Can we measure effort in cognitive tasks?*

Examining the application of Additive Conjoint Measurement and the Rasch model

James Steele

2023-01-22

Abstract

'Effort' is a concept of interest in cognitive psychology and neuroscience where many theories include it as a postulate. Despite its intuitiveness it is difficult to define such that its operationalisation follows a logical derivation chain. Recently I have proposed conceptual definitions of both actual effort, and the perception of effort, as the ratio of task demands to capacity to meet task demands, both actual and perceived respectively. Clear conceptual definitions are key for determining whether a given operationalisation meets the necessary and sufficient conditions adequately. For physical tasks valid operationalisation, and indeed measurement, of actual effort is often trivial. But a problem arises for operationalisation of actual effort in cognitive tasks where the underlying capacity that disposes an individual to be able to attempt, and perhaps complete, the task is not directly observable nor is the demand the task presents. However, a solution may lie in applications of Additive Conjoint Measurement to determine conditions where classical measurement may be possible, and the Rasch model as a measurement operation of capacity and demands to derive effort. A key aspect of the Rasch model is that it posits and estimates from data two latent constructs that I accept here as conceptually equivalent to capacity and demands respectively in my definition of effort; first a characteristic of the individual (ability), and second a characteristic of the test or item (difficulty). As such, applications of these methods might provide a measurement operation of actual effort in cognitive tasks to enable more precise formulations and testing of theories that employ the concept. In this work I explore these ideas using simulation and analogical abduction of a task where effort is known and examine an empirical dataset. Finally, I discuss the conditions under which these methods may be suitable and their inherent limitations.

1 Introduction

'Effort' is a concept of interest in a diverse range of fields, and particularly so in cognitive psychology and neuroscience where many theories include it as a postulate and empirical studies attempt to operationalise it[1]. Yet, in spite of this interest there have been few attempts to define the concept in a manner such that its operationalisation, and indeed measurement, follows a logical derivation chain. Clear conceptual definitions are important for theoretical science (Scheel et al., 2021) determining whether a given operationalisation of those definitions (i.e., our ways of defining those theoretical variables in the context of our empirical research) meet the necessary and sufficient conditions adequately for the theoretical unit of interest (Dubin, 1969). In an 'effort' to address this relative lack of thoughtful conceptual analysis many have recently begun to explore potential definitions of the concept (Bermúdez & Massin, 2023; Massin, 2022, 2017; Shepherd, 2022). In particular, I have presented formal conceptual definitions of both the actual effort required for attempted task performance, and the perception of that effort (Steele, 2020). In explicating both definitions I differentiate

*Preprint (DOI: 10.31234/osf.io/6pvht), not peer-reviewed. Invited paper submitted to Cortex special issue "Strengthening derivation chains in cognitive neuroscience". All data, materials, and code is available in the GitHub (see https://github.com/jamessteeleii/cognitive_effort) repository or Open Science Framework (see https://osf.io/yfam5/) for this project. Address for correspondence: james.steele@solent.ac.uk Affiliation: Department of Sport and Health, Solent University, UK

[1]Indeed, a cursory search within the journal Cortex for the term 'effort' yields over a thousand articles most of which are empirical research.

between the *actual* effort and the phenomenology of that effort (i.e., the conscious *experience of*, *perception of*, or *feeling of* that effort[2]).

The definition for actual effort is as follows:

> "Effort; *noun*; That which must be done in attempting to meet a particular task demand, or set of task demands, and which is determined by the current task demands relative to capacity to meet those demands, though cannot exceed that current capacity."

And more specifically following the Set Theoretical approach of defining concepts (intensional population independent) and constructs (extensional population dependent - the removal of the w term below renders the definition of the construct) from Markus (2008):

> "Effort (concept); $E_A(p, t, C_A, D_A, w, T_{Any})$ is the actual effort for any individual p at time t where $C_A(p, t, x_C)$, and $D_A(i, t, x_D)$ are the actual capacity and actual demands respectively, and x_C and x_D are the magnitudes of those respectively for individual p at time t, where w denotes all possible states of affairs (i.e. combinations of p, t, C_A, and D_A), and T_{Any} denotes the boundary conditions noting it as intensional to all possible types of tasks."[3]

Where the subscript A denotes that it is the *actual* effort. In this sense the demands of a task and the capacity to meet those demands of an individual actor are both necessary and sufficient conditions for effort: a task imposes demands if it is attempted and, while task demands exceeding capacity can be *attempted*, capacity must be some positive value greater than zero to allow the actor to even make an attempt i.e., the capacity is the disposition to be able to attempt the task.

This can be expressed as a derived ratio (or even a percentage given how common it is to think of effort in this manner e.g., "they gave X% in that attempt") given that I conceptualise the primitives, capacity and demands, as being quantitative attributes having natural origins[4] and hypothesised to relate to magnitudes of one another ordinally and additively (Michell, 1997). The effort E_{Api} for person p and task i expressed as a percentage then is:

$$E_{Api} = \frac{D_{Ai}}{C_{Ap}} \times 100, \text{ if and only if } D_{Ai} \leq C_{Ap}; \text{ else } E_{Api} = 100\% \tag{1}$$

Indeed, it could be said that effort is merely a measure of the demands of a task given in capacity units.

Now, the T_{Any} argument is important for what I am about to present. The conceptual definitions I have proposed are at this stage deliberately agnostic of the type of task being performed. To derive a measurement of effort we merely need to perform a measurement operation of a persons capacity to perform a given task, and one for the demands that the task being attempted presents. However, in practice it is not easy to apply the definition to all kinds of tasks in terms of some measurement operation. For attempted performances of cognitive tasks in particular, which are thought to be behavioural reflections of relations between unobservable psychological dispositions of the actor and unobservable aspects of the task, operations of measurement are fraught with difficulty (Michell, 1997). In essence, it is difficult to derive the effort for a given individual attempting a given task because we haven't got good operationalisations of the magnitudes of C_A and D_A.

[2]Note, in my conceptual analysis paper (Steele, 2020) I use the terminology *objective* and *subjective* to differentiate *actual* efforts from the *phenomenology of* efforts. However, recent discussion from Bermudez and Massin (2023) have convinced me that this language is inappropriate and I intend to change this moving forward, and will update the conceptual analysis paper in due course to reflect this.

[3]I'll use the subscript p here to denote the individual (i.e., 'person') instead of i as I do in the original paper, to be in keeping with the typical notation used in Item Response Theory models that will follow i.e., the i subscript is used for the 'item'.

[4]The capacity to meet task demands can be zero which would mean that it is impossible to even attempt the task; the person does not have the dispositional characteristic of being able to perform the task. Demands too can be zero but in this case such a condition is equivalent to their being no task to attempt. For effort to be present though neither can be less than zero; an actor has to have some disposition towards capacity to perform a task, which they attempt, and by dint of attempting a task that task must represent some demands.

In some fields however the measurement of C_A and D_A is in fact somewhat trivial. Let us consider an example of a physical task common to research in the sport and exercise sciences and one which has constituted the experimental paradigm for much of my work[5]; lifting a weight i.e., resistance training. For a physical task such as this it is simple to think about how one would go about exploring effort. I gave an example in Steele (2020) of the definition in play for such a task:

> "In a physical task the role of differential demands and capacity are easily considered in that actual effort is determined by the task demands relative to the current capacity to meet task demands. As such, if two individuals were attempting to pick up the same specific absolute load (e.g. 80 kg) the stronger of the two would initially require less actual effort to complete this task. If they had both performed prior tasks that had resulted in a reduction in their maximal strength, then each would require a greater actual effort to complete the task than compared with when their capacity was not reduced. And further, if both continued performing repetitions of this task their maximal strength might continue to reduce insidious to continued attempts to maintain a particular absolute demand, and thus require an increasingly greater actual effort with every individual or continued attempt to meet the task demands. Correspondingly, if the absolute task demands were increased then both individuals would also require greater actual effort to complete the task. Yet for both, the continued attempted performance of the task with fixed absolute demands and insidious reduction of capacity, or the increase of absolute demands, task performance would be capped by their maximum capacity at which maximum effort is required. With training though that maximum strength might be increased such that a given absolute task demand now represents relatively less and so requires less actual effort. Further, biomechanical alterations to the task might reduce the absolute demands and thus the actual effort."

We in essence have a *causal* system here involving three variables (C_A, D_A, and E_A) akin to classical measurement (Stenner et al., 2013), a kind of derived measurement (Campbell, 1920), wherein we can trade off the causal effects of manipulating either the task demands (i.e., we could make the weight lighter or heavier) and the persons capacity (i.e., we could have a stronger and weaker individual each perform the task, or fatigue or train an individual and have them perform the task again) in a manner that produces known effects on the effort required to perform the task. For example, a person (p) attempts to lift a load (i) in a particular exercise (e.g., elbow flexion) that was 10 kg, and further the maximum load the person could actually lift in this task was 20 kg (often referred to as maximal strength and measured using a one-repetition maximum test; 1RM). From this it is simple to calculate the actual effort required:

$$D_{Ai} = 10, C_{Ap} = 20$$

$$E_{Api} = \frac{D_{Ai}}{C_{Ap}} \times 100$$

$$50\% = \frac{10}{20} \times 100$$

So, the amount of actual effort required by the person to lift the load in this task is 50%. If the person were twice as strong it would be 33.33%, if the load were halved it would be 25% etc. The task of deriving a measurement of effort here is simple because, in the case of physical tasks such as lifting a weight, we have knowledge of the quantitative nature of the attributes of capacity and demands given that they are physical variables and are amenable to classical measurement, and we have available to us operations for measuring them i.e., the numerical estimation of the ratio of their magnitudes to a unit of the same attribute (e.g., kilograms; Michell (1997)).

[5]The benefit, or perhaps bias, of having begun my interest in the topic of effort from the vantage point of being a sport and exercise scientist should not be overlooked in the context of discovery. The analogies that I have been able to draw upon from this field, where it is almost trivial to employ my conceptualisations of effort, have of course shaped my approach to the topic here.

We can conceivably apply my definition to cognitive tasks if we assume that such tasks present some demands that must be met if attempted, and that we have some capacity to meet them. I offer a similar set of parallel examples for them (Steele, 2020):

> "Similar examples could be provided for cognitive tasks. For example, if two individuals were attempting to hold a fixed number of items in their working memory, the one who has the larger working memory of the two would require less actual effort to complete this task. However, both individuals would again require greater actual effort to do so in the presence of lingering reduction in cognitive capacity from prior tasks, or from continued attempts to meet the task demands, or from increased absolute task demands (i.e., more items to be held in working memory). Again, training may also improve maximal capacity. Also, cognitive processing alterations (i.e., heuristics; Shah & Oppenheimer (2008)) might reduce task demands and thus the actual effort."

The problem however is actually measuring the capacity being used to perform specific cognitive tasks, and the demands of those tasks. There has been debate as to whether such attributes are even quantitative at all that continues to this day (Franz, 2022a, 2022b; Michell, 2022; Tafreshi, 2022; Trendler, 2022) which if they are not would render the derivation of effort (at least based upon my conceptual definition) impossible[6]. Unlike many physical tasks where operationalisation is fairly trivial and indeed even key assumptions about whether capacity and demands are quantitative are widely accepted, it is not quite so simple for tasks where the underlying capacity that disposes an individual to be able to attempt and perhaps complete the task is not directly observable nor is the demand that the task presents, and indeed it is still an untested assumption as to whether either are quantitative in the first place.

The problem of deriving the effort required for cognitive tasks is not trivial. This is perhaps why typically only the task demands themselves are considered and manipulated in experimental work in cognitive psychology in an ordinal fashion (i.e., task A is harder than task B which is harder than task C etc. but it is not exactly clear by what magnitudes they differ). But this misses the crucial element of the individual actor themselves and the plain intuition that many have; some tasks require more or less effort dependent on who is attempting them.

How might we overcome the barriers to derivation of effort in cognitive tasks? That is the question that I address in the present work. Given the nature of my conceptual definition of effort I speculate that a possible solution lies in existing psychometric theory; the paradigm typically referred to as Item Response Theory (IRT).

2 Item Response Theory

I will provide a very brief introduction to key aspects of IRT relevant to this exposition. IRT is typically taken to be an improvement over Classical Test Theory in modern psychometrics due to its ability to, for instance, equate tests, explore item bias, and develop computer adaptive tests. IRT models assume that a persons response to each item of a test can be related to an underlying latent trait or disposition e.g., an 'ability'. It is also assumed that the items possess at least one underlying latent variable typically referred to as its 'difficulty'[7]. The notation used for the two key parameters of these models is θ_p for a given person p's ability, and β_i for a given item i's difficulty. Both person and item latent variables are assumed to consist of the same attribute with instances of this attribute for both being comparable on the same underlying uni-dimensional scale. Further, it is assumed that responses to items are locally independent once the effect of the latent person variable is accounted for. The probability that a given person will respond in a given way (e.g., answering an item correctly or incorrectly) is specified by a link function expressing a monotonic relationship between their ability and the items difficulty. However, different IRT models assume different

[6]Though this may not be an issue for those with a cost-based conceptualisation of effort (Westbrook & Braver, 2015)

[7]Sometimes it is referred to as 'easiness' and the parameter is expressed with an opposite sign. I will use 'difficulty' here however to be in closer keeping with the 'demands' terminology used in my conceptual definition.

functional forms for this relationship which can be deterministic or stochastic, discrete or continuous, linear or non-linear.

Generally speaking, the application of an IRT model allows for test data to be decomposed into an estimate of a characteristic of the individual (i.e., their 'ability', θ_p). Additionally, the other parameters of such models which reflect the item characteristics can also be estimated such as the items location on the scales (i.e., their 'difficulty', β_i). It is these two parameters that interest me and which I will draw upon. Two of the most simple and closely related models both incorporating ability and difficulty parameters are the Guttman (1944) and Rasch (1960) models.

The Guttman model is most restrictive according to which a person is successful in attempting a particular item if and only if their level of ability at least matches the difficulty of the item. In essence:

$$P_{pi}(success) = 1 \text{ if and only if } \theta_p \geq \beta_i;\ \text{else } P_{pi}(success) = 0 \tag{2}$$

Where $P_{pi}(success)$ denotes the probability of a person p successfully performing the item i attempted. The relationship between the probability of success and a persons ability level is represented by a step function (see figure 1A) and this model is said to describe a deterministic process.

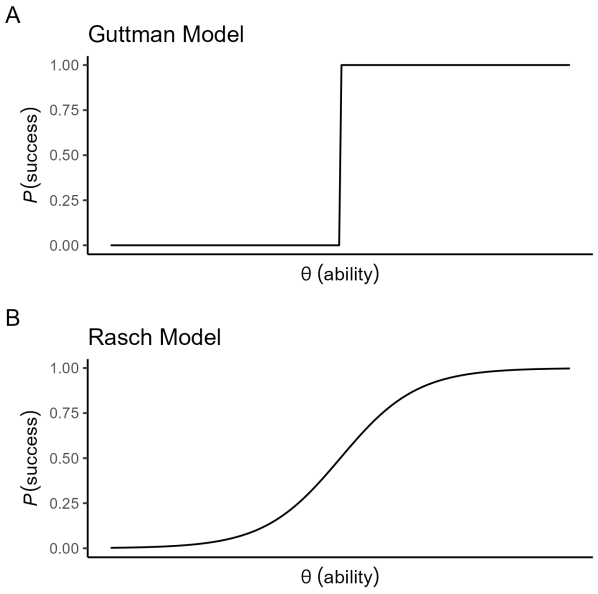

Figure 1: Example response functions for (A) Guttman and (B) Rasch models.

The Rasch model is essentially a stochastic version of the Guttman model where the relationship between the probability of success and a persons ability level is represented by a logistic function (see figure 1B). It is represented formally as:

$$P_{pi}(success) = \frac{e^{(\theta_p - \beta_i)}}{1 + e^{(\theta_p - \beta_i)}} \tag{3}$$

Perhaps the parallels between my conceptual definition of effort and these IRT models are already clear to the astute reader. In essence, the argument I will develop here is that the two primitives C_{Ap} and D_{Ai} which I have assumed are necessary and sufficient for the derived measurement of the concept E_{Api} are exactly what IRT models such as the Guttman and Rasch models take as their own underlying assumed parameters; namely θ_p and β_i respectively. I am proposing that there is extensional equivalence between the concepts

of 'capacity' and 'ability', and 'demands' and 'difficulty' and stipulating that they are merely idiosyncratic synonyms[8].

