, the property of invariance is highlighted as one of the main advantages of IRT models compared to classical test theory (CTT). For example, Hambleton et al. (1991) argue that measurement invariance "is the cornerstone of IRT and its major distinction from classical test theory" (p. 19). Reise, Ainsworth, and Haviland (2005) consider invariance as one of the main characteristics of IRT models because, without this property, it would be "virtually impossible to administer a common measure to different groups, compute raw scores, and make meaningful comparisons" (p. 97). Embretson and Reise (2000) listed the property of invariance as one of the new rules of measurement which emerged with the IRT framework. Thus, IRT advocates seem to believe that invariance is a goal achievable using IRT models for test and scale development, and this is regarded as a revolution in the field of psychometrics: as stated by Wright, "a new measurement in psychology has emerged from a confluence of scientific and social science methodology" (1999, p. 65).

Interestingly, the authors who consider invariance to have been achieved do not provide a formal definition for the concept, although they seem to regard it as a universal or unconditional property of IRT models which enables the complete independence of estimates from the samples of participants, populations of individuals, and sets of items used in the assessment. For example, some authors claim that IRT estimates are "population-free" and "test-free" (Embretson, 1999, p. 8), which is a highly desirable and useful property "because it frees the practitioner from the specific characteristics of the instrument and samples used" (De Ayala, 2009, p. 409), enabling participant parameters to be "estimated independently of the particular test items" (Hambleton & Russell, 1993, p.

42), and making the assessment independent of the characteristics of any particular population (Reise et al., 2005). Furthermore, they suggest that invariance in IRT enables the calibration of tests with biased samples of the target population, arguing for example that "unbiased estimates of item properties may be obtained from unrepresentative samples" (Embretson & Reise, 2000, p. 23), or that "sample invariance inherent within IRT means that test developers do not need a representative sample of the examinee population to calibrate test items" (Hambleton & Russell, 1993, p. 45).

From the arguments cited above, it seems possible to infer that IRT enables the validation and estimation of the properties of a test using a biased sample which, for example, may over or under-represent any group or subpopulation (e.g., a sample that under-represents a particular ethnic group) or a sample of participants with biased levels of ability in the latent trait (e.g., a sample comprised only of participants with high levels of ability). It should be clarified that the property of invariance in IRT does not imply that exactly the same item parameter estimates will be obtained when fitting a model to different samples (e.g., samples of participants with high and low ability). Due to the indeterminacy of the estimation (i.e., arbitrary values for the mean and scale of the latent trait), parameter estimates only will be linearly related (DeMars, 2010; Rupp & Zumbo, 2004). As a result, IRT estimates "are invariant only within a linear transformation" (Reise et al., 2005, p. 96). Only after equating the parameters to exhibit the same metric will estimates be equivalent.

The property of invariance is a potentiality which is materialized for a data set if and when the IRT model fits the data. As Reise and Haviland (2005) assert, "any advantages that IRT modeling may have relative to CTT can only be realized in practice when data are judged appropriate for IRT models and the estimated IRT model parameters fit the observed data" (p. 230). In keeping with this, De Ayala (2009) argues, "theoretically, IRT item parameters are invariant … However, whether invariance is realized in practice (i.e., with parameter estimates) is contingent on the degree of model-data fit" (p. 61). As a consequence, given that invariance is a property of real or formal systems, within IRT models, invariance is only a potentiality of its mathematical function, and will remain in this state until model-data fit is proven. In that scenario, what are the limits (if any) of invariance when participants' responses to a set of items fit a given model? In the sections that follow, we will address these points.

Invariance in IRT: Questioning the answer

As mentioned before, IRT models are regarded as a paradigm shift in psychometrics, as well as methodological tools that yield invariant parameter estimates when evidence of model-data fit exists. Thus, if an IRT model fits to a population of items and participant responses, any sample or subsample of items drawn from that population will produce the same participant parameter estimates (after equating the metric for the subtests) and, consistently, equivalent item parameters will be obtained when calibrating the test with any sample or subsample of participants drawn from that population for which the model fits. However, it is not possible to assume that invariance will remain if the same set of items is applied to another sample drawn from a different population of participants, or from a subpopulation that behaves in a different fashion with regard to the latent trait

because, in that scenario, the IRT model may not fit the data or, even more importantly, may fit according to different parameters.

