Archives of Rehabilitation Research and Clinical Translation (2020) **2**, 100055

ACRM
AMERICAN CONGRESS OF
REHABILITATION MEDICINE

Archives of Rehabilitation Research and Clinical Translation

Archives of Rehabilitation Research and Clinical Translation 2020;2:100055

Available online at www.sciencedirect.com

Original Research

The Template for Intervention Description and Replication as a Measure of Intervention Reporting Quality: Rasch Analysis

Marcel P. Dijkers, PhD, FACRM [a],
Scott R. Millis, PhD, ABPP, CStat [a,b]

[a] *Department of Physical Medicine and Rehabilitation, Wayne State University, Detroit, Michigan*
[b] *Department of Emergency Medicine, Wayne State University, Detroit, Michigan*

KEYWORDS
Guideline adherence;
Rehabilitation;
Reproducibility of results;
Research report

Abstract *Objective:* To determine whether the 12 items of the Template for Intervention Description and Replication (TIDieR) can be combined into a single summary score reflecting intervention reporting completeness and quality.

Design: Systematic review and reanalysis of published data. After a systematic search of the published literature, 16 review articles were retrieved with 489 sets of 12 TIDieR ratings of experimental intervention, comparator, or the 2 combined as reported in primary studies. These 489 sets were recoded into a common format and analyzed using Rasch analysis for binary items.

Setting: Not applicable.

Participants: Not applicable.

Interventions: Not applicable.

Main Outcome Measures: Psychometric qualities of a Rasch Analysis-based TIDieR summary score.

Results: The data fit the Rasch model. Infit and outfit values were generally acceptable (range, 0.70-1.45). TIDieR was reasonably unidimensional in its structure. However, the person (here: study) separation ratio was 1.25 with a corresponding reliability of 0.61. In addition, the confidence interval around each estimate of reporting completeness was wide (model standard error of 0.78)

Conclusion: Several Rasch indicators suggested that TIDieR is not a strong instrument for assessing the quality of a researcher's reporting on an intervention. It is recommended that it be used with caution. Improvements in TIDieR itself may make it more helpful as a reporting tool.

List of abbreviations: CONSORT, Consolidated Standards of Reporting Trials; MNSQ, mean square; N/A, not applicable; NM, no modifications; RCT, randomized controlled trial; TIDieR, Template for Intervention Description and Replication.

Disclosures: none.

Cite this article as: Arch Rehabil Res Clin Transl. 2020;2:100055.

https://doi.org/10.1016/j.arrct.2020.100055

Complaints about the quality of the reporting of the results of medical and health care research were first published in the early 1980s,[1-5] and the stream of critique has not abated. Poor reporting is relevant to science because it interferes with research users' ability to distinguish poor research (with presumably less reliable or even irrelevant findings) from good research. In a worst-case scenario, well-performed groundbreaking research is reported so poorly that it never has any effect.

One solution to poor reporting that has been attempted is the creation of checklists that tell authors which elements they should include in their papers, in what specific places. Some journals have made obligatory using and submitting the checklist appropriate for one's study design.[6] Checklists for randomized controlled trials (RCTs) were the first to be developed. The initial version of the Consolidated Standards of Reporting Trials (CONSORT) was published in 1997[7] and was followed by 18 subsidiary CONSORT standards, for example, for types of study design,[8] classes of interventions,[9] and report components.[10] For biobehavioral research investigating the effects of *complex interventions*, the CONSORT standards, premised on drug versus placebo studies, turned out to be deficient, and a CONSORT version for nonpharmacological trials[11] was published in 2008 and updated in 2017.[12] Even CONSORT version for nonpharmacological trials was judged to be insufficient to guide authors, and in 2014 the TIDieR *checklist and guide*, was published, offering guidance for describing rehabilitation, psychotherapy, and all other treatments that cannot be simply communicated with a drug name and dosage.[13] TIDieR offers a list of 12 items (table 1) which (added to CONSORT or to SPIRIT [Standard Protocol Items: Recommendations for Interventional Trials],[14] another *parent* to TIDieR) together are expected to "improve the reporting of interventions and make it easier for authors to structure accounts of their interventions, reviewers and editors to assess the descriptions, and readers to use the information"[13(p1)]

Even though CONSORT et al were created to assist authors in writing research reports, it did not take long for researchers to use these tools as measures to evaluate the reporting quality of entire batches of published reports.[15] There are now dozens of papers in the literature reporting on the quality of published papers, in general or before and after the publication of CONSORT or another landmark reporting checklist. The same fate has befallen TIDieR—there now are more than 3 dozen published papers that assess, using TIDieR, the literature in a certain area.