I suggest that we may be able to use specific IRT models in order to estimate θ_p and β_i and then use them to calculate an estimate of the E_{Api} required for each individual as they attempt each item[9]. That is to say, we can take our function for E_{Api} above and say in the case of an IRT model:

$$E_{Api} = \frac{\beta_i}{\theta_p} \times 100, \text{ if and only if } \beta_i \leq \theta_p; \text{ else } E_{Api} = 100\% \tag{4}$$

However, as explained, I postulate that the capacity/ability to perform a task and the demands/difficulty that a task represents should be quantitative attributes if we are to be able to perform an operation yielding measurement of them in order to derive measurement of the effort for required for a given person performing a given task. As such, whilst there may be little to disabuse the idea of treating θ_p and β_i as extensionally equivalent to C_{Ap} and D_{Ai} in principle, there may only be specific empirical conditions under which the application of specific kinds of IRT models might yield measurement such that we can indeed derive E_{Api}.

3 Classical measurement in science and application of additive conjoint measurement

I should note at this stage for clarity that I adopt a realist ontology regarding the existence of theoretical concepts (McMullen, 2011) including those theorised and described up to this point (i.e., capacity/ability, demands/difficulty, and effort), and adopt a similar view with respect to the task of measurement following Michell (1997). The entities I have postulated are defined such that I hypothesise them to sustain a quantitative structure; that is to say that instances of them are related to one another ordinally and additively. Further, this ontological stance also entails a commitment to the existence of numbers[10]. As such, I am behooved to endeavour to engage with the scientific task of discovering whether, and perhaps when and where, this is indeed the case with my postulated entities such that classical measurement of them is possible and we can engage in the instrumental task of constructing a measurement operation.

Classical measurement as it pertains to the fundamental measurement of extensive quantities involves the demonstration of concatenation of objects to be measured, the comparison of effects of arbitrary combinations of quantities of a single specified kind (Campbell, 1920), and was first axiomatized by Otto Hölder in 1901 (see translation by Michell & Ernst (1996)). However, Luce and Tukey (1964) introduced a new axiomatized system of fundamental measurement based on the observation of simultaneous conjoint entities and the comparison of effects of pairs formed from two specified quantities; additive conjoint measurement (ACM). This new type of fundamental measurement guaranteed that interval scales can be constructed for systems of three variables where one variable is a non-interactive[11] function of two other variables and where its axioms are met, and which usually can be converted naturally to ratio scales (Krantz et al., 1971; Luce & Tukey,

[8] I just happened to use 'capacity' and 'demands' when forming my conceptual definitions prior to giving much thought to the possibilities of IRT models to explore applications in cognitive tasks.

[9] Note, it should hopefully be obvious that I am using 'item' also as synonymous with 'task'. Saying a person attempted an item on a test is no different than saying that they attempted a task on a test. In fact, my conceptualisation of a task is in line with the definition from Kunzell et al. (2018). They note a task iss impersonal and merely represents "what has to be done" by any participant given the attempted performance of it. It is only once a given person accepts that task as a goal, and indeed performs the action of attempting its performance, that the interaction between the person and task gives rise to the actual effort required for that person attempting that task.

[10] Reports of measurements under realism are considered to be true if and only if things are literally as reported. Consider this example, adapted from Michell (2005) and Domingue (2014); at a given time and place we report a barbell to weigh 100 kg. This is true if and only if the weight of the barbell at that time and place really is 100 kg. But, for this to be true the number '100' must be something actually existing. This differs from the representationalist view that numbers are Platonic entities. Within classical measurement, which is consistent with a realist view, *numbers* are thought to actually exist and the labels we give to them are typically referred to as *numerals* instead (Michell, 1990).

[11] Transformations of the function (f) are permissible here. For example, if $f(x, y) = \frac{x}{y}$ then $\log(\frac{x}{y}) = \log(x) - \log(y)$ based on the Quotient rule and yields an acceptable transformation.

1964). Consider two natural attributes X and Y where it is unknown if either are continuous quantities. Let x_1, x_2, and x_3 be independent identifiable levels of X, and let y_1, y_2, and y_3 be independent identifiable levels of Y. A third attribute Z is a function f of X and Y consisting of the nine ordered pairs of the levels of X and Y i.e., $f(x_1, y_1), f(x_2, y_1), \ldots, f(x_3, y_3)$. Quantification of the attributes depends on the relations holding on levels of Z. These are presented as the following axioms:

1. Independence (or single cancellation)

- If $f(x_1, y_1) \geq f(x_2, y_1)$ then for all y_2 in Y $f(x_1, y_2) \geq f(x_2, y_2)$.

2. Double cancellation

- If $f(x_1, y_2) \geq f(x_2, y_1)$ and $f(x_2, y_3) \geq f(x_3, y_2)$ then $f(x_1, y_3) \geq f(x_3, y_1)$.

3. Solvability

- For all x_1 in X, y_1, y_2 in Y, there exists x_2 such that $f(x_1, y_1) = f(x_2, y_2)$.

4. Archimedean condition

- No value of a quantitative variable is infinitely larger than any other value ensuring comparability between any two values given they can only be a finite distance apart.

Michell (1990) gives a more comprehensive, yet mathematically simplified, introduction to ACM and what these axioms entail exactly which the reader is referred to for more detail. The final two axioms, due to involving infinitistic concepts, are impossible to directly verify in most cases[12] and as such are typically assumed[13].

IRT models may present possible operations on conjoint systems of persons, items, and probabilities that might yield fundamental measurement. The Guttman and Rasch models already introduced involve systems of three variables: ability, difficulty, and the probability of a successful response[14]. However, given the axioms noted, only one of these (in fact the only IRT model that appears to meet ACMs axioms) is compatible.

4 Guttman, Rasch, and additive conjoint measurement

Considering the probabilities of success for a person with ability θ_p and item with difficulty β_i we might observe tables looking something like tables (1 and 2). In the Rasch model, the axioms of ACM are satisfied merely by the fact that its mathematical form incorporates them. Thus, assuming data fit the model, the resulting scale meets the requirements for interval level measurement allowing us to not only order persons and items, but also to interpret differences between them. In fact, many consider the Rasch model to be a stochastic version of ACM (Bond et al., 2020; Brogden, 1977; Embretson & Reise, 2000; Fischer, 1995; Fischer & Molenaar, 1995; Perline et al., 1979; Rasch, 1960). However, this does not mean that the direct checking of the axioms of ACM should not be performed on the data itself as indeed data can fit the Rasch model whilst seemingly violating the axioms (Domingue, 2014).

[12]For example, as Michell (1990) explains, finding x_2 that satisfies solvability may be impractical due to resource constraints. Also, the steps to show a certain value is bounded and restricted by the Archimedian condition may be practically impossible; for example, demonstrating the sun is a finite distance from the earth using a common 1 foot ruler (Domingue, 2014).

[13]Though Scott's (1964) finite set of cancellation conditions can be used to directly test these axioms, though they are still empirically determined regarding the extent to which they test them.

[14]This to some extent depends on whether the probabilities involved are seen to be real empirical objects (Borsboom & Scholten, 2008), though as a realist I am considering them in their conceptualisation as part of the object language as opposed to the meta-language concept (Carnap, 1945).

Table 1: Example of success probabilities for a Guttman model

	β_1	β_2	β_3
θ_1	1	0	0
θ_2	1	1	0
θ_3	1	1	1

Table 2: Example of success probabilities for a Rasch model

	β_1	β_2	β_3
θ_1	0.5000000	0.2689414	0.1192029
θ_2	0.7310586	0.5000000	0.2689414
θ_3	0.8807971	0.7310586	0.5000000

The Guttman model however can only yield at best the ordering of persons and items on the latent attribute, meaning their differences cannot be compared meaningfully. This does not mean that the attribute necessarily lacks quantitative structure, but the Guttman model can not yield a quantitative scale if it does. Why the Guttman model does not meet the axioms of ACM can be shown with the case of double cancellation called the Luce-Tukey condition (Luce & Tukey, 1964; Michell, 1988) (see figure 2). Formally, if and only if the antecedents $(x_2, y_1) \geq (x_1, y_2)$ and $(x_3,_y 2) \geq (x_2, y_3)$, then the consequent is $(x_3, y_1) \geq (x_1, y_3)$. Considering table 2 we can see this in the Rasch model probabilities. However, double cancellation will be violated in many of the 3×3 matrices that are possible for Guttman probabilities.

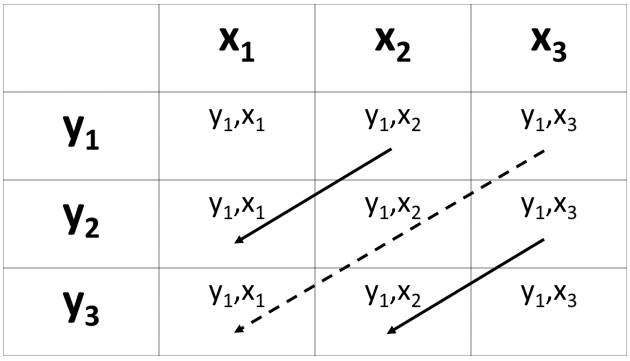

Figure 2: A Luce-Tukey condition of double-cancellation. The consequent inequality (the broken line arrow) does not contradict the direction of both antecedent inequalities (solid arrows). Reproduced from Wikipedia.

It seems then that the only current option available to us for the purpose of measurement of person ability/capacity and item difficulty/demand, such that we can derive the effort for a person attempting a cognitive task, is to look to situations where data meet the axioms of ACM and conform to the Rasch model. In essence, we need to operate in the confirmatory approach of the Rasch paradigm testing the contexts in which the Rasch model is not falsified. Where data can be shown to pass both checks of the assumptions of ACM, and the Rasch model provides a good fit to the data, it is reasonable to operate under the assumption that the hypothesis of quantitative structure has been tested and corroborated and that the resulting scale provided by the model yields interval properties. I will briefly demonstrate through simulations, and an empirical dataset, an application of this in the derivation of effort. But, before doing so we must consider the scale itself that the Rasch model yields and its admissibility in deriving effort as interval scale data do not have the characteristics that they can be expressed as ratios, and effort as noted could be said to essentially be a measure of the demands of a task given in capacity units i.e., the ratio of demands to capacity.

5 Scaling in the Rasch model

We face an initial problem here for my proposed solution to a measurement operation to derive effort; β and θ are typically estimated in the Rasch, and other IRT models, on the *logit* scale which ranges $(-\infty, \infty)$ and thus does not have ratio properties[15]. Fortunately, the choice of which scale to place β and θ on is somewhat arbitrary.

Most common IRT models can undergo transformations from the *logit* scale that both preserves the underlying probabilities for a given person with ability θ_p completing a given item with difficulty β_i. For example, the Rasch model can undergo classical linear transformation methods, such as regression or the mean-sigma method, without altering the underlying mathematical model (Hambleton et al., 1991). These approaches however rely on some other quantitative scale with which to anchor upon.

Another approach maybe to consider the odds formulation of the Rasch model (Freund, 2019). When using the logit (i.e., log odds) form of the Rasch model the absolute difference nature of the interval scale formed by the logit unit makes the value of the subtractive difference between two respondents' estimated ability values (e.g., a 1 logit difference) conceivable as an increase in predicted log odds of success for any item. By exponentiating to the odds form of the Rasch model estimates can be compared through ratios, which provides potential justification for conclusions such as one person has twice the ability of another. Alternatively it has been argued that the logit scale itself might be used to provide a derivation of effort from ability and difficulty estimates merely by their difference (i.e., $\beta_i - \theta_p$; Ehrich et al. (2021)). The difference between β and θ could provide a measurement of effort, albeit on the logit difference scale. Based on the Quotient rule though there is an equivalence between these two methods.

A final approach I will consider here I refer to as the 'logit-shift' and involves considering the range of empirical estimates for the Rasch model across the logit scale and 'shifting' them to re-scale the lower bound to zero and artificially create an absolute ratio scale (see figure 3). Similar approaches have been used with conjoint measurement models (Fisher et al., 1994). However, this transformation requires us to make certain assumptions about the nature of the items and persons used to develop the Rasch measurement model and estimate their parameters. A range that as near as possible approaches the limit of the easiest possible task, though does not result in a Guttman-like response pattern, is likely needed. Further, a wide evenly spaced range of items enhances the precision of the measurement (Scholten, 2011).

Some have questioned, given that the Rasch model is in essence the Guttman model with the addition of measurement error, whether this leads to a paradox i.e., that by adding error we can move from mere ordering to an interval scale (Michell, 2008, 2014, 2009), or even a ratio scale. However, even in the physical sciences it has been shown that the addition of error can in fact improve measurement precision and the same phenomenon occurs in the case of the Rasch model demonstrated by manipulation of the so called 'discrimination' of items i.e., α (I will vary this assumption in simulations explored in the next section). Indeed, applications of the Rasch model have been used to recover the scales of physical quantities such as length and density (Pelton & Bunderson, 2003; Stephanou & Fisher, 2013). Such quantities are known to have quantitative structure and the properties of ratio scaling. Given the value of analogical abduction in what has been referred to as the 'prototheory' stage of theory development (Borsboom et al., 2021), and the insights that metaphors have been argued to yield in psychometrics (Bramley, 2020), these examples of scale recovery in physical quantities coupled with my earlier example of the physical task of lifting a weight offers an interesting simulation to explore.

6 Using additive conjoint measurement and Rasch models with simulated weight lifting tasks

Returning to the analogy earlier of a physical task involving lifting of weights, we can extend this to explore through simulation the application of ACM and Rasch models for deriving effort. As noted, in such physical

[15] Ontologically, whilst as intensional concepts infinite sets might be fine with regards definition, it seems odd to think of the attributes of ability and difficulty as extensional constructs as having no lower floor or origin.

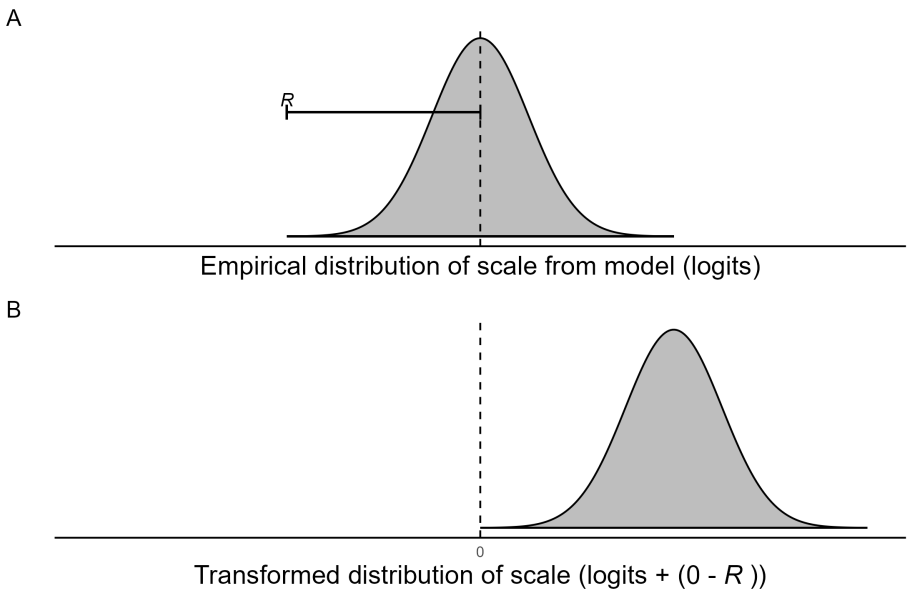

Figure 3: Shifting the range of empirical logit values (A) from empirical estimates to artificially create an absolute ratio scale (B).

tasks it is trivial to measure the difficulty of a task (we merely weigh the loads to be lifted with a set of scales), and also to measure the ability an individual has in that particular task (we can perform a 1RM test as noted earlier). If we know a particular persons strength capacity and the load they are attempting to lift we can quite simply derive the effort required for that attempt. Further, given that the outcome of attempting to lift a given load can be considered in a binary manner (that is to say a person either can, or cannot, lift the load given their strength ability), we could fit a Rasch model to such data. Practically speaking this kind of example is highly likely to exhibit a Guttman-like pattern. If we know the maximum load that an individual can lift once, then it's a good assumption to think that this means they can also lift any load weighing less than this. If their 1RM was 100 kg and we asked them to lift 50 kg they'd almost certainly be able to. If we asked them to lift 90 kg, whilst it would be a lot more demanding to do so, they'd still almost certainly be able to do so. But if we asked them to lift 110 kg they'd almost certainly not be able to do so. In reality, it is not guaranteed that a given individual will necessarily be able to lift a load at a given time as small fluctuations in other factors might impact the probability that they succeed (e.g., motivation, technique alterations etc.). For the purpose of these simulations though it does not matter a great deal as the primary point of interest is scale recovery and derivation of effort under conditions where we actually know the underlying capacity/ability and demands/difficulty to compare it to[16]. Thus, I conducted simulations of weightlifting tasks where it was possible to know the actual effort required for attempting the task of lifting a given load for a given person and explored testing of the axioms of ACM, and the use of Rasch models to produce a measurement operation in order to recover the actual effort required.