To illustrate this point, a Monte Carlo simulation, similar to those conducted when differential item functioning (DIF) phenomena are analyzed, was conducted according to the Rasch model (although its generalization to other IRT models is trivial) for unidimensional dichotomous tests of 40 items, applied to samples of 2,000 participants where θ values were drawn from a standard normal distribution. Five experimental conditions were created by manipulating the values of difficulty parameters for participants with high levels of ability, thus generating a slightly different Rasch model for one of the ability subgroups. Condition I was generated as the baseline. Here all participants—regardless of their level of ability—responded to the items according to exactly the same Rasch model (i.e., a model with the same parameters) with item difficulty parameters (i.e., parameter *b*) drawn from a standard normal distribution. In conditions II through V, we first generated the difficulty parameters for all items according to a standard normal distribution. Then the difficulty parameters were modified by random numbers drawn either from a Uniform (-1.5, 1.5) distribution (in conditions II and III) or a $\chi^2(1)$ distribution (in conditions IV and V) for items with a level of difficulty greater than the mean difficulty of the test (i.e., $b > 0$), and for participants with levels of ability greater than the mean population ability (i.e., $\theta > 0$). For conditions II and IV, random numbers were summed to the original difficulty parameters, and for conditions III and V, the random values were multiplied by the original parameters. This simulation implies that in conditions II through V, all participant responses fit to a Rasch model, but the population of respondents with higher levels of ability answered the test according to slightly different difficulty parameters than participants with lower levels of ability.

A total of 500 replicates were created for each condition using the software R 2.15.2 (R Development Core Team, 2012). After data generation, each sample was split into two subsamples (i.e., A and B) of low and high values in the latent trait. Each subsample was calibrated independently using the Rasch model implemented in the *LTM* package (Rizopoulos, 2006). The correlation between parameter estimates in subsamples A and B was assessed for each condition to evaluate invariance of results.

The average Pearson correlation among difficulty parameter estimates in subsamples A and B across replicates equals .994 for condition I and .845, .598, .773, and .738 for conditions II through V, respectively. To illustrate these results, Figure 1 depicts the observed relationship between parameter estimates in subsamples A and B for a randomly selected replicate. The upper row in Figure 1 illustrates the strong linear relationship between parameter estimates in the two subsamples in condition I, while the second and third row depict a lower correlation of estimates in both subsamples in conditions II, III, IV, and V. Please note that the weaker relationships and the greater variability observed in conditions II through V result from true population differences in difficulty parameters among participants with low and high levels of ability, and are not the result of a larger estimation error.

The almost perfect correlation observed in condition I demonstrates that when all data is generated from the same Rasch model, parameter estimates are invariant even if estimations are conducted on samples with biased (e.g., high or low) levels of ability. In contrast, the smaller correlation observed in conditions II through V demonstrates that

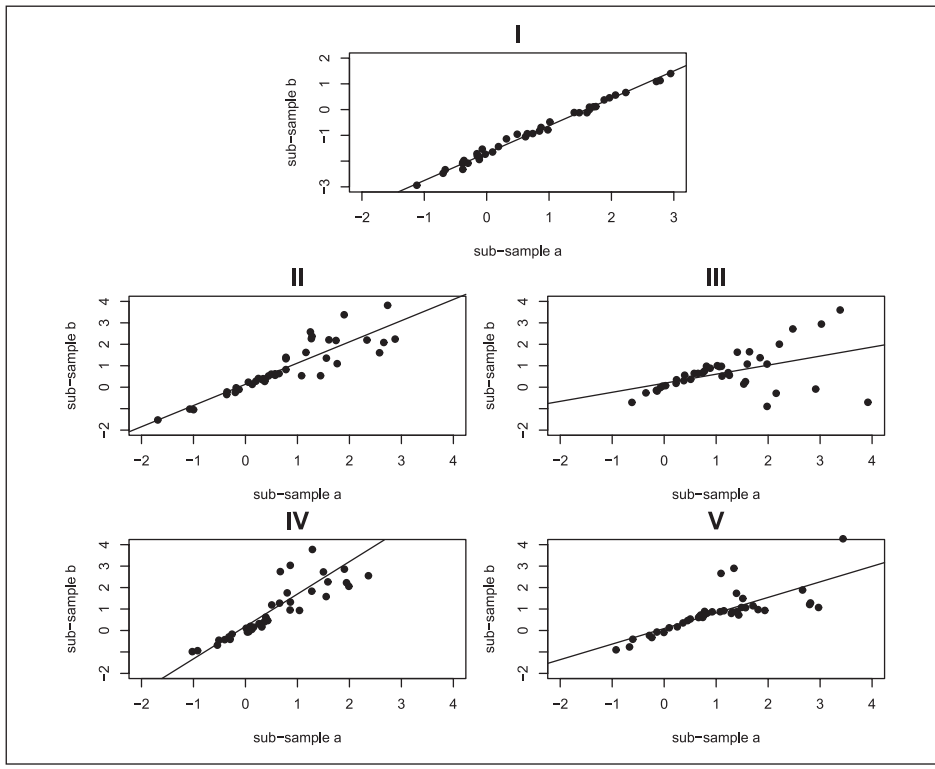

Figure 1. Observed relationship between parameter estimates in subsamples A and B for a randomly selected replicate.

invariance is not achieved when participant responses are produced by a slightly different Rasch model, estimated on samples of participants with different levels of ability. This means that invariance will not hold whenever two or more populations—in this case, two populations defined by their level of θ—exhibit certain characteristics correlating with the probability of endorsing some of the items in the test given θ, even if the model fits the data within each population.