6.1 Simulated datasets

A variety of datasets representing varying assumptions were simulated. This included varying of the error included (no errors and thus a deterministic Guttman-like response pattern; a Rasch response with normally

[16]If preferable for the analogy consider that the response of success or failure is somewhat determined by the observer and it is not 100% clear to them whether or not a person has indeed succeeded or not. For example, in powerlifting competitions judges determine whether a given lift was successful or not given the required conditions of the task. Scholten (2011) gives a similar example using the Rasch model to recover lengths where dependent on the person eyeballing the lengths and comparing pairs of them there will be more or less error (represented by the α parameter).

distributed error and $\alpha = 1$; a Rasch response with normally distributed but substantial error and $\alpha = 0.1$; and a Rasch response with normally distributed but low error and $\alpha = 5$), the sample size of persons (50, 100, 250, 500, and 1000) and items (i.e., loads; 5, 10, and 20), and the range over which the items were spaced (narrow ranging from 60 kg to 140 kg, and wide ranging from 10 kg to 200 kg). For each person and item pair I simulated 500 trials (with the exception of for the deterministic model). Person strengths (i.e., 1RMs) where simulated as sampled from a $Normal(100, 25)$ distribution. The probabilities for the trials were calculated using the Rasch model equation as noted albeit with the inclusion of the *alpha* parameter as I varied this:

$$P_{pi}(success) = \frac{e^{\alpha(\theta_p - \beta_i)}}{1 + e^{\alpha(\theta_p - \beta_i)}} \tag{5}$$

The differences between the person strength and item load (θ_p - β_i) is related to the natural log of the ratios of true strengths (S_p) and loads (L_i) by the Quotient rule. The log of the strength of a person (θ_p) corresponds to their strength 'ability' and the log of the load of an item (β_i) corresponds to an items 'difficulty' on the logit scale:

$$log(\frac{S_p}{L_i}) = log(S_p) - log(L_i) = \theta_p - \beta_i \tag{6}$$

And so the probability function of the Rasch model can be restated in terms of strengths and loads as:

$$P_{pi}(success) = \frac{S_p^\alpha}{S_p^\alpha + L_i^\alpha} \tag{7}$$

In order to simulate responses from the simulated strength and load values on the kilogram scale.

6.2 Checks, transformations, and analyses

The presence of a Guttman-like response in the simulated datasets was checked by calculating the Coefficient of Reproducibility (C_R; Guttman (1944)) for the first trial of each dataset. Given that it was known *a priori* that the Guttman dataset would not meet the axioms of ACM, and further that the Rasch model cannot be fit the perfect Guttman-like responses[17] I did not conduct further analysis with these datasets. For the remaining Rasch datasets I first checked the cancellation conditions for the first trial of each using the *ConjointChecks* package which employs a Bayesian sampling approach to testing the axioms (Domingue, 2014). With a 95% credible region it would be expected that $\leq 5\%$ of checks would show violations if their really is quantitative structure. Both unweighted estimates of the proportion of violations, and estimates weighted to the cell sample sizes, were examined. Following this I fit Rasch models within a generalised linear mixed effects model framework using the *lme4* package and maximal likelihood (Boeck et al., 2011; Doran et al., 2007; Lamprianou, 2013) to each dataset[18].

From each model the estimates of $\hat{\theta_p}$ and $\hat{\beta_i}$ were rescaled such that they could be compared directly to the original strength and load values simulated. Anchoring on both the strengths and loads separately a mean-sigma transformation was used to transform the estimated log scales to the true log scale and then exponentiated to generate estimates on the true strength and load scale (i.e., back to kilograms):

[17]There is no maximum likelihood solution for perfect responses for the Rasch model. Various approaches have been suggested to handle the existence of perfect responses where these are limited in number. However, Bayesian estimation methods may be a possible solution (Swaminathan & Gifford, 1982; Wright, 1986). I considered refitting the models used in the simulations in this paper using Bayesian hierarchical regression models for IRT (Bürkner, 2020) however due to the computational time opted to not do so for now (I needed to meet the already extended submission deadline!). I may explore whether Bayesian estimation can yield scale recovery well even in the presence of perfect response patterns at some future time.

[18]Note, as the data have been simulated from the Rasch model I do not examine any fit statistics here, save for the aforementioned check of Guttman-like patterns cropping up by chance. In the empirical example later however I do explore the fit of the data to the Rasch model.

$$\hat{S}_p = e^{(((\hat{\theta_p} - \hat{\mu_p})/\hat{\sigma_p})\sigma_p + \mu_p)} \tag{8}$$

$$\hat{L}_i = e^{(((\hat{\beta_i} - \hat{\mu_p})/\hat{\sigma_p})\sigma_p + \mu_p)} \tag{9}$$

Where $\hat{\mu_p}$ and $\hat{\sigma_p}$ describe the distribution of strengths or loads on the estimated log kilogram scale, and μ_p and σ_p describe the distribution of strengths or loads on the true log kilogram scale. Both strengths and loads were used for anchoring in order to compare the impact they have on scale recovery as with real datasets we can only know in advance some potential scale with respect to the items (e.g., the integer of an N-Back test) and the aim of the test is to estimate a persons ability. Anchoring on items, loads in this case, then allows me to explore whether we can recover person abilities, strengths in this case, from them on the same scale. The mean of the absolute deviation proportion (MADP) was calculated for the difference between the true values (i.e., those simulated) and the estimated values both on the kilogram scale and presented as a percentage for both N strengths over each trial, and K loads:

$$MADP_S = \left[\sum_{n=1}^{N} \frac{abs(S_p - \hat{S}_p)}{S_p} \right] / N \tag{10}$$

$$MADP_L = \left[\sum_{k=1}^{L} \frac{abs(L_p - \hat{L}_p)}{L_p} \right] / K \tag{11}$$

Lastly, for each dataset and model the actual effort E_{Api} was calculated for each person attempting each load based on the simulated strengths and loads as the capacities and demands respectively:

$$E_{Api} = \frac{L_i}{S_p} \times 100, \text{ if and only if } L_i \leq S_p; \text{ else } E_{Api} = 100\% \tag{12}$$

This was then compared with the previously mentioned means of scaling to derive effort from the Rasch model parameter estimates for each pairing of person and item (i.e., strength and load).

The derivation of effort as a percentage from the mean-sigma transformations of the scales for strength and load back to kilograms:

$$E_{Api,mean-sigma} = \frac{\hat{L}_i}{\hat{S}_p} \times 100, \text{ if and only if } \hat{L}_i \leq \hat{S}_p; \text{ else } E_{Api,mean-sigma} = 100\% \tag{13}$$

The odds ratio from the odds formulation of the Rasch model:

$$E_{Api,odds} = \frac{e(\hat{\beta}_i)}{e(\hat{\theta_p})} \times 100, \text{ if and only if } \hat{\beta}_i \leq \hat{\theta_p}; \text{ else } E_{Api,odds} = 100\% \tag{14}$$

The difference in logits between strength and load estimates, except here in comparison with the approach of simply taking the difference $\beta_i - \theta_p$ as per Ehrich et al. (2021), I take the $\theta_p - \beta_i$ difference including a conditional statement such that if the difficulty is greater than the ability we estimate effort on the logit difference scale as being zero (i.e., equivalent to 100% effort) and take the inverse so that the tendency is for increasing logit differences for more difficult items:

$$E_{Api,logit-diff} = (\hat{\theta_p} - \hat{\beta}_i) \times -1, \text{ if and only if } \hat{\theta_p} \geq \hat{\beta}_i; \text{ else } E_{Api,logit-diff} = 0 \tag{15}$$

And finally, the logit-shift method rescaling the logit estimates to have an origin at zero based on the minimum value (R) of either $\hat{\beta}_i$ or $\hat{\theta_p}$ and then deriving effort from this:

$$R = min(\hat{\theta_p}),\ \text{if and only if } min(\hat{\theta_p}) < min(\hat{\beta_i});\ \text{else } min(\hat{\beta_i}) \tag{16}$$

$$\hat{\theta_{p,logit-shift}} = \hat{\theta_p} + (0 - R) \tag{17}$$

$$\hat{\beta_{i,logit-shift}} = \hat{\beta_i} + (0 - R) \tag{18}$$

$$E_{Api,logit-shift} = \frac{e(\hat{\beta_{i,logit-shift}})}{e(\hat{\theta_{p,logit-shift}})} \times 100,\ \text{if and only if } \hat{\beta_{i,logit-shift}} \leq \hat{\theta_{p,logit-shift}};\ \text{else } E_{Api,odds} = 100\% \tag{19}$$

I explored the relationships visually between the actual effort calculated from known strengths and loads (i.e., equation (12)) and each of the approaches to deriving effort from the Rasch model parameter estimates. For the approaches to deriving effort that yielded ratios of demands to capacity (i.e., equations (13), (14), and (19)) such that they were on the scale $[0, 100]$ comparably with equation (12), I also calculated the MADP for the efforts calculated for each strength and load pairing Q between:

$$MADP_E = \left[\sum_{q=1}^{Q} \frac{abs(E_{Api} - \hat{E}_{Api})}{E_{Api}} \right] / Q \tag{20}$$

Where $\hat{E}_{Api}$ is the derived effort (i.e., equations (13), (14), or (19)) dependent on which scaling procedure is used.

6.3 Results

6.3.1 Coefficient of reproducibility and cancellation checks

The C_R values for the simulated Rasch datasets were all below 0.9, except in the case of low error (i.e., $\alpha = 5$) and a wide range of loads (i.e., 10 to 200 kg), suggesting that perfect Guttman-like patterns had largely been avoided (see https://osf.io/rpnmh). When checking the cancellation axioms of ACM in the datasets it was evident at larger sample sizes and typically lower item numbers there was evidence of quantitative structure in the data. Indeed, for low error models at larger sample sizes this was very clear (see figure 4 for mean violation proportions and https://osf.io/6nyjt for individual item proportion plots). Given that the datasets were simulated to deliberately involve variables of quantitative structure conforming to the Rasch model this is encouraging prior to application of the Rasch model and scale recovery approaches as measurement operations. However, it does suggest that detecting quantitative structure when it is indeed present (as was simulated here) in the data may require large samples and low error tests though with fewer items to most reliably detect it, something to consider in applications to empirical data.

6.3.2 Recovery of the true kilogram scale

The mean-sigma approach to recovering the true scale values in kilograms for the Rasch model estimates of ability and difficulty yielded results in line with those seen in prior investigations of recovery of physical quantities (Pelton & Bunderson, 2003; Scholten, 2011; Stephanou & Fisher, 2013). The accuracy of estimates was typically increased where sample size was larger, item number was larger albeit spread across a narrow range of difficulties about the mean of person ability, and with lower errors. Figures (5) and (6) show the scale recovery and the MADPs for both person strength values and item load values with anchoring based on the strength values, and figures (7) and (8) based on the load values. With wide load ranges estimates

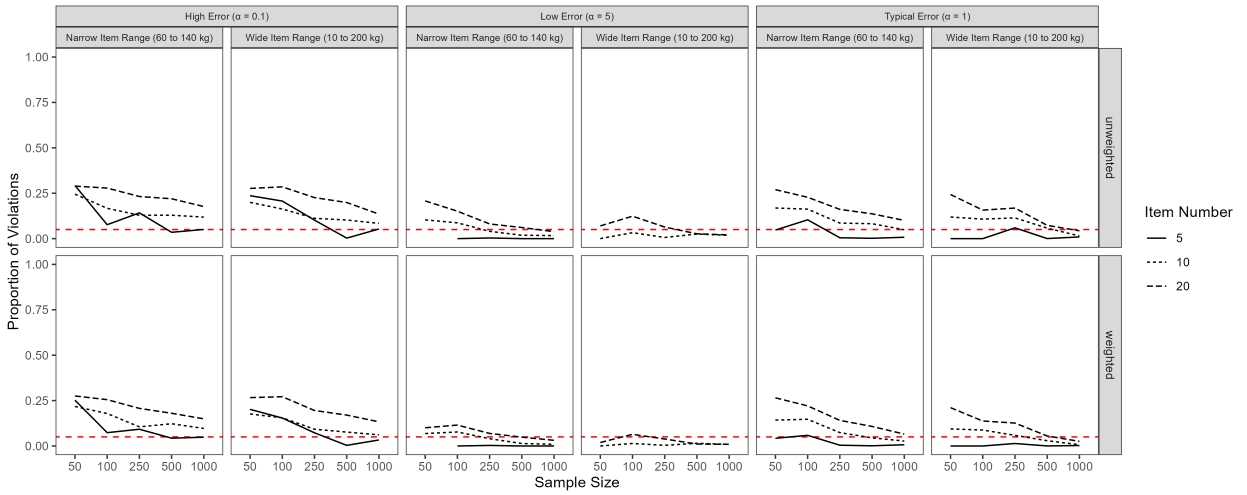

Figure 4: Unweighted and weighted mean violation proprotions for checks of the cancellation axioms of additive conjoint measurement in each dataset.

were biased dependent upon whether the strengths or loads were used for anchoring the transformations; however, for narrow ranges there was little obvious bias and lower error resulted in more accurate scale recovery independently of the anchoring process used. For example, with models displaying typical error (i.e., $\alpha = 1$) or low error (i.e., $\alpha = 5$) the MADPs are at most around ~5%.

6.3.3 Derivation of effort from Rasch model estimates

The four methods of deriving effort from the Rasch model estimates yielded quite different estimates of the effort required for a given person-item pair (i.e., person of a given strength lifting a given load).

6.3.3.1 Mean-sigma scale recovery method ($E_{Api,mean-sigma}$) In line with the mean-sigma scale recovery itself this method for deriving effort (i.e., $E_{Api,mean-sigma}$) showed improved estimates independent of using strength or load anchoring with increased sample size, item number, and for the narrow item ranges (the wide item ranges bias in scale recovery impacting effort estimation also). Again, the for models displaying typical error or less the MADPs for effort are <5%. Figures (9) and (10) shows the scale recovery and the MADPs for the mean-sigma recovery with anchoring based on the strength values, and figures (11) and (12) based on the load values.

6.3.3.2 Odds formulation method ($E_{Api,odds}$) Interestingly, with the odds formulation approach (i.e., $E_{Api,odds}$) the models with typical error clearly performed best with both higher and lower error displaying biased estimates. Within the typical error models the narrower item range resulted in the least error with little relative impact of even sample size and item number (all MADPs <5%). Figures (13) and (14) show the scale recovery and the MADPs for the odds formulation approach.

6.3.3.3 Logit difference method ($E_{Api,logit-diff}$) The logit difference method (i.e., $E_{Api,logit-diff}$), despite being somewhat directionally related to the actual effort in certain models, did not appear to perform well. Further, due to it being on the logit scale it was not possible to examine its error compared with the actual effort leaving only the directional associations. Figure (15) shows the logit differences plotted against the actual effort.

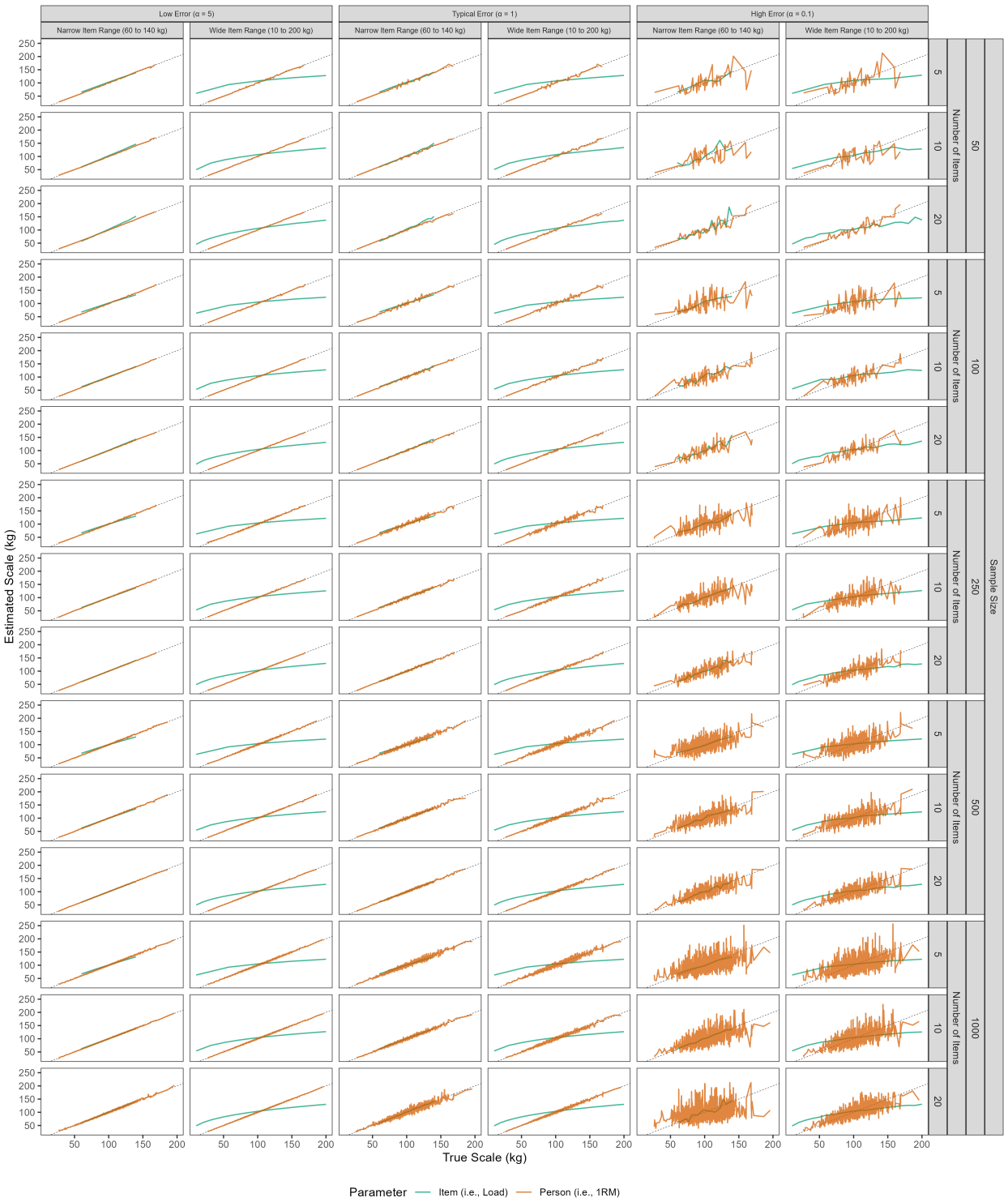

Figure 5: Comparison of Rasch estimates of kiolgram scale to the true kiolgram scale in each dataset when anchored on person strength (i.e., one repetition maximum [1RM]).

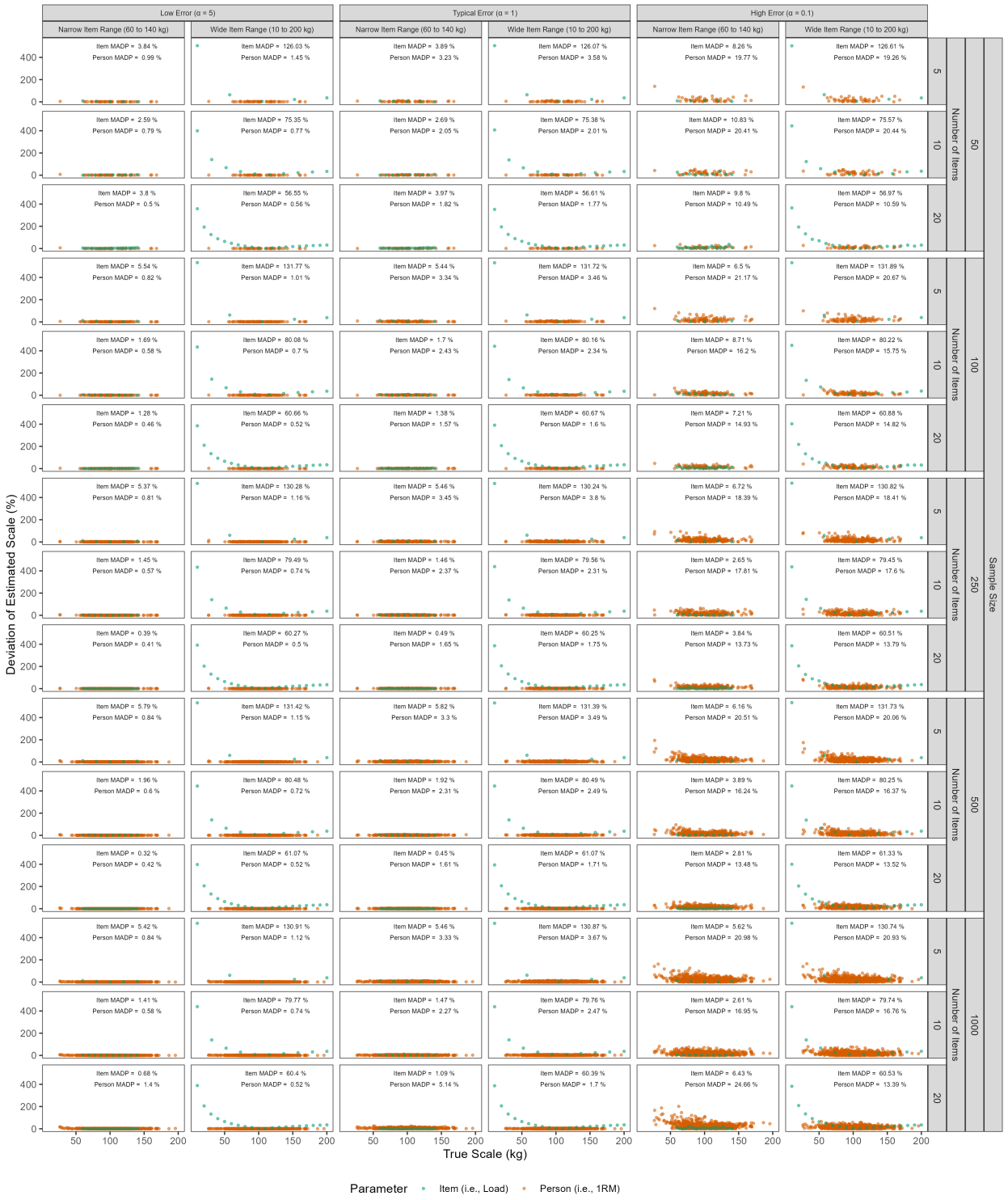

Figure 6: Absolute deviation proportion and mean absolute deviation proportion (MADP) values from comparison of Rasch estimates of kiolgram scale to the true kiolgram scale in each dataset when anchored on person strength (i.e., one repetition maximum [1RM]).

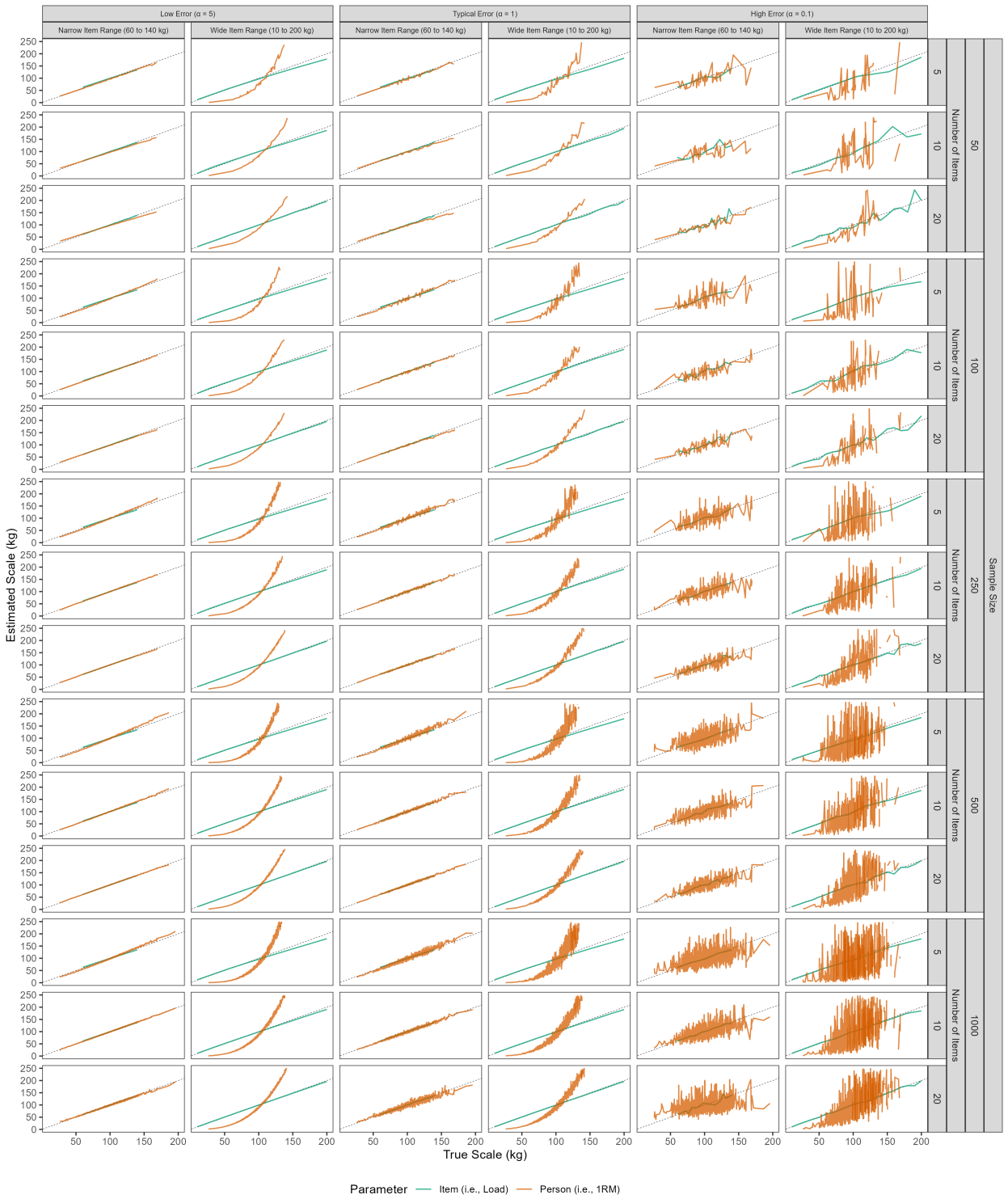

Figure 7: Comparison of Rasch estimates of kiolgram scale to the true kiolgram scale in each dataset when anchored on item load (i.e., load to be lifted).

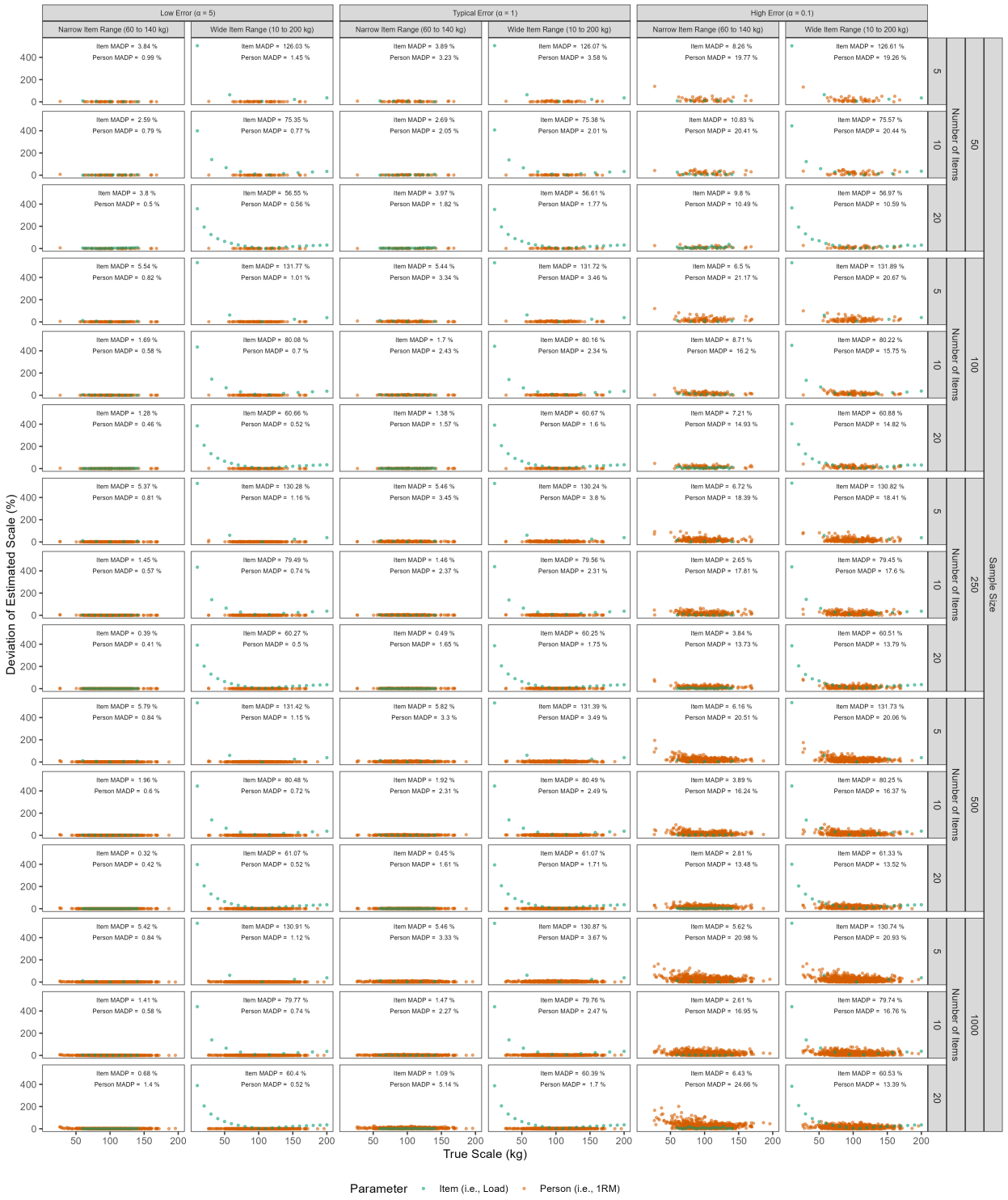

Figure 8: Absolute deviation proportion and mean absolute deviation proportion (MADP) values from comparison of Rasch estimates of kiolgram scale to the true kiolgram scale in each dataset when anchored on item load (i.e., load to be lifted).