This example demonstrates an important limitation on the possibility of generalizing the invariance of IRT models, namely the difference between *internal* and *external invariance*. IRT models are *internally invariant* because they will yield the same parameter estimates *within* the single population of items and participants (and samples drawn from that population) for which evidence of fit exists. However, this property will not remain if a group in the population assessed has at least one characteristic that interferes with the participants' conditional responses to the test, or if a different population (and samples) of items and participants is assessed. In other words, contrary to conventional wisdom promulgated by IRT advocates, IRT models themselves cannot support the assumption of *external invariance* of the results (i.e., the invariance *between* populations and samples).

The distinction between internal and external invariance seems unnoticed in previous literature on psychometrics, so much so that demonstrations found in papers and well-known handbooks to support the idea of invariance in IRT models only demonstrate internal invariance, and appear simply to assume external consequences. There follows a presentation and discussion of three examples retrieved from the most-cited IRT handbooks, illustrating this situation.

The first example is taken from De Ayala's (2009) work. The author aims to demonstrate that invariance "is not present in the application of CTT, but it is exhibited in IRT" (p. 409). He conducted a small Monte Carlo experiment according to the one-parameter logistic model for a dichotomous test of 20 items and a sample of 1,000 participants. He split the test into two subsets of difficult and easy items, and then computed the raw score (RS) of each participant on each subtest. The correlation between both series of RS was equal to .713, which (in the author's opinion) demonstrates that CTT participant estimates are not invariant. Interestingly, IRT participant estimates from the two subsets of items exhibited a correlation equal to .745, which is evidently close to the correlation value observed for RS and does not demonstrate invariance in the IRT. However, De Ayala argues that in this scenario, the measurement error increases as a consequence of using short subtests, and so he repeated the exercise using two subsets of 50 items each, finding a correlation between IRT participant estimates equal to .933. He regarded the results as proof of invariance in IRT.

Although this result may indeed be correct, we should bear in mind that both IRT and CTT estimates are affected by test lengths, and RS should also yield a stronger correlation in a longer subtest, consequently exhibiting greater degrees of invariance in those situations. Unfortunately, this information was not reported by the author.

To evaluate this hypothesis, we replicated De Ayala's (2009) work in a Monte Carlo study using 500 replicates. The average correlation between IRT participant estimates was equal to .773 and the correlation between RS was equal to .717 for subtests of 10 items. When subsets of 50 items were used, the correlation between IRT participant estimates reached .924, and the correlation between RS reached .82, which appears to support the idea of greater degrees of invariance in IRT if compared to CTT. However, if instead of computing RS we replace them with the Z-value of the proportion of correct responses to the test of each participant—as Fan (1998) has suggested doing in order to avoid ceiling and/or floor effects generated by the metric of RS—the average correlation between the two subtests increased to .914. Therefore, after this simple transformation, CTT estimates achieve a level of invariance equivalent to IRT estimates.

The second example is taken from the work of Embretson and Reise (2000). The authors created the responses to a test of 30 items using a Rasch model, in a sample of 3,000 participants and simulated participant abilities and difficulty parameters according to a normal distribution. The authors split the sample into two groups (using the median of θ) of low and high level of ability, and calibrated each subsample independently. Based on the large correlation ($r = .997$) between difficulty parameter estimates of both samples, they concluded that IRT estimations are invariant. They discarded invariance in CTT scoring because the relationship between the proportion of correct responses of each item (i.e., parameter p) in both samples was monotonic but not linear ($r = .8$). Interestingly, the lack of linearity in p-parameters seems to be the consequence of ceiling

and/or floor effects inherent to the *p*-metric. If instead of computing the correlation between p-parameters we replace them by the Z-value of *p*, nonlinearity will tend to vanish, and CTT parameters will also exhibit high levels of invariance. To evaluate this argument, we reproduced Embretson and Reise's (2000) study using 500 replicates and found an average correlation between IRT difficulty parameters equal to .962, an average correlation of *p*-parameters equal to .814, and an average correlation between Z-values of *p*-parameters equal to .996. This reveals that a simple transformation of the metric of estimates in CTT may result in greater degrees of invariance.

The last example is taken from Hambleton et al. (1991). They evaluated the invariance of IRT using survey data and suggested that a reasonable approach to assessing invariance would be to split the sample twice: first by randomly assigning participants to two groups, and second by splitting the sample according to the median ability estimated with the IRT model. They found a correlation of ability estimates equal to .86 when the groups were created by random assignment, and a correlation equal to .80 when groups were created according to median ability. According to their criteria, these results demonstrated invariance. Although the authors may be correct in their interpretation of the results, it is important to stress that their exercise only demonstrates what we have termed internal invariance, and not external invariance, since the authors, instead of comparing estimates between populations, compare estimates within one sample for which evidence of fit to a single IRT model was found.

All three examples described above are demonstrations of internal invariance of estimates, thus conclusions about invariance derived from them cannot be generalized beyond the set of items and sample of participants used in the estimation; for instance, to another set of items developed to measure the same construct for which another IRT model may fit, or to another population of participants that may take the test. Indeed, even though in the above examples, samples were split into subsamples with contrasting levels of ability, this demonstration of invariance is fairly restricted. In the first two examples, the demonstration of invariance was a tautology, as datasets were created from a single IRT model, with no population characteristic interfering with the probabilities of endorsing any of the items. As a consequence, invariance observed in the results was caused by the simulation procedure, not by the IRT model used in the analyses. Similarly, in the third example, the demonstration of invariance in real data was conducted according to the magnitude of θ, which does not guarantee that further segmentations of the sample according to other variables (e.g., age, gender, etc.) will produce invariant results. Therefore it can be seen that external invariance was neither demonstrated nor considered in these examples which are so frequently cited in support of the use of IRT models as a means of prevention of the influence of other population characteristics, that is, against the lack of external invariance.

Invariance, model-fit, and (sub) sample size: Exploring the boundaries

As mentioned earlier, the property of invariance in IRT depends on the possibility of establishing the degree of model-data fit; in fact, for some authors "invariance and model-data fit are equivalent concepts" (Hambleton, 1994, p. 540). However, because

invariance is also a property of the sample for which evidence of model-fit exists, inferences to a broader population depend on the design of the sampling procedure. Thus, examining the limits of the property of invariance requires a discussion of the problems relating to model-data fit and to sampling design.

Regarding the evaluation of model-data fit and invariance in IRT models, it is known that invariance is a property of the model parameters (Hambleton, 1994), which "only holds when the fit of the model to the data is exact in the population" (Hambleton et al., 1991, p. 23); however, in real-life applications varying degrees of misfit and lack of invariance will always be found due to the probabilistic nature of IRT models and therefore invariant parameters are unlikely to ever be found in these scenarios (DeMars, 2010). Applied researchers should rely on statistical tests to assess model-data fit under the assumption that relative-fit is sufficient to achieve an acceptable degree of invariance. Unfortunately, it is a highly complex task to evaluate when the degree of misfit is meaningful enough to reject an IRT model, because it is known that some goodness-of-fit statistics exhibit unacceptable power and Type I error rates (Liu & Maydeu-Olivares, 2013; Orlando & Thissen, 2000), or tend to reject correct models as a consequence of minor degrees of misfit when the sample size is large (Embretson & Reise, 2000; Hambleton et al., 1991). Thus, statistical tools to assess model-data fit in the IRT seem to be insufficiently developed to enable a straightforward and effective decision-making process for applied research aiming to assess invariance.

The most frequently cited IRT handbooks (e.g., Embretson & Reise, 2000; Hambleton et al., 1991; van der Linden & Hambleton, 1997) tend to elaborate on the merits of a particular IRT model and the complexity of parameter estimation, but do not discuss in detail their requirements in terms of sample design. Indeed, when sample issues are mentioned, authors briefly indicate that samples need to be large and heterogeneous (Hambleton & Russell, 1993), or that samples do not need to be random (DeMars, 2010) or representative of the population of interest (Embretson & Reise, 2000; Hambleton & Russell, 1993) because of the property of invariance in IRT. However, modern sampling theory demonstrates that sample-based inferences to the population are supported only when samples are representative, namely, when the sample accurately reflects the characteristics of interest in the population. Thus, even though a large sample size allows for smaller variation around estimates (i.e., smaller standard errors), this should not detract attention from the fact that "large unrepresentative samples can perform as badly as small unrepresentative samples. A large unrepresentative sample may do more damage than a small one because many people think that large samples are always better than small ones" (Lohr, 2009, pp. 8–9).

Highlighting sample size as the single key feature for inference based on IRT models neglects the importance of sample design and representativeness for research on invariance. Inferences regarding the external invariance of results may hold only if satisfactory evidence of internal invariance has been obtained from a representative sample. Thus, assuming that a biased sample will yield unbiased parameter estimates means believing—without empirical proof—that none of the characteristics of the misrepresented groups in the sample are associated with the conditional probabilities of endorsing the items, and this belief is unrealistic.

In addition, most human populations are heterogeneous and could be understood as the aggregation of several subpopulations, which may exhibit characteristics associated with the conditional probabilities of endorsing the items in a test. Therefore, it is difficult to establish the set of subpopulations for which invariance holds, because even if the overall sample size is large and representative of the general population, the sample size of—at least—one subpopulation, for which a different model may fit, might not be large enough to produce meaningful levels of misfit in the overall sample.