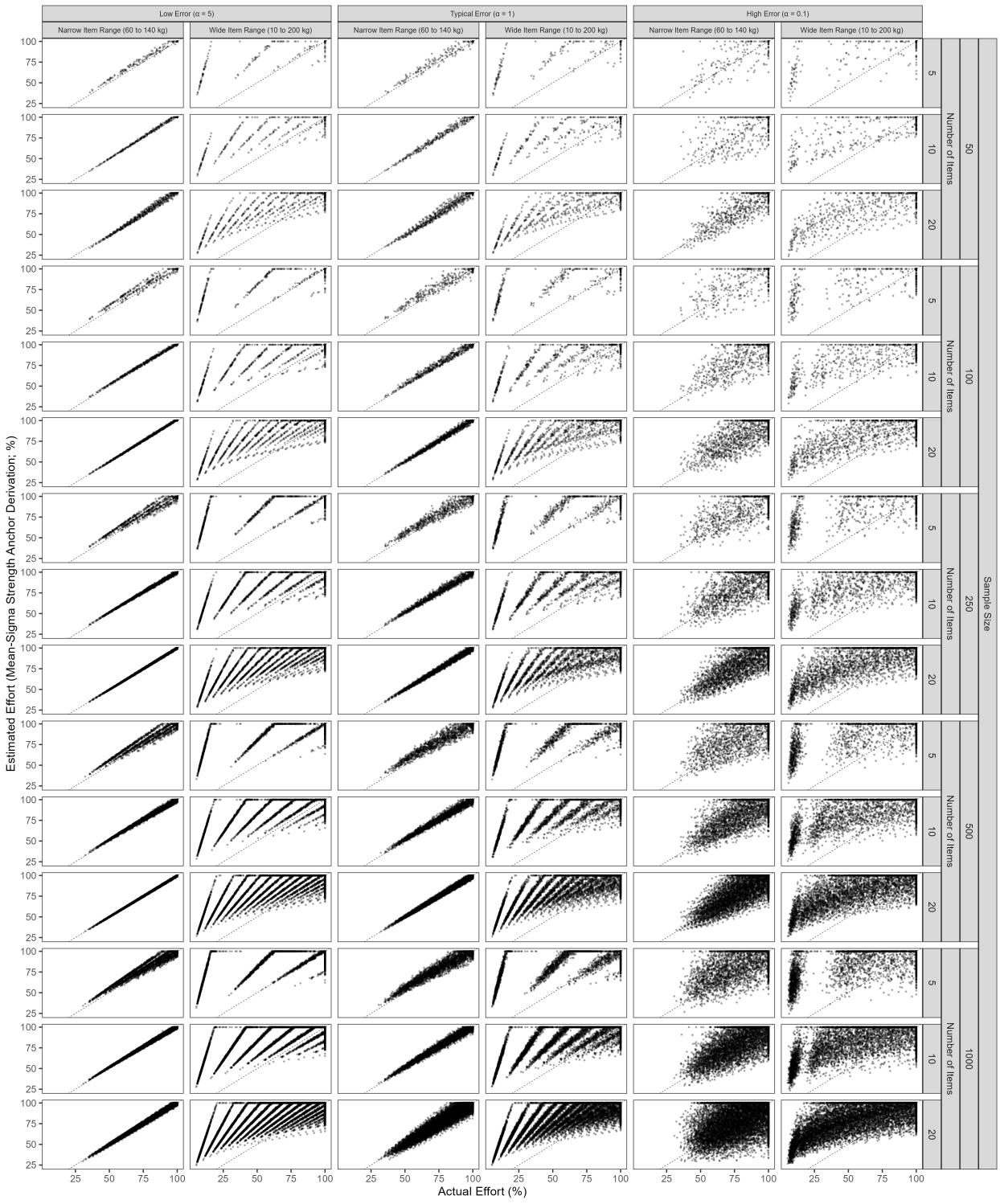

Figure 9: Estimation of effort compared to actual effort when using the mean-sigma scale recovery anchored on person strength (i.e., one repetition maximum [1RM]).

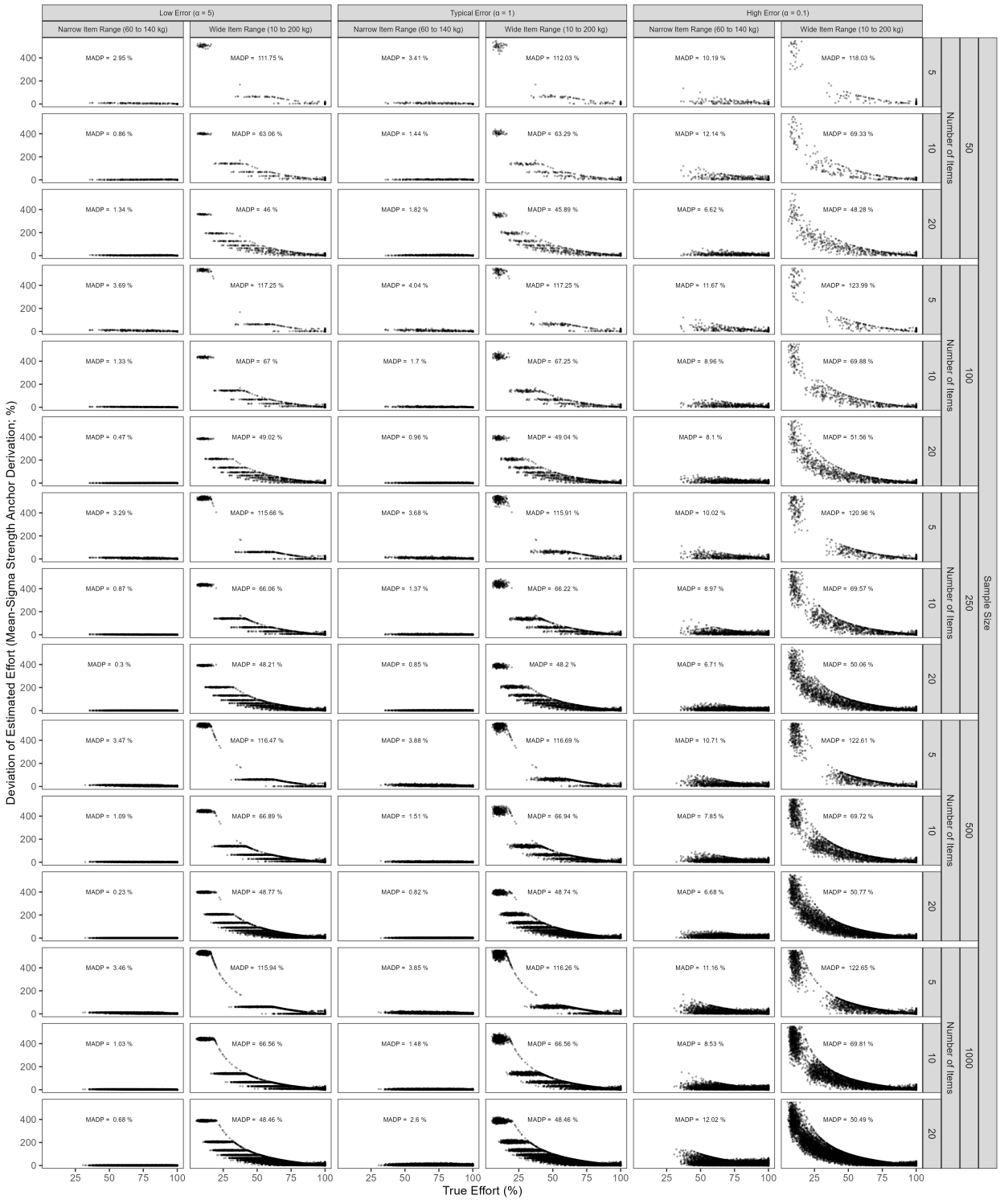

Figure 10: Absolute deviation proportion and mean absolute deviation proportion (MADP) values from comparison of effort estimates to the actual effort using the mean-sigma scale recovery anchored on person strength (i.e., one repetition maximum [1RM]).

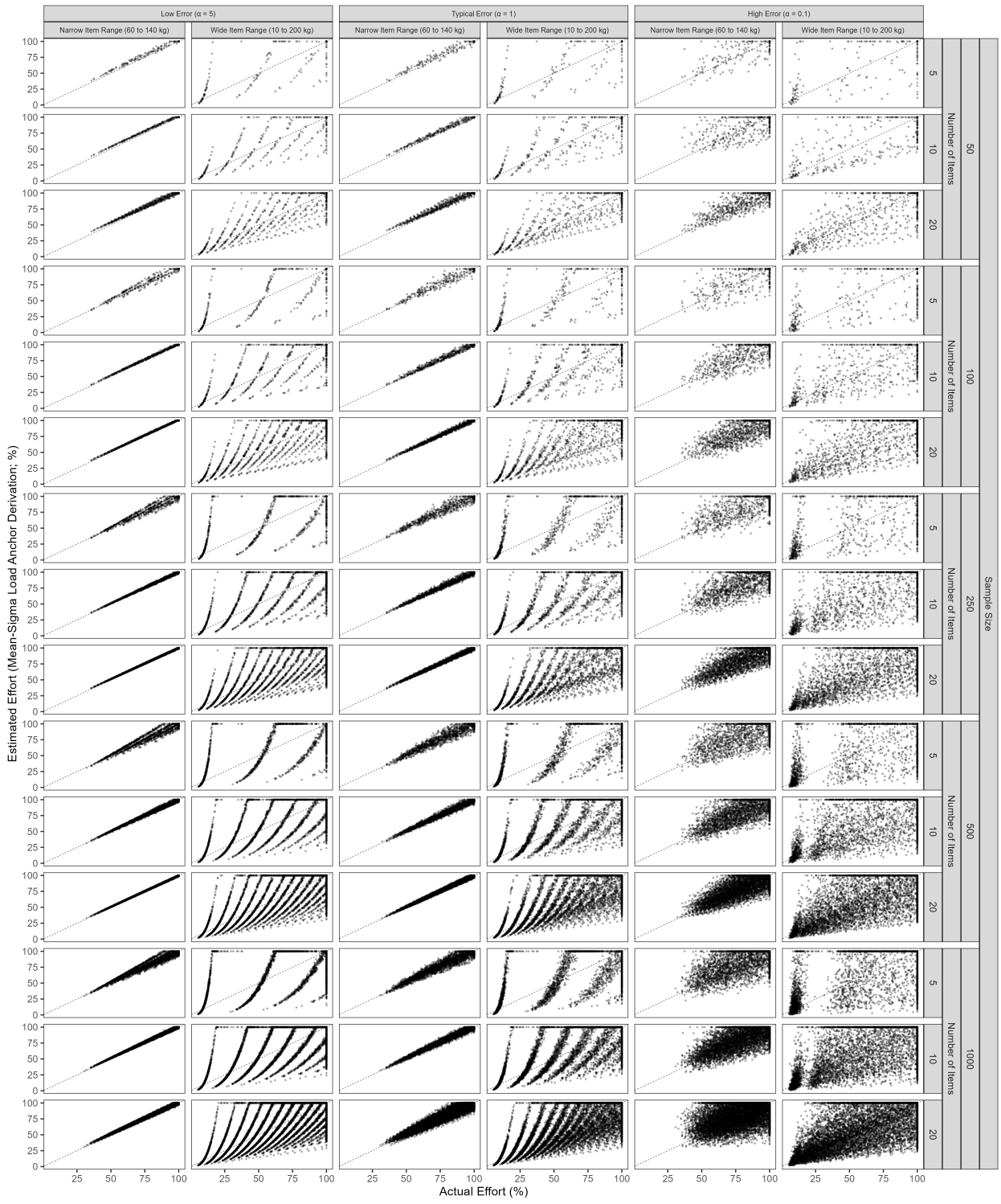

Figure 11: Estimation of effort compared to actual effort when using the mean-sigma scale recovery anchored on item load (i.e., load to be lifted).

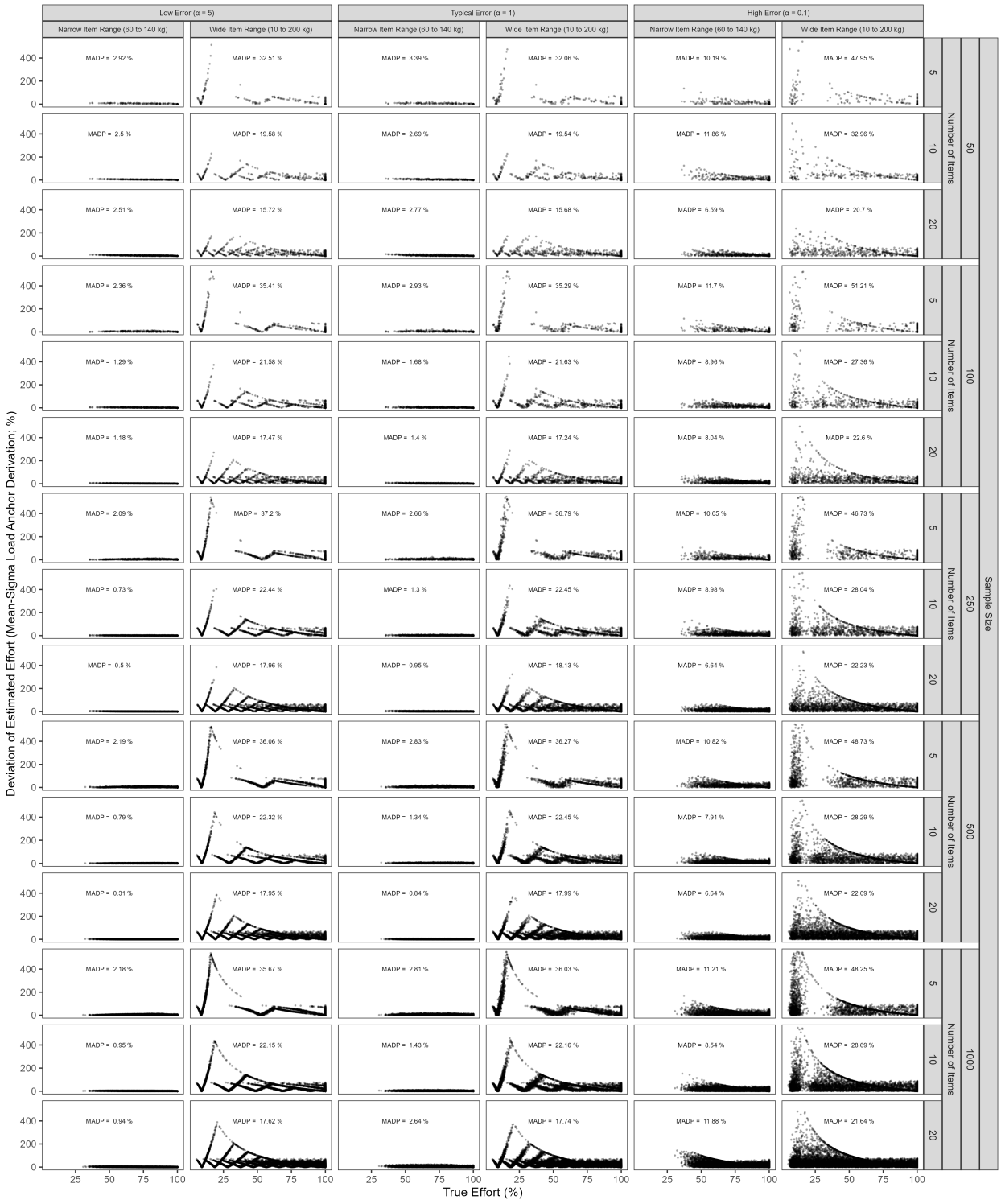

Figure 12: Absolute deviation proportion and mean absolute deviation proportion (MADP) values from comparison of effort estimates to the actual effort using the mean-sigma scale recovery anchored on item load (i.e., load to be lifted).

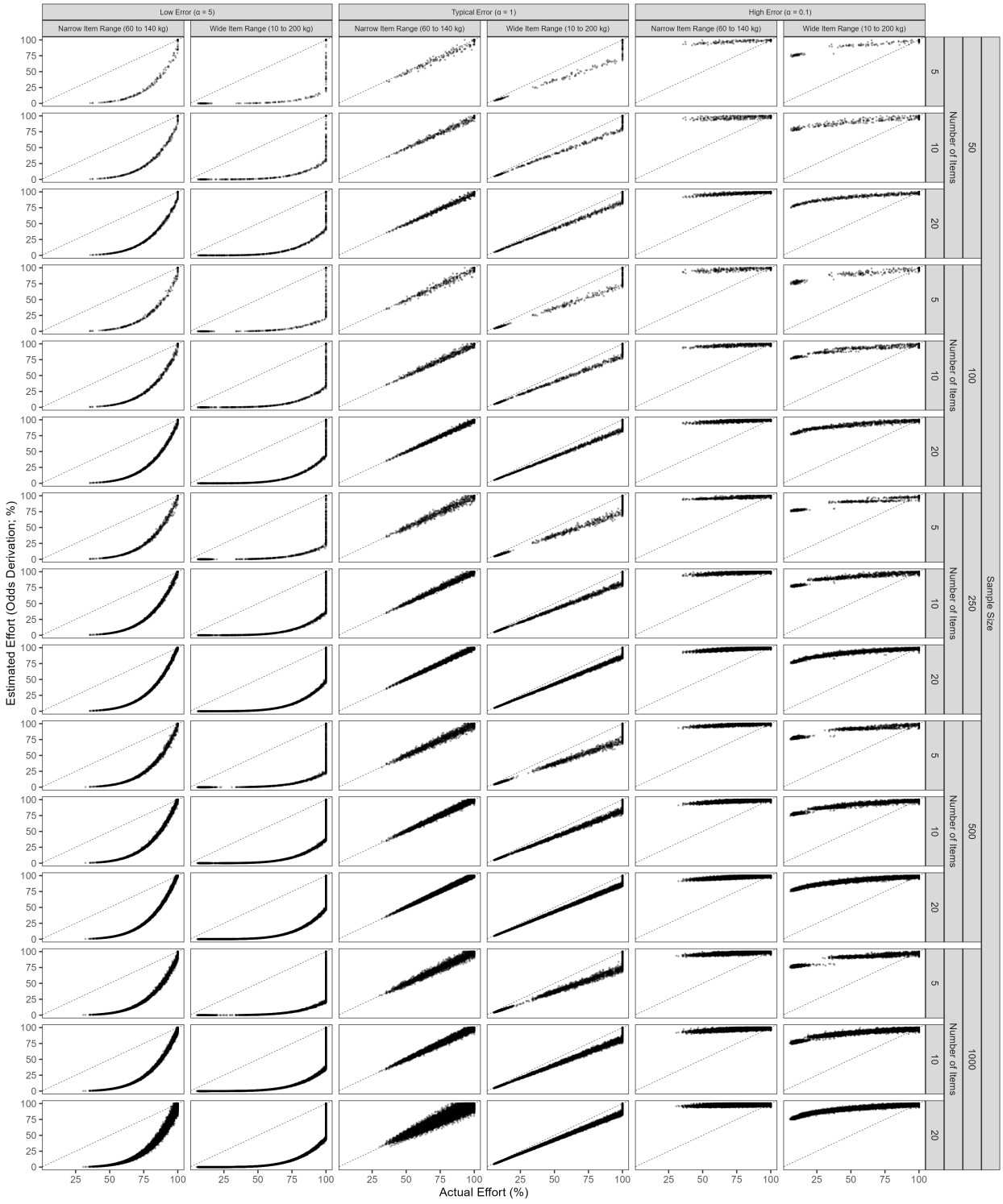

Figure 13: Estimation of effort compared to actual effort when using the odds formulation method.

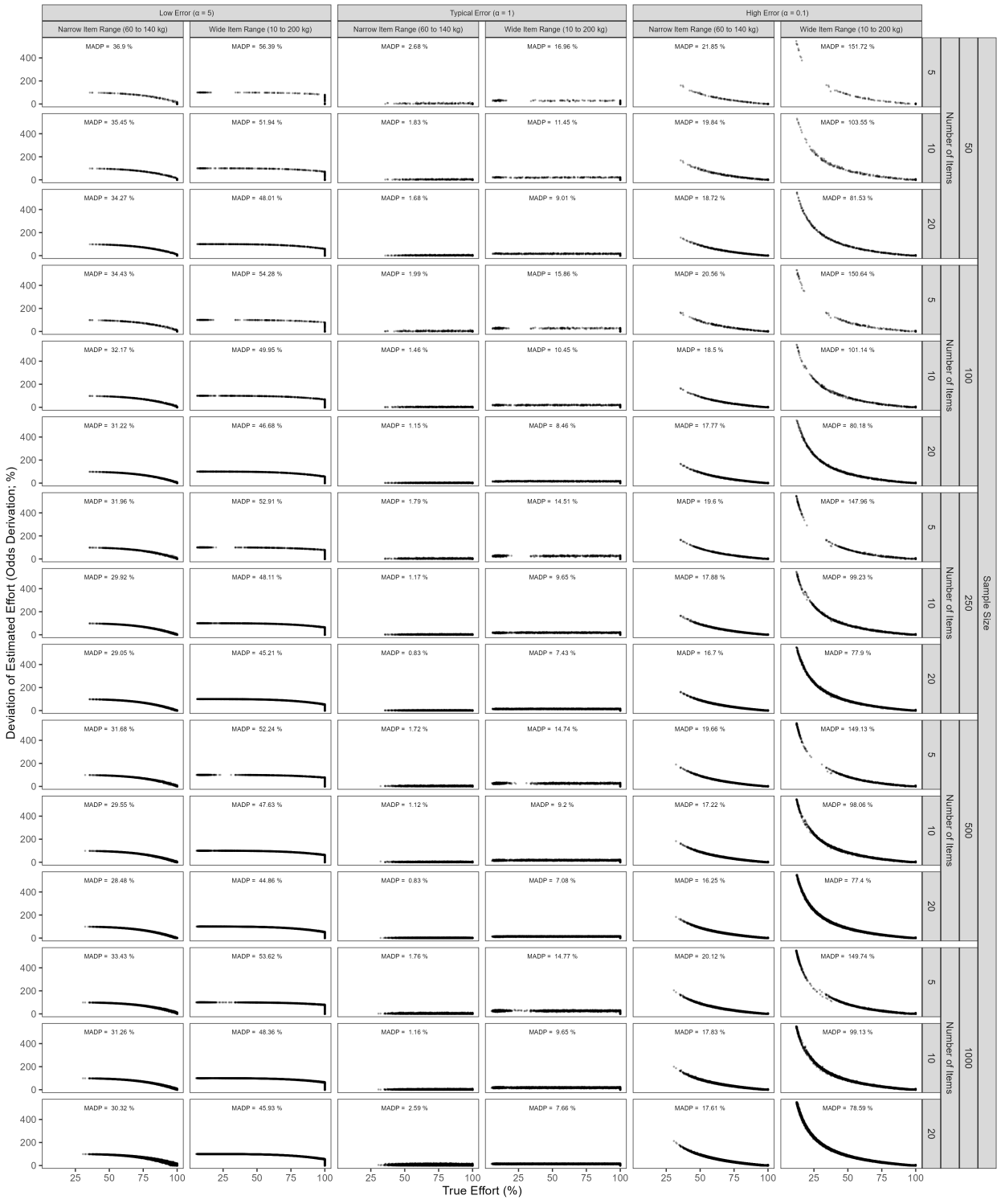

Figure 14: Absolute deviation proportion and mean absolute deviation proportion (MADP) values from comparison of effort estimates to the actual effort when using the odds formulation method

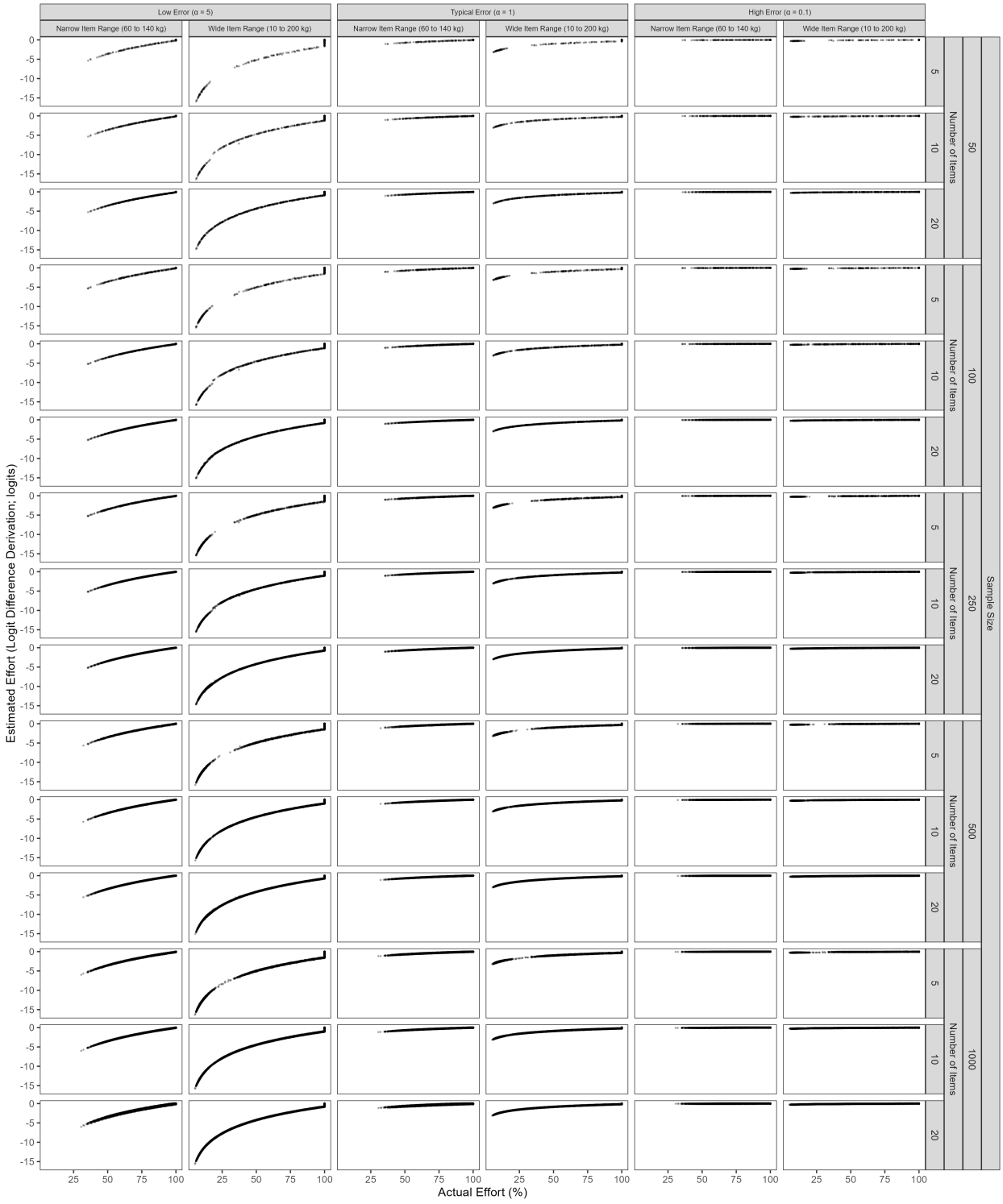

Figure 15: Estimation of effort compared to actual effort when using the logit difference method.

6.3.3.4 Logit shift method ($E_{Api,logit-shift}$) With the logit shift method (i.e., $E_{Api,logit-shift}$) both typical and low error models performed similarly well with MADPs <5%. Within these models the narrower item range again resulted in the least error with little relative impact of sample size and item number. However, it appeared as though the logit shift method tended to show some bias at lower effort levels which were not included in the narrow item range (lowest loads were 60 kg) and so at higher efforts performance appeared similar between narrow and wide item ranges. The bias at lower efforts appears to be because, in the wide item range datasets, R was set according to the lowest item difficulty resulting in some efforts to be calculated as zero. Figure (16) and (17) shows the scale recovery and the MADPs for the logit shift approach.

6.4 Considerations for methods of effort derivation

Regarding the performance of each method the logit difference clearly appeared to be unsuitable for derivation of effort whereas, at least in terms of mean errors, the mean-sigma, odds formulation, and logit shift all performed fairly well for typical Rasch models.

The mean-sigma method worked best when using a larger sample size, item number, and for the narrow item ranges in addition to test conditions where error is low. This appeared to be largely independently of whether the mean-sigma transformations of the scale were anchored to known item or person variables. This is useful to know as in most cognitive test conditions, whilst we may in some cases have some quantitative scale to anchor upon with respect to item characteristics, we do not know the corresponding value for the person parameters on this scale. As such, application of this approach relies on the use of some cognitive task for which there is already a simple quantitative scale for items, or where theory is developed sufficiently such that a quantitative scale can be developed for items (e.g, the integer value for an N-back or simple span task, or Lexile's (Stenner et al., 2013), or hierarchical ordering of complexity (Commons et al., 2008)). Where such an item scale is present it can be used to anchor and transform underlying logit scores for persons into this scale such that effort can be derived for person-item pairs.

In the absence of tests where quantitative scales exist for items to anchor upon, the odds and logit shift approaches may be suitable alternatives. Although, the logit-shift method may present some bias at lower actual effort values whereas, in the case of the typical Rasch error (i.e., $\alpha = 1$) and a narrow item range, the odds formulation showed no bias and minimal error. This was also impacted relatively little by item number and sample size. As such, it seems likely to be an appropriate and practical method for derivation of effort from cognitive tasks conforming to the standard Rasch model.

Before some concluding remarks and limitations to consider, I will demonstrate the mean-sigma, odds, and logit shift methods with a real dataset.

7 Example using a real dataset of N-back tasks

I was able to locate several studies that had employed N-back tasks and either had openly available data or the authors were able to share their data (Beh et al., 2021; Blacker & Curby, 2013; Tiberghien et al., 2017; Westbrook et al., 2020; Westbrook et al., 2013; Zerna et al., 2022). These studies had used a variety of N-back task items; for example, some used only one task (i.e., 3-back; Blacker & Curby (2013); Tiberghien et al. (2017)) whereas others used a range of items (i.e., 0-,2-, and 3-back; Beh et al. (2021); 1- to 4-back; Westbrook et al. (2020); Zerna et al. (2022); or 1- to 6-back; Westbrook et al. (2013)). Thus, some people in the combined dataset had missing data for certain N-back items. The Rasch model can however, in the context of generalised linear mixed effects model framework, handle missing data in its estimation of parameters[19] (Waterbury, 2019).

First I explored whether Guttman patterns might be present by calculating the coefficient of reproducibility, and also whether there were violations of the cancellation conditions of ACM in the data itself. Due to the sample size impact on checks for cancellation, and also to ensure that sufficient rows (i.e., total scores)

[19]Assuming no systematic non-random reasons for the fact that certain studies decided to examine certain N-back items.

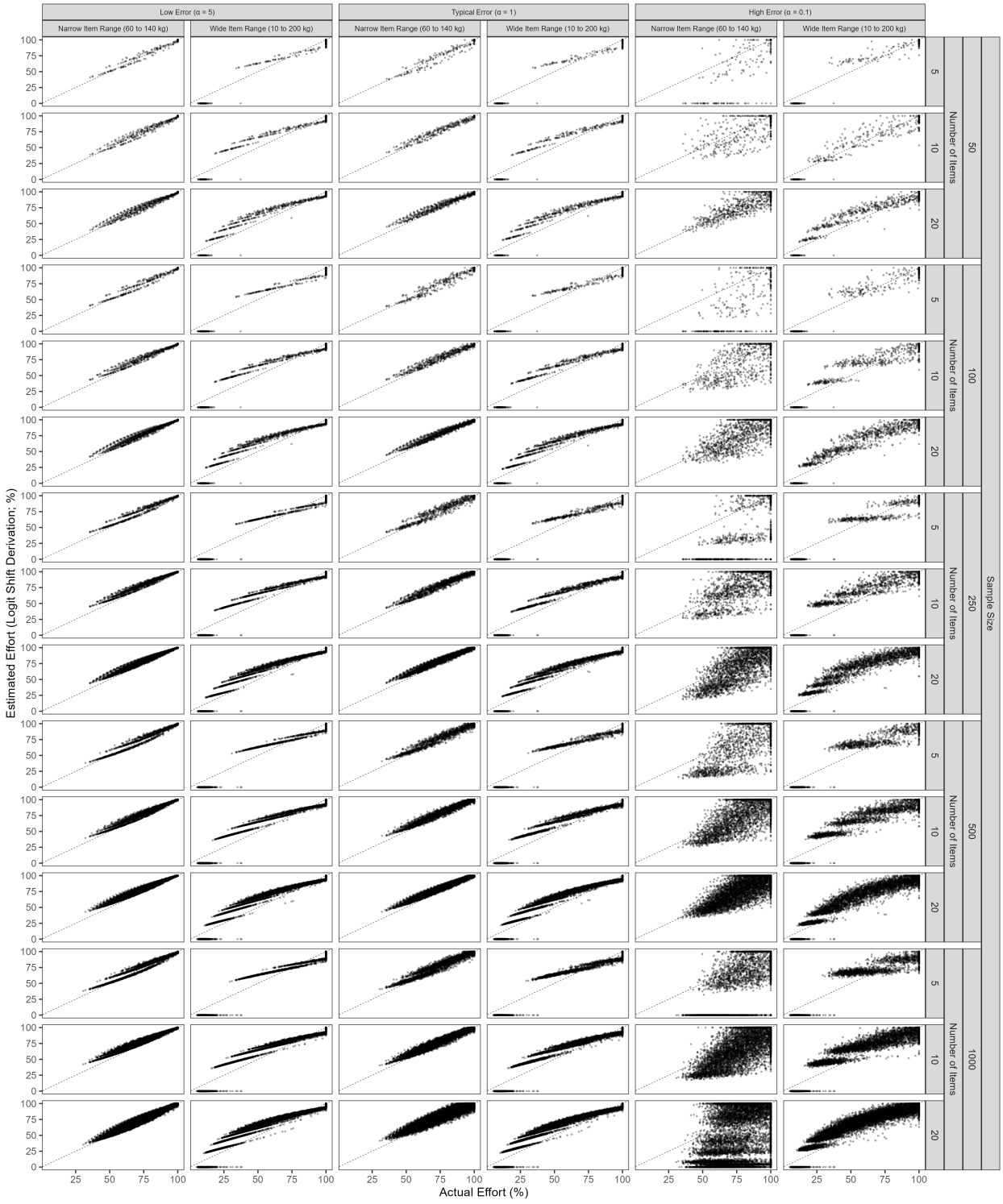

Figure 16: Estimation of effort compared to actual effort when using the logit shift formulation method.