To illustrate this point, a Monte Carlo simulation was conducted on three conditions. We created a total of 500 replicates for each condition, considering samples of 1,000 participants and responses to tests comprised of 40 items, according to a two-parameter IRT model where participants' abilities and difficulty parameters were sampled from a standard normal distribution. Discrimination parameters were set according to a Uniform (0.5, 2.5) distribution. In condition I, all participants belong to one population (i.e., there were no subpopulations) where a two-parameter IRT model fit, whereas in conditions II and III, participants were randomly assigned to two subpopulations (A and B). In condition II, 95% of the participants were assigned to subpopulation A, and 5% were assigned to subpopulation B. In condition III, 55% of the participants were assigned to A, and 45% were assigned to B. In conditions II and III, discrimination parameters were the same for all items and participants in both subpopulations, but difficulty parameters were different for each subpopulation (although they were drawn from the same type of distribution). With this design, we aimed to represent heterogeneous populations comprised of subgroups of different sizes, which respond to a test according to different parameters which belong to the same generic IRT model. Data analyses were conducted for the total sample of each condition, ignoring the subpopulations, in order to evaluate the effectiveness of goodness-of-fit statistics in detecting this lack of invariance.

Table 1 displays the mean goodness-of-fit statistics across the 500 replicates for each condition. Results revealed a good fit of participants' responses in condition I and an acceptable fit in condition II. In condition III, a significant proportion of items exhibited a misfit, especially in the analysis of doublets and triplets of items: this could be misinterpreted as a problem of local dependence, whereas in fact it is a problem of lack of invariance.

These results confirm that when populations are heterogeneous and comprised of subpopulations whose parameters are different, the overall sample goodness-of-fit statistics only yield evidence on subpopulations for which the sample size is large enough to generate meaningful misfits in the global sample. If the subpopulation with different parameters is small (or severely underrepresented), its differences from the larger groups are likely to be overlooked. Thus, researchers should bear in mind that evaluating model-data fit to establish the possibility of measurement invariance in heterogeneous populations is a complex task that requires research of its own (Hambleton et al., 1991), and only when representative and large samples of each relevant subpopulation are available would it be possible to provide empirical evidence on measurement invariance for each relevant subpopulation in order to support inferences to the global population. This is no different from any other research in the social and behavioral sciences.

Table 1. Goodness-of-fit statistics for each condition.

Goodness-of-fit statistics	Condition		
	I	II	III
Mean χ^2 statistic per item	9.31	9.86	10.89
Percentage of items that yield misfit according to item χ^2 statistic	6.9%	9.6%	14.0%
Mean χ^2 statistic per doublets of items	1.03	2.02	9.96
Percentage of items yielding misfit according to χ^2 of doublets	0.2%	4,0%	25.6%
Mean χ^2 statistic per triplets of items	3.64	7.78	32.26
Percentage of items yielding misfit according to χ^2 of triplets	0.4%	9.7%	47.5%
Mean Lz statistic per subject	0.23	0.23	0.24
Percentage of subject yielding misfit according to Lz	2.3%	5.1%	2.0%

Note. Lz = Standardized version of L_0 statistic (Levine & Rubin, 1979).

Invariance in IRT: Stating the consequences

Given the evidence presented so far, the reader may wonder how advocates of IRT can suggest that estimations within this framework are sample-free, that representative samples are not required, or why no explicit advice has been given with regard to sample requirements to ensure the external validity of results. These concerns are shared by some psychometricians, and some criticism has emerged regarding the hypothetical property of invariance in IRT models.

For example, McDonald (1999) considers invariance to be a mathematical tautology rather than a property of IRT models because "if the item parameters from two groups cannot be rescaled so as to coincide … we can always use population membership as a 'latent trait' and make a model whose parameters are tautologically invariant" (p. 326). Rupp and Zumbo (2004, 2006) argue that invariance is a relational property of multiple populations which is meaningless when a single population is assessed, as indeed seems to be the case in most research examples where the property of invariance is demonstrated. Millsap (2008) regards invariance as a theoretical property which has "little role to play in any actual investigation" (p. 196), because "invariance is an empirical property of items that may or may not hold, but is not mandated by the structure of a particular latent variable model" (p. 197). Moreover, Hambleton et al.'s (1991) statement regarding the equivalence between invariance and model-fit could be interpreted as an implicit acknowledgment that invariance is restricted to samples and populations for which evidence of fit exists, and may not be generalized beyond that evidence. Muñiz and Hambleton (1992) also suggest the internal meaning of invariance in IRT when they claim that invariance can refer only to those tests comprising items that belong to an item-bank and which are calibrated on the same scale. Otherwise there is no such invariance, to the extent that without item-banks, IRT does not yield any meaningful difference compared to CTT.

These statements reveal that in applications involving real data, invariance is not guaranteed and each subpopulation must be empirically assessed, as recommended by the tradition of factor analysis (Maydeu-Olivares, Morera, & D'Zurilla, 1999) and DIF studies (Camilli & Shepard, 1994; Holland & Wainer, 1993). However, these cautions are in contrast with the widespread belief that the property of invariance in IRT allows measurements to be test-free, sample-free, and population-free, and with the idea that invariance (i.e., internal and external invariance as defined previously) is guaranteed when evidence of model-data fit is available.