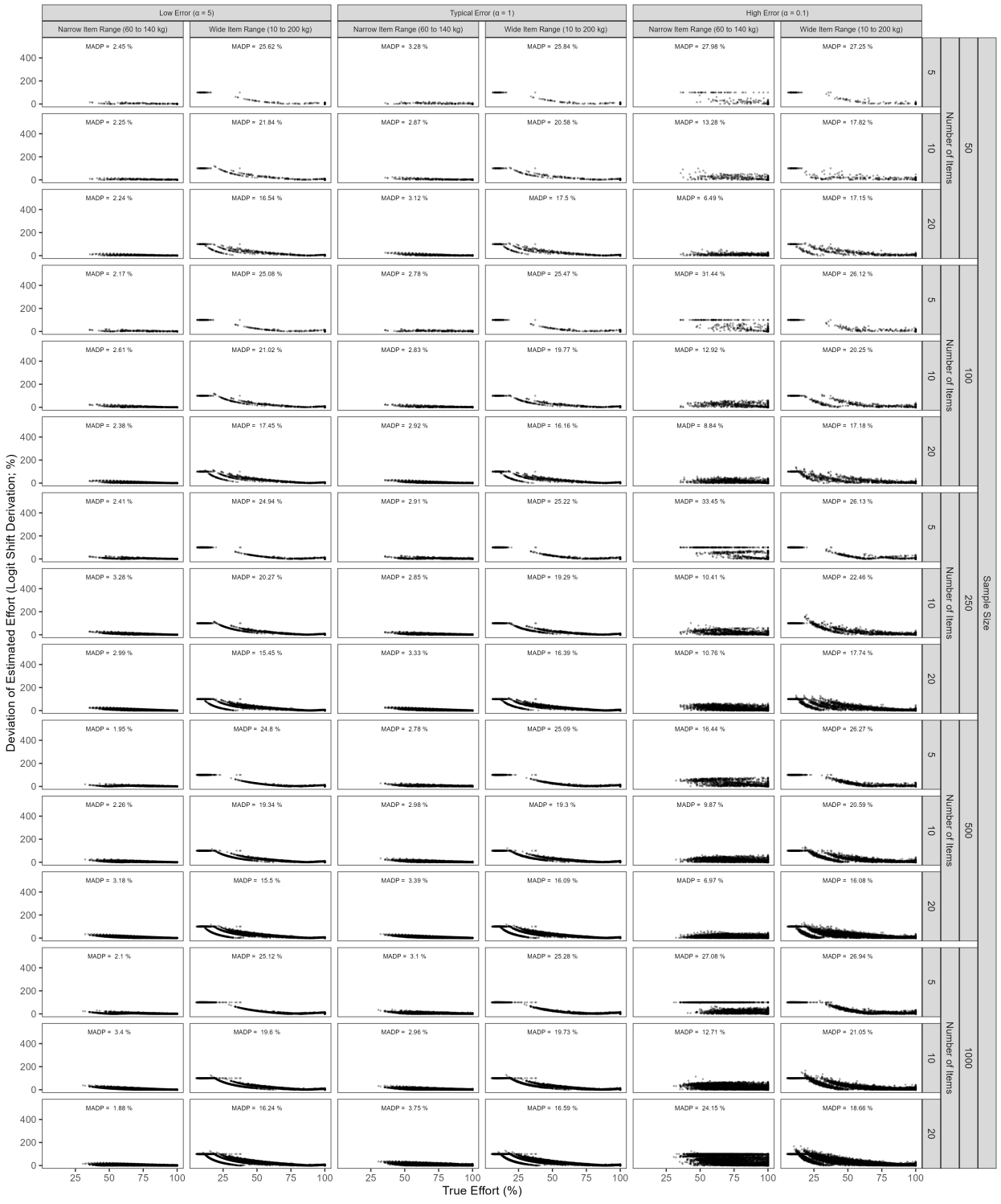

Figure 17: Absolute deviation proportion and mean absolute deviation proportion (MADP) values from comparison of effort estimates to the actual effort when using the logit shift formulation method

were produced to provide sufficient matrix sizes for checks[20], I limited checks for cancellation to the largest dataset I could create from these studies. This meant that checks were conducted in the 1- to 4-back items of Westbrook et al. (2020; 2013) and Zerna et al. (2022) combined data. This resulted in a dataset of 149 participants. Each individual study had varying numbers of trials per item also, however all three had at least 30 target trials and so cancellation checks were performed for each of these 30 in order (i.e., the first target trial to appear to participants, then the second etc.). As can be seen in figure (18), the coefficients of reproducibility in these trials were all $< 90\%$ and the majority of checks for cancellation conditions all showed proportions of violations $< 5\%$. This suggests that the N-back tasks yields data of a quantitative structure and thus may be amenable to fundamental measurement operations through the Rasch model.

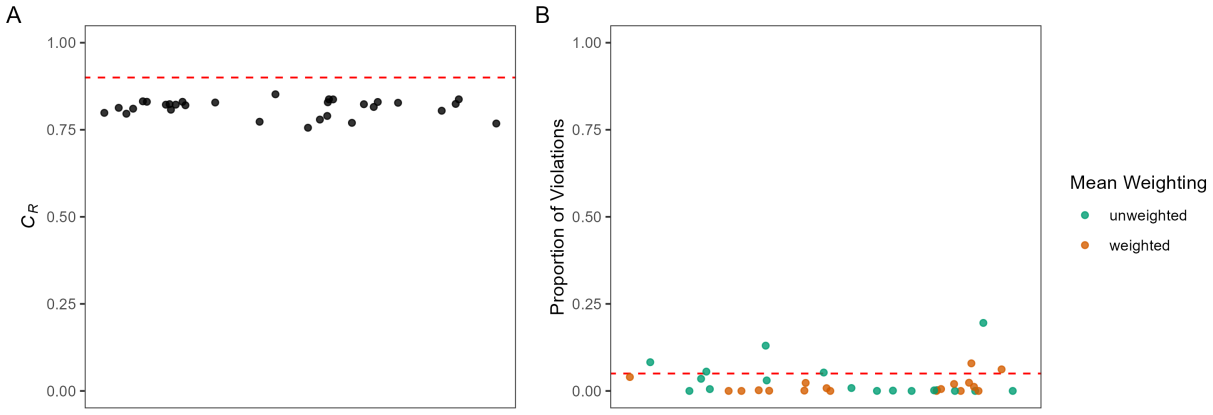

Figure 18: Cofficient of reproducibility and unweighted and weighted mean violation proportions for checks of the cancellation axioms of additive conjoint measurement in N-back data

Following this I fit a Rasch model to the full dataset and looked to see how well a mean-sigma transformation anchored on the items recovered the N-back integer values. One dataset included a 0-back task which, although item fit statistics (see https://osf.io/rwhsj) seemed to be within reasonable bounds (Wright & Linacre, 1994), appeared to be so incredibly easy that it affected scale recovery (see https://osf.io/vu76d). As such, I refit the Rasch model excluding the 0-back task. Although there were a small number of misfitting persons with very low infit and outfit based on mean-square values the fit was reasonable for all items (see https://osf.io/pajkx). Further, there was reasonable recovery of the integer scale without any obvious bias, though not as accurately as in the earlier simulations with MADP of 8.84% (see figure 19).

Lastly, I explored application of the different effort operationalisations to the Rasch estimates (omitting the logit difference method as it did not perform well in the simulations). Interestingly, in this dataset we see some stark differences between the methods (figure 20) that might be unexpected given the simulations above.

Whereas in the simulations all three methods tended to give reasonable estimates with no obvious bias and little error, here the mean-sigma method reveals higher effort estimates, followed by the logit shift method, and then the odds formulation. Exploring the distribution of item and person estimates (see https://osf.io/4a9cg) suggests that the item difficulties are in fact quite low (1- to 6-back appear relatively easy) whilst the person ability estimates are quite high (persons in the sample appear to be quite good at the N-back). In order to understand how this might impact derivation of effort I returned to one of the simulated datasets with 1000 persons, 10 items, narrow item range, and typical error. I adapted it to reflect the situation in this empirical dataset to see how it affected scale recovery and effort estimation by removing items that were above the simulated mean of ability (i.e., 100 kg), and also persons that were below this. In situations such as this it seems that when recovering the true scale using the mean-sigma method item difficulties are overestimated relative to person abilities (see https://osf.io/edfa4) potentially explaining why the mean-sigma method

[20]Note, eight of the trials yielded matrices that could not be used for checks of the cancellation axioms. Thus, the checks performed were for 22 N-back target trials.

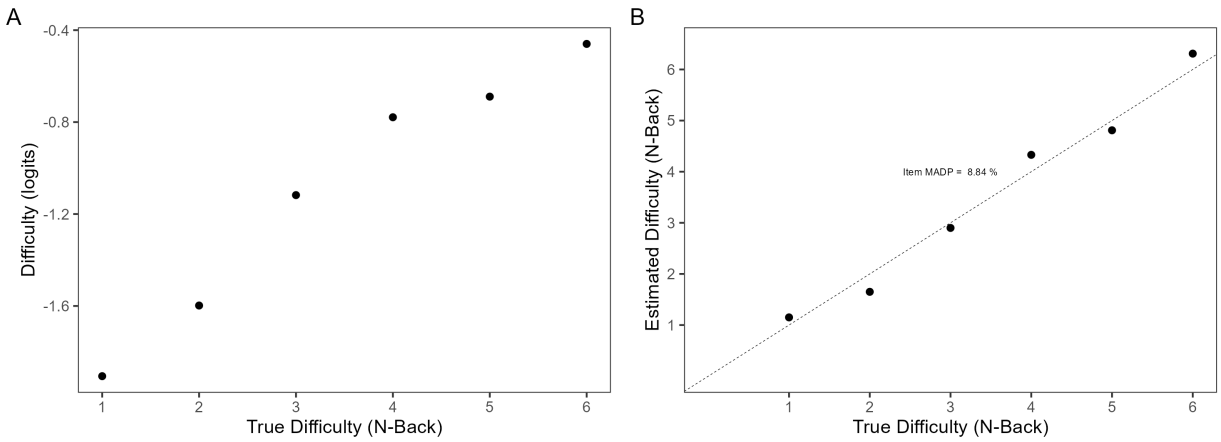

Figure 19: Comparison of Rasch estimates of N-back integer scale to the true N-back integer scale.

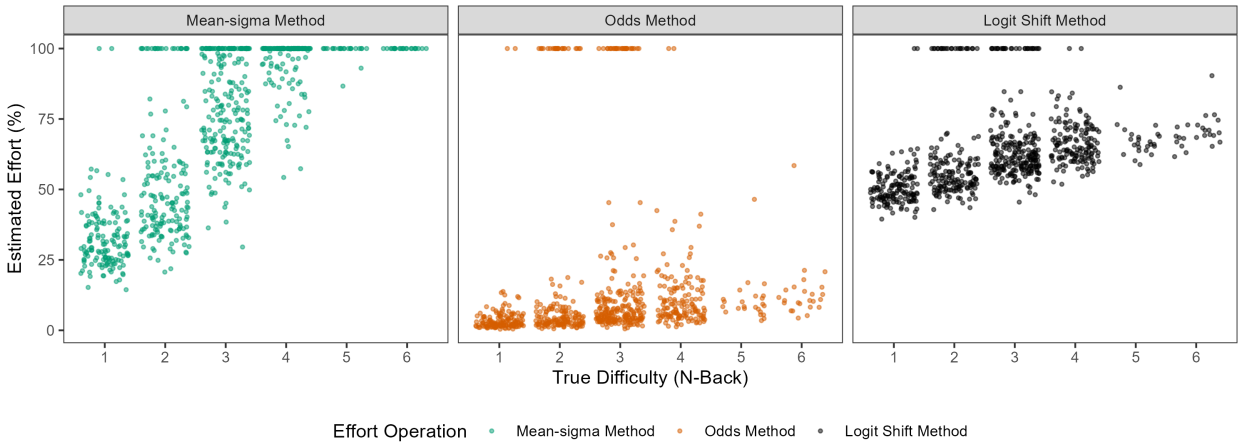

Figure 20: Comparison of Rasch estimates of N-back integer scale to the true N-back integer scale.

results in higher effort estimates for the N-back data. Indeed, this bears out in the estimated effort using the mean-sigma method, and further the odds and logit shift methods also produce lower effort estimates (see https://osf.io/5n4hq).

Two of the N-back datasets included self-report ratings of the perception of effort required to perform the N-back task items using the NASA-TLX effort subscale (Beh et al., 2021; Westbrook et al., 2013). I fit an ordered beta regression model, appropriate to bounded data such as effort (Kubinec, 2022; Steele et al., 2022), of the perception of effort with each actual effort operationalisation using random intercepts and slopes for each person. Examining the relationship between the estimated actual effort and the perception of effort here would likely lead to different conclusions about the nature of the psychophysical relationship (figure 21). Indeed, given the simulations and the results of analysing this empirical dataset, it would certainly seem that specific empirical conditions might be required to employ these approaches to deriving effort for cognitive tasks.

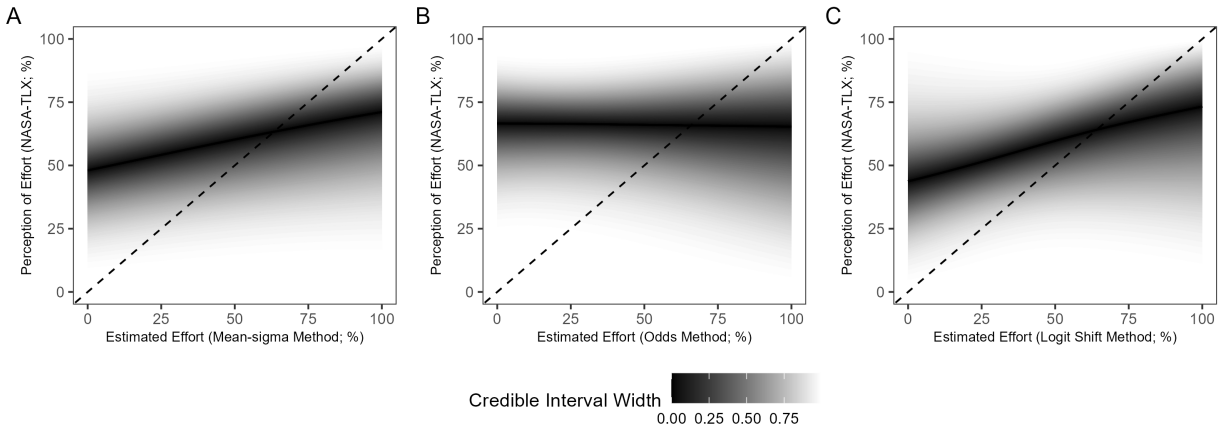

Figure 21: Ordered beta regression models of perception of effort with each actual effort operationalisation as a predictor during the N-back task.

8 Concluding remarks

Given how widely effort as a concept is employed in the cognitive sciences it seems that there should be interest in more precisely defining it, and as such improving the ability to consider how appropriate proposed operationalisations of it are. The conceptual definition I have proposed is explicit in hypothesising effort to be a variable derived as the ratio of two quantitative variables; the demands of a task and capacity of the actor to meet the task demands. As such, I contend that approaches to its operationalisation should consider in the case of cognitive tasks whether the latent dispositional capacity to perform a task, and the demands a task present, are indeed quantitative in structure. Following support in this scientific task, measurement operations might be possible and application of the Rasch model may be amenable to a variety of methods for the derivation of effort from its estimates of person abilities and item difficulties.