The lack of clarity regarding the meaning of the property of invariance in IRT and the boundaries of inference surrounding it, has generated at least three important negative consequences.

First, it has obscured the real differences between CTT and IRT estimates. Indeed, some authors have tried to emphasize the differences between CTT and IRT to create the impression of a paradigm shift, however, differences between the two approaches are only related to differences in the procedures for estimating participant abilities, item properties, metrics thereof (i.e., bounded in CTT, unbounded in IRT) and the capacity to model the relationship between a latent trait and the participants' responses to the items in IRT. Therefore, instead of highlighting their differences, it might be more productive to address their similarities, as other authors have done (cf. Holland & Hoskens, 2003).

Second, it has generated confusion among social and behavioral scientists attempting to provide empirical proof of the advantage of IRT models, and has made them more likely to force their data in order to confirm their expectations using misleading analyses. For instance, Adedoyin, Nenty, and Chilisa (2008) attempted to demonstrate the invariance of IRT estimations by comparing the mean of participant parameters across several pairs of samples, disregarding the fact that IRT software arbitrarily fixes the mean of the latent trait at around zero and the standard deviation at one. As a consequence, all samples will yield the same mean, and therefore such comparison does not prove any property within the model.

Third, it has led some researchers to overlook the probabilistic and statistical nature of research in social and behavioral sciences by misinterpreting invariance in IRT as a protection against the impact of population characteristics, or as a property that frees the researcher from the need to use a representative sample when examining the validity of tests. For example, Breithaupt and Zumbo (2002) argue that IRT models "are not theoretically sensitive to examinee characteristics unrelated to ability (such as gender, or average group performances)" (p. 391), and Chernyshenko, Stark, Drasgow, and Roberts (2007) developed a scale intended for the general population and conducted a study of a sample of students, claiming that "whereas it is true that college samples likely show higher means on order than does the general U.S. population, it is important to note that IRT item parameters are subpopulation invariant" (p. 95). But, as we have shown, only when samples are representative of the population can they be used to estimate population characteristics with a known degree of accuracy (Lohr, 2009).

In this paper, we have argued that population characteristics (e.g., age, gender, language, nationality, race, or even differences in participants' abilities) might be related to the probability of endorsing one or more items conditional on θ, and that using IRT models (or any other statistical model) does not provide any protection against their

influence. Therefore, to ensure that a certain population characteristic (e.g., gender) does not interfere with the measurement of a construct, the sample size should be large enough to enable demonstrations of model-data fit in the overall sample, and the assessment of model-data fit, DIF, or differential test functioning (DTF) in all subpopulations resulting from that characteristic.

The main negative consequences of the ambiguous definition of measurement invariance in IRT are, on the one hand, overlooking the restrictions for population inference when working with samples with limited or no representativity of the population of interest (e.g., a sample of college students when developing an instrument intended for the general population) and, on the other hand, overlooking the need to employ large and representative samples for each relevant subpopulation assessed with the instrument. These omissions have reduced attention to the limits of test validation, and might have negative consequences for participants assessed using tests that were calibrated on a different population or with a biased sample.

Concluding remarks: The sirens' call and self-delusion in IRT

This article has analyzed the concept of invariance in IRT and its theoretical and empirical support. It has demonstrated its empirical limitations and discussed the consequences of unclear definitions of invariance for applied research in the behavioral sciences.

Despite the fact that many advocates of IRT have promoted the belief that this framework yields unconditionally invariant measures (i.e., internal and external invariance) this paper has provided evidence to support the idea that IRT is only internally invariant, and not externally invariant per se. This means that inferences regarding the invariance of results are restricted to the populations of items and participants that are accurately represented in the sample of participants and items used in the calibration of the test, provided that the model fits the data. Thus, the properties observed in a given sample are not generalizable to other populations or samples without further evidence.

Examples such as those presented in this paper, and the phenomena of DIF and DTF, demonstrate this limitation and the risk of simply assuming external invariance without empirical evidence. Therefore, in order to be able to claim that measurement results are invariant across different populations and/or instruments (i.e., external invariance), comprehensive research should be conducted for all relevant populations and instruments.

In recognition of the difficulties involved in carrying out valid comparisons among participants which may belong to numerous groups or populations, researchers are advised to act prudently, avoiding generalizations beyond the evidence in the study because, even if a careful analysis of DIF or DTF with regard to certain characteristics (e.g., gender) is conducted, other characteristics not considered in the analyses (e.g., age, race, language, etc.) may still interfere with measurement results.

Although the idea of internal invariance in IRT may sound less appealing than the idea of unconditional invariance, the empirical consequences of internal invariance are relevant for research. Internal invariance of IRT enables, for example, the development of computer-adaptive tests (CATs) by guaranteeing that any subset of items drawn from an item-bank for which evidence of fit exists will yield equivalent results to any other subset

of items drawn from the same bank. Naturally, a valid employment of CATs is always restricted to the population of participants for which calibrated parameters can be extrapolated, and there is no reason to believe that different groups or populations will yield the same results when assessed with items from the bank, since there may be characteristics associated with the conditional probabilities of responses to the items in the bank that need to be assessed before assuming further external validity and external invariance of the bank.

The information and evidence provided here enable us to suggest that the belief that IRT models are unconditionally invariant (i.e., internally and externally) is an example of the phenomenon of self-delusion in science (Gratzer, 2001), which illustrates the difficulties experienced by scientific communities in the construction of knowledge, where pressure for meaningful achievements occasionally produces negative impacts on critical thinking skills, and renders researchers vulnerable to misinterpretation of the evidence, a phenomenon that is potentiated when the scholars involved in the misinterpretation are highly competent and prestigious.

The phenomenon of self-delusion is not unusual in science. For instance, Feyerabend (1975) described how Galileo's detractors had valid reasons to doubt what he claimed he saw through the telescope, and how Galileo used argumentative strategies unattached to the data to convince the scientific community of his ideas. Even though later evidence demonstrated that Galileo was right, this should not obscure the fact that he could have been wrong, in which case his argumentative skills would have caused a delay in scientific progress, at least until the error had been exposed. We believe that arguments suggesting that IRT estimations are unconditionally or universally invariant are likely to fall into this same category.

Achieving true scientific measurements in social and behavioral sciences is highly desirable, and this desirability has led advocates of IRT to mischaracterize the scope of invariance in this psychometric framework, or at least to be ambiguous enough to allow misinterpretations and misconceptions about invariance in the field. While it is true that this confusion gave greater face-validity and legitimacy to IRT when compared to CTT, we believe that focusing its outreach on an incorrect interpretation of the property of invariance created more difficulties than benefits by blurring the most meaningful differences and similarities between both psychometric frameworks, generating confusion among researchers intending to prove the superiority of IRT estimations, and failing to provide clarity as to the situations and objectives for which IRT is more efficient than CTT. More importantly, however, misconceptions and misinterpretations regarding invariance in IRT have led some researchers to become overconfident in the IRT framework itself, overlooking the fact that IRT models are statistical modeling tools that do not replace the need for representative samples of all target populations and subpopulations when designing and validating tests in order to reach valid and generalizable research conclusions.

The history of science shows that long-term results arise after having clear knowledge of the limitations of measurement tools, and not as a consequence of self-delusion about their properties. Believing that IRT models are methodological tools that yield internally and externally invariant measurement results means surrendering to the beauty of the sirens' call, but just as Ulysses tied himself firmly to the mast we must resist the call for the sake of our disciplines, by employing our best critical practices.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was partially supported by the Chilean National Commission for Scientific and Technological Research (CONICYT) "Becas Chile" program (Grant N°26081114FIC and N°72120061).

References

Adedoyin, O. O., Nenty, H. J., & Chilisa, B. (2008). Investigating the invariance of item difficulty parameter estimates based on CTT and IRT. *Educational Research and Reviews*, *3*(2), 83–93.

Breithaupt, K., & Zumbo, B. D. (2002). Sample invariance of the structural equation model and the item response model: A case study. *Structural Equation Modeling*, *9*(3), 390–412. doi: 10.1207/S15328007SEM0903_5

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.

Chernyshenko, O. S., Stark, S., Drasgow, F., & Roberts, B. W. (2007). Constructing personality scales under the assumptions of an ideal point response process: Toward increasing the flexibility of personality measures. *Psychological Assessment*, *19*(1), 88–106. doi: 10.1037/1040–3590.19.1.88

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

DeMars, C. (2010). *Item response theory*. Oxford, UK: Oxford University Press.

Embretson, S. E. (1996). Item response theory models and spurious interaction effects in factorial ANOVA designs. *Applied Psychological Measurement*, *20*(3), 201–212. doi:10.1177/014662169602000302

Embretson, S. E. (1999). Issues in the measurement of cognitive abilities. In S. Embretson & S. Hershberger (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 1–15). Mahwah, NJ: Psychology Press.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Engelhard, G. (1984). Thorndike, Thurstone, and Rasch: A comparison of their methods of scaling psychological and educational tests. *Applied Psychological Measurement*, *8*(1), 21–38. doi: 10.1177/014662168400800104

Engelhard, G. (2008). Historical perspectives on invariant measurement: Guttman, Rasch, and Mokken. *Measurement*, *6*(3), 155–189. doi: 10.1080/15366360802197792

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement*, *58*(3), 357–381. doi: 10.1177/0013164498058003001

Feyerabend, P. (1975). *Against method: Outline of an anarchistic theory of knowledge*. London, UK: New Left Books.

Fischer, G. H., & Molenaar, I. W. (Eds.). (1995). *Rasch models: Foundations, recent developments, and applications*. New York, NY: Springer.

Gratzer, W. (2001). *The undergrowth of science: Delusion, self-deception, and human frailty*. Oxford, UK: Oxford University Press.

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction* (Vol. IV, pp. 60–90). Princeton, NJ: Princeton University Press.