However, as the simulations presented here and examination of an empirical dataset reveal, specific conditions may be required in order to achieve this in practice. Arguably though, this is in keeping with the confirmatory, rather than exploratory, philosophical tradition underlying the application of Rasch models for fundamental measurement more generally (Salzberger, 2013; Stone & Stenner, 2014). It seems that when a tasks data conform to the typical Rasch model, and the items employed have difficulties reasonably similar to the distribution of person abilities, it is possible to achieve measurement such that we can utilise transformations of the person and item parameter estimates in order to derive the efforts required for persons performing each item. Further, given that formulations for deriving effort that do not rely on a developed theory of item

difficulty (e.g., Stenner et al. (2013); Commons et al. (2008)) or existing quantitative scale for items (e.g., N-back or simple span integers), such as the odds or logit shift approach, produce accurate and unbiased derivations of effort under such conditions, it seems that research programs could be aimed at testing whether a broad range of existing cognitive tests used in effort research involve quantitative structure and conform to the Rasch model, or upon development of tests that do.

I will make a final remark regarding the inherent trouble with all of this for psychological science. Even if, as I do, we adopt a realist ontological stance towards the concepts postulated, and we endeavour to establish whether the tests we employ produce data of quantitative structure and that conform to the Rasch model. Even in cases where we can be confident that the differences between task demands (i.e., item difficulties) are indeed quantitative, it remains possible that the dispositional cognitive capacities that permit the observed behavioural responses to attempted task performances are in fact heterogenous (Michell, 2013; Richters, 2021). The capacities may not be homogeneous (i.e., unidimensional) in the manner that is a fundamental assumption of the approaches I have described even where the task itself may be. For at least some unidimensional tasks, as item difficulty increases the corresponding increases in cognitive capacity required for successful attempted performance are mutually qualitatively heterogeneous. Even where item differences are quantitative, qualitatively different strategies may lead to different levels of performance, or even the same level of performance. If so, even if the test data conform to the axioms of ACM and fit the Rasch model, we still cannot be sure instances of the capacity or attribute itself constitute degrees of a quantitative structure. The assumption of psychological homogeneity is hard to reconcile (Richters, 2021), and I do not confess to having any particular solution to it in general of course. However, I reiterate Michell's (2013) suggestion which I think should accompany, and perhaps precede, my suggestion above regarding test development; the structure of the attributes underlying cognitive test performance will be best understood by drawing inferences from the character of the phenomena involved in such testing which is to say by theorising regarding these concepts and testing the deductively entailed hypotheses.

9 References

Beh, W.-K., Wu, Y.-H., An-Yeu, & Wu. (2021). *MAUS: A Dataset for Mental Workload Assessmenton N-back Task Using Wearable Sensor*. arXiv. https://doi.org/10.48550/arXiv.2111.02561

Bermúdez, J. P., & Massin, O. (2023). Efforts and their feelings. *Philosophy Compass*, *18*(1), e12894. https://doi.org/10.1111/phc3.12894

Blacker, K. J., & Curby, K. M. (2013). Enhanced visual short-term memory in action video game players. *Attention, Perception, & Psychophysics*, *75*(6), 1128–1136. https://doi.org/10.3758/s13414-013-0487-0

Boeck, P. D., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The Estimation of Item Response Models with the lmer Function from the lme4 Package in R. *Journal of Statistical Software*, *39*, 1–28. https://doi.org/10.18637/jss.v039.i12

Bond, T., Yan, Z., & Heene, M. (2020). Applying the Rasch Model: Fundamental Measurement in the Human Sciences. In *Routledge & CRC Press*. https://www.routledge.com/Applying-the-Rasch-Model-Fundamental-Measurement-in-the-Human-Sciences/Bond-Yan-Heene/p/book/9780367141424

Borsboom, D., Maas, H. L. J. van der, Dalege, J., Kievit, R. A., & Haig, B. D. (2021). Theory Construction Methodology: A Practical Framework for Building Theories in Psychology. *Perspectives on Psychological Science*, *16*(4), 756–766. https://doi.org/10.1177/1745691620969647

Borsboom, D., & Scholten, A. Z. (2008). The Rasch Model and Conjoint Measurement Theory from the Perspective of Psychometrics. *Theory & Psychology*, *18*(1), 111–117. https://doi.org/10.1177/0959354307086925

Bramley, T. (2020). Metaphors and the psychometric paradigm. *Assessment in Education: Principles, Policy & Practice*, *27*(2), 178–191. https://doi.org/10.1080/0969594X.2020.1731421

Brogden, H. E. (1977). The rasch model, the law of comparative judgment and additive conjoint measurement. *Psychometrika*, *42*(4), 631–634. https://doi.org/10.1007/BF02295985

Bürkner, P.-C. (2020). *Bayesian Item Response Modeling in R with brms and Stan*. arXiv. https://doi.org/10.48550/arXiv.1905.09501

Campbell, N. R. (1920). *Physics: The Elements*. Cambridge University Press.

Carnap, R. (1945). The Two Concepts of Probability: The Problem of Probability. *Philosophy and Phenomenological Research*, *5*(4), 513–532. https://doi.org/10.2307/2102817

Commons, M. L., Goodheart, E. A., Pekker, A., Dawson, T. L., Draney, K., & Adams, K. M. (2008). Using Rasch scaled stage scores to validate orders of hierarchical complexity of balance beam task sequences. *Journal of Applied Measurement*, *9*(2), 182–199.

Domingue, B. (2014). Evaluating the Equal-Interval Hypothesis with Test Score Scales. *Psychometrika*, *79*(1), 1–19. https://doi.org/10.1007/s11336-013-9342-4

Doran, H., Bates, D., Bliese, P., & Dowling, M. (2007). Estimating the Multilevel Rasch Model: With the lme4 Package. *Journal of Statistical Software*, *20*, 1–18. https://doi.org/10.18637/jss.v020.i02

Dubin, R. (1969). *Theory Building*. Free Press.

Ehrich, J. F., Howard, S. J., Bokosmaty, S., & Woodcock, S. (2021). An Item Response Modeling Approach to Cognitive Load Measurement. *Frontiers in Education*, *6*. https://www.frontiersin.org/articles/10.3389/feduc.2021.648324

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory*. Psychology Press. https://doi.org/10.4324/9781410605269

Fischer, G. H. (1995). Some neglected problems in IRT. *Psychometrika*, *60*(4), 459–487. https://doi.org/10.1007/BF02294324

Fischer, G. H., & Molenaar, I. W. (1995). *Rasch Models: Foundations, Recent Developments, and Applications*. Springer US. https://link.springer.com/book/10.1007/978-1-4612-4230-7

Fisher, A. G., Bryze, K. A., Granger, C. V., Haley, S. M., Hamilton, B. B., Heinemann, A. W., Puderbaugh, J. K., Linacre, J. M., Ludlow, L. H., Mccabe, M. A., & Wright, B. D. (1994). Applications of conjoint measurement to the development of functional assessments. *International Journal of Educational Research*, *21*(6), 579–593. https://doi.org/10.1016/0883-0355(94)90012-4

Franz, D. J. (2022a). "Are psychological attributes quantitative?" Is not an empirical question: Conceptual confusions in the measurement debate. *Theory & Psychology*, *32*(1), 131–150. https://doi.org/10.1177/09593543211045340

Franz, D. J. (2022b). Psychological measurement is highly questionable but the details remain controversial: A response to Tafreshi, Michell, and Trendler. *Theory & Psychology*, *32*(1), 171–177. https://doi.org/10.1177/09593543211062868

Freund, R. (2019). *Rasch and Rationality: Scale typologies as applied to Item Response Theory* [PhD thesis, UC Berkeley]. https://escholarship.org/uc/item/1vh141kq

Guttman, L. (1944). A Basis for Scaling Qualitative Data. *American Sociological Review*, *9*(2), 139–150. https://doi.org/10.2307/2086306

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. SAGE.

Krantz, D. H., Krantz, D. M., Suppes, P., Luce, R. D., & Tversky, A. (1971). *Foundations of Measurement: Additive and polynomial representations*. Academic Press.

Kubinec, R. (2022). Ordered Beta Regression: A Parsimonious, Well-Fitting Model for Continuous Data with Lower and Upper Bounds. *Political Analysis*, 1–18. https://doi.org/10.1017/pan.2022.20

Künzell, S., Broeker, L., Dignath, D., Ewolds, H., Raab, M., & Thomaschke, R. (2018). What is a task? An ideomotor perspective. *Psychological Research*, *82*(1), 4–11. https://doi.org/10.1007/s00426-017-0942-y

Lamprianou, I. (2013). Application of single-level and multi-level Rasch models using the lme4 package. *Journal of Applied Measurement*, *14*(1), 79–90.

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, *1*(1), 1–27. https://doi.org/10.1016/0022-2496(64)90015-X

Markus, K. A. (2008). Constructs, Concepts and the Worlds of Possibility: Connecting the Measurement, Manipulation, and Meaning of Variables. *Measurement: Interdisciplinary Research and Perspectives*, *6*(1-2), 54–77. https://doi.org/10.1080/15366360802035513

Massin, O. (2022). *Defining Physical Efforts*. PsyArXiv. https://doi.org/10.31234/osf.io/qmg5j

Massin, O. (2017). Towards a definition of efforts. *Motivation Science*, *3*, 230–259. https://doi.org/10.1037/mot0000066

McMullen, T. (2011). *Critiques and Developments "Out there," not "in here": A Realist account of concepts*. Brill. https://doi.org/10.1163/9789004194878_010

Michell, J. (2008). Conjoint Measurement and the Rasch Paradox: A Response to Kyngdon. *Theory &*

Psychology, *18*(1), 119–124. https://doi.org/10.1177/0959354307086926
Michell, J. (2014). The Rasch paradox, conjoint measurement, and psychometrics: Response to Humphry and Sijtsma. *Theory & Psychology*, *24*(1), 111–123. https://doi.org/10.1177/0959354313517524
Michell, J. (2022). Denying Descartes and wary of Wittgenstein: Response to Franz. *Theory & Psychology*, *32*(1), 151–157. https://doi.org/10.1177/09593543211046204
Michell, J. (2013). Constructs, inferences, and mental measurement. *New Ideas in Psychology*, *31*(1), 13–21. https://doi.org/10.1016/j.newideapsych.2011.02.004
Michell, J. (1988). Some problems in testing the double cancellation condition in conjoint measurement. *Journal of Mathematical Psychology*, *32*(4), 466–473. https://doi.org/10.1016/0022-2496(88)90024-7
Michell, J. (2005). The logic of measurement: A realist overview. *Measurement*, *38*(4), 285–294. https://doi.org/10.1016/j.measurement.2005.09.004
Michell, J. (1990). *An Introduction To the Logic of Psychological Measurement*. Psychology Press.
Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, *88*(3), 355–383. https://doi.org/10.1111/j.2044-8295.1997.tb02641.x
Michell, J. (2009). The psychometricians' fallacy: Too clever by half? *British Journal of Mathematical and Statistical Psychology*, *62*(1), 41–55. https://doi.org/10.1348/000711007X243582
Michell, J., & Ernst, C. (1996). The Axioms of Quantity and the Theory of Measurement: Translated from Part I of Otto Hölder's German Text "Die Axiome der Quantität und die Lehre vom Mass." *Journal of Mathematical Psychology*, *40*(3), 235–252. https://doi.org/10.1006/jmps.1996.0023
Pelton, T. W., & Bunderson, C. V. (2003). The recovery of the density scale using a stochastic quasi-realization of additive conjoint measurement. *Journal of Applied Measurement*, *4*(3), 269–281.
Perline, R., Wright, B. D., & Wainer, H. (1979). The Rasch Model as Additive Conjoint Measurement. *Applied Psychological Measurement*, *3*(2), 237–255. https://doi.org/10.1177/014662167900300213
Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Nielsen & Lydiche.
Richters, J. E. (2021). Incredible Utility: The Lost Causes and Causal Debris of Psychological Science: Basic and Applied Social Psychology: Vol 43, No 6. *Basic and Applied Social Psychology*, *43*(6), 366–405. https://www.tandfonline.com/doi/full/10.1080/01973533.2021.1979003
Salzberger, T. (2013). *Reporting a Rasch Analysis*. 347–362. https://doi.org/10.1002/9781118574454.ch19
Scheel, A. M., Tiokhin, L., Isager, P. M., & Lakens, D. (2021). Why Hypothesis Testers Should Spend Less Time Testing Hypotheses. *Perspectives on Psychological Science*, *16*(4), 744–755. https://doi.org/10.1177/1745691620966795
Scholten, Z. (2011). *The Guttman-Rasch Paradox in Item Response Theory* [PhD thesis]. https://www.semanticscholar.org/paper/The-Guttman-Rasch-Paradox-in-Item-Response-Theory-4-Scholten-A./815732abb7cfa43b6a585adb7db08d97e79e4116
Scott, D. (1964). Measurement structures and linear inequalities. *Journal of Mathematical Psychology*, *1*(2), 233–247. https://doi.org/10.1016/0022-2496(64)90002-1
Shah, A. K., & Oppenheimer, D. M. (2008). Heuristics made easy: An effort-reduction framework. *Psychological Bulletin*, *134*(2), 207–222. https://doi.org/10.1037/0033-2909.134.2.207
Shepherd, J. (2022). Conscious cognitive effort in cognitive control. *WIREs Cognitive Science*, *n/a*(n/a), e1629. https://doi.org/10.1002/wcs.1629
Steele, J. (2020). *What is (perception of) effort? Objective and subjective effort during attempted task performance*. PsyArXiv. https://doi.org/10.31234/osf.io/kbyhm
Steele, J., Pinto, M., Nosaka, K., & Nuzzo, J. L. (2022). *Perceptions of capacity, fatigue, and their psychophysics: Examining construct equivalence and the relationship between actual capacity and perception of capacity during resisted elbow flexion tasks*. PsyArXiv. https://doi.org/10.31234/osf.io/46vn5
Stenner, A., Fisher, W., Stone, M., & Burdick, D. (2013). Causal Rasch models. *Frontiers in Psychology*, *4*. https://www.frontiersin.org/articles/10.3389/fpsyg.2013.00536
Stephanou, A., & Fisher, W. P. (2013). From Concrete to Abstract in the Measurement of Length. *Journal of Physics: Conference Series*, *459*(1), 012026. https://doi.org/10.1088/1742-6596/459/1/012026
Stone, M. H., & Stenner, A. J. (2014). Comparison is key. *Journal of Applied Measurement*, *15*(1), 26–39.
Swaminathan, H., & Gifford, J. A. (1982). Bayesian Estimation in the Rasch Model. *Journal of Educational Statistics*, *7*(3), 175–191. https://doi.org/10.2307/1164643
Tafreshi, D. (2022). Sense and nonsense in psychological measurement: A case of problem and method passing

one another by. *Theory & Psychology*, *32*(1), 158–163. https://doi.org/10.1177/09593543211049371

Tiberghien, K., Notebaert, W., Smedt, B. D., & Fias, W. (2017). Reactive and Proactive Control in Arithmetical Strategy Selection. *Journal of Numerical Cognition*, *3*(3), 598–619. https://doi.org/10.5964/jnc.v3i3.124

Trendler, G. (2022). Is measurement in psychology an empirical or a conceptual issue? A comment on David Franz. *Theory & Psychology*, *32*(1), 164–170. https://doi.org/10.1177/09593543211050025

Waterbury, G. T. (2019). Missing Data and the Rasch Model: The Effects of Missing Data Mechanisms on Item Parameter Estimation. *Journal of Applied Measurement*, *20*(2), 154–166.

Westbrook, A., Bosch, R. van den, Määttä, J. I., Hofmans, L., Papadopetraki, D., Cools, R., & Frank, M. J. (2020). Dopamine promotes cognitive effort by biasing the benefits versus costs of cognitive work. *Science*, *367*(6484), 1362–1366. https://doi.org/10.1126/science.aaz5891

Westbrook, A., & Braver, T. S. (2015). Cognitive effort: A neuroeconomic approach. *Cognitive, Affective, & Behavioral Neuroscience*, *15*(2), 395–415. https://doi.org/10.3758/s13415-015-0334-y

Westbrook, A., Kester, D., & Braver, T. S. (2013). What Is the Subjective Cost of Cognitive Effort? Load, Trait, and Aging Effects Revealed by Economic Preference. *PLOS ONE*, *8*(7), e68210. https://doi.org/10.1371/journal.pone.0068210

Wright, B. D. (1986). *Bayes' Answer to Perfection.* https://www.rasch.org/memo38.pdf

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370. https://www.rasch.org/rmt/rmt83b.htm

Zerna, J., Scheffel, C., Kührt, C., & Strobel, A. (2022). *When easy is not preferred: A paradigm for estimating load-independent task preference.* PsyArXiv. https://doi.org/10.31234/osf.io/ysh3q