Hambleton, R. K. (1994). Item response theory: A broad psychometric framework for measurement advances. *Psicothema*, *6*(3), 535–556.

Hambleton, R. K., & Russell, J. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12*(3), 38–47. doi: 10.1111/j.1745–3992.1993.tb00543.x

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. New York, NY: Springer.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Holland, P. W., & Hoskens, M. (2003). Classical test theory as a first-order item response theory: Application to true-score prediction from a possibly nonparallel test. *Psychometrika*, *68*(1), 123–149. doi: 10.1007/BF02296657

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Jones, L. V. (1960). Some invariant finding under the method of successive intervals. In H. Gulliksen & S. Messick (Eds.), *Psychological scaling: Theory and applications* (pp. 7–20). New York, NY: John Wiley & Sons.

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of educational statistics*, *4*(4), 269–290.

Liu, Y., & Maydeu-Olivares, A. (2013). Local dependence diagnostics in IRT modeling of binary data. *Educational and Psychological Measurement*, *73*(2), 254–274. doi: 10.1177/0013164412453841

Lohr, S. (2009). *Sampling: Design and analysis*. Boston, MA: Cengage Learning.

Lord, F. (1952). A theory of test scores. *Psychometric monographs, No.7*. Richmond, VA: Psychometric Corporation.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Mahwah, NJ: Lawrence Erlbaum Associates.

Maydeu-Olivares, A., Morera, O., & D'Zurilla, T. J. (1999). Using graphical methods in assessing measurement invariance in inventory data. *Multivariate Behavioral Research*, *34*(3), 397–420. doi: 10.1207/S15327906MBR3403_5

McDonald, R. P. (1982). Linear versus models in item response theory. *Applied Psychological Measurement*, *6*(4), 379–396. doi: 10.1177/014662168200600402

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

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika*, *58*, 525–543. doi: 10.1007/BF02294825

Millsap, R. E. (2008). Model-implied invariance in psychometrics: Be skeptical when theory suggests data are not needed. *Measurement: Interdisciplinary Research and Perspectives*, *6*(3), 195–197. doi: 10.1080/15366360802265979

Muñiz, J., & Hambleton, R. K. (1992). Medio siglo de teoría de respuesta a los ítems [Half a century of item response theory]. *Anuario de Psicología*, *52*, 41–66.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*(1), 50–64. doi: 10.1177/01466216000241003

R Development Core Team. (2012). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* (Expanded ed.). Chicago, IL: University of Chicago Press. (Original work published 1960)

Reise, S. P., Ainsworth, A. T., & Haviland, M. G. (2005). Item response theory: Fundamentals, applications, and promise in psychological research. *Current Directions in Psychological Science*, *14*(2), 95–101. doi: 10.1111/j.0963–7214.2005.00342.x

Reise, S. P., & Haviland, M. G. (2005). Item response theory and the measurement of clinical change. *Journal of Personality Assessment*, *84*(3), 228–238. doi: 10.1207/s15327752jpa8403_02

Rizopoulos, D. (2006). LTM: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, *17*(5), 1–25. doi: 10.18637/jss.v017.i05

Rupp, A. A., & Zumbo, B. D. (2004). A note on how to quantify and report whether IRT parameter invariance holds: When Pearson correlations are not enough. *Educational and Psychological Measurement*, *64*(4), 588–599. doi: 10.1177/0013164403261051

Rupp, A. A., & Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement*, *66*(1), 63–84. doi: 10.1177/0013164404273942

Thorndike, E. L. (1922). On finding equivalent scores in tests of intelligence. *Journal of Applied Psychology*, *6*, 29–33.

Thurstone, L. L. (1927). The unit of measurement in educational scales. *Journal of Educational Psychology*, *18*(8), 505–524. doi: 10.1037/h0072880

van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York, NY: Springer.

Wright, B. D. (1968). Sample-free test calibration and person measurement. In *Proceedings of the 1967 invitational conference on testing problems* (pp. 85–101). Princeton, NJ: Educational Testing Service.

Wright, B. D. (1999). Fundamental measurement for psychology. In S. Embretson & S. Hershberger (Eds.). *The new rules of measurement: What every psychologist and educator should know* (pp. 65–104). Mahwah, NJ: Psychology Press.

Author biographies

Rodrigo A. Asún is a sociologist with a doctorate in Research Methodology. He is an assistant professor at the University of Chile. His lines of research focus on latent variable modeling for categorical data, social movements, and education. Email: rasun@uchile.cl

Karina Rdz-Navarro is a sociologist with a doctorate in Research Methodology. She is an assistant professor at the University of Chile. Her lines of research focus on structural equation models and latent variable modeling for categorical and continuous data. Email: rdznavarro@uchile.cl

Jesus M. Alvarado is a psychologist with a doctorate in Psychology. He is a professor of research methods at the Complutense University of Madrid. His lines of research focus on test validity and measurement models. Email: alvarado@psi.ucm.es