One of these papers, by Yamato et al,[16] used a sample of 200 reports of RCTs to assess the completeness of physical therapy research reporting, separately for the experimental intervention and the comparator. The authors concluded that reporting was typically incomplete, for both arms. In a more recent paper, they reanalyzed their data, exploring whether the 12 TIDieR items can be summed to create an interval scale of reporting completeness.[17] They argued that a simple summary score would be helpful in synthesizing the results of individual critical reviews[17] and facilitating the evaluation of strategies to improve intervention reporting. They rated each TIDieR item on a 0-1-2 scale, assigning 2 points if both intervention and comparator were described adequately, 1 point if either was, and 0 points if neither was. Then they conducted a Rasch analysis of the 2400 evaluations (12 ratings for 200 papers) and found that the data fit the Rasch model, targeted the sample well, with the items progressing in a logical order. Even so, they concluded, "The TIDieR summary score requires validation in an independent data set and, this could be carried out using the data generated in other evaluations of cohorts of articles using the TIDieR checklist."[17(p34)] The purpose of this study is to do so, using TIDieR ratings for papers in a variety of health care areas, including various rehabilitation disciplines.

Methods

In July 2019, we searched PubMed, Embase, PsycINFO, Web of Science, and CINAHL for any paper that in title or abstract used the term *Template for Intervention Description and Replication* or *TIDieR*. All abstracts were screened for the likelihood of authors using TIDieR as a tool to rate the reporting quality of full-text articles published in the peer-reviewed literature. The full texts of these were examined for reporting, either in a table in the text, or as supplemental appendix S1 (available online only at http://www.archives-pmr.org), ratings on all or most TIDieR items, for multiple papers. We found 16 studies with ratings of altogether 489 individual papers (table 2).

Most of these secondary studies applied TIDieR to the experimental intervention, although that was not always clearly stated. In most instances, communication with first authors allowed us to clarify what arm(s) TIDieR had been applied to; see supplemental appendix S1.

All authors used essentially a *reported [1] vs. not [adequately] reported [0]* rating scheme, with the exception of Picariello et al,[29] who used a *partial* rating in addition. We recoded the partial scores to 0. Other than Yamato et al,[17] we could not create a 0-1-2 coding scheme, because with 2 exceptions, only scores for 1 arm (generally, the treatment arm) were reported by the 16 authors. In addition, we considered that in future studies of the quality of intervention reporting in the literature, it would be useful to have a summary score for the comparator arm, separate from one for the intervention arm.

The primary studies used many codes of *not applicable* (N/A), especially for TIDieR items 9 (tailoring) and 10 (modifications). We decided that if there was no tailoring or

Table 1 The TIDieR checklist and Rasch analysis results for items

Item	Difficulty	Model SE	Infit MNSQ	Outfit MNSQ
T1 BRIEF NAME: Provide the name or a phrase that describes the intervention.	−3.72	0.25	0.95	0.95
T2 WHY: Describe any rationale, theory, or goal of the elements essential to the intervention.	−1.62	0.12	1.24	1.45
T3 WHAT: Materials: Describe any physical or informational materials used in the intervention, including those provided to participants or used in intervention delivery or in training of intervention providers. Provide information on where the materials can be accessed (eg, online appendix, URL).	0.34	0.10	1.02	1.24
T4 WHAT: Procedures: Describe each of the procedures, activities, and/or processes used in the intervention, including any enabling or support activities.	−0.92	0.11	1.04	1.00
T5 WHO PROVIDED: For each category of intervention provider (eg, psychologist, nursing assistant), describe their expertise, background, and any specific training given.	0.56	0.10	0.96	0.90
T6 HOW: Describe the modes of delivery (eg, face-to-face or by some other mechanism, such as internet or telephone) of the intervention and whether it was provided individually or in a group.	0.05	0.10	0.97	0.99
T7 WHERE: Describe the type(s) of location(s) where the intervention occurred, including any necessary infrastructure or relevant features.	0.38	0.10	1.03	1.04
T8 WHEN and HOW MUCH: Describe the number of times the intervention was delivered and over what period of time including the number of sessions, their schedule, and their duration, intensity, or dose.	−0.15	0.10	0.83	0.76
T9 TAILORING: If the intervention was planned to be personalized, titrated, or adapted, then describe what, why, when, and how.	1.06	0.11	1.06	1.03
T10 MODIFICATIONS: If the intervention was modified during the course of the study, describe the changes (what, why, when, how).	2.45	0.16	1.01	1.10
T11 HOW WELL—Planned: If intervention adherence or fidelity was assessed, describe how and by whom, and if any strategies were used to maintain or improve fidelity, describe them.	1.04	0.12	0.90	0.83
T12 HOW WELL—Actual: If intervention adherence or fidelity was assessed, describe the extent to which the intervention was delivered as planned.	0.53	0.11	0.99	0.94
Mean	0.00	0.12	1.00	1.02
SD	1.48	0.04	0.10	0.18

Table 2 Key information on the 16 studies contributing data

Author Name	Domain of Intervention	Study Designs Included	Years of Publication of Papers Extracted	Number of Experimental Intervention Arms Rated	Number of Comparator Intervention Arms Rated	Number of Combined Experimental and Comparator Arms Rated	Number of Original TIDieR Items Used	TIDieR Rating Criterion
Baron et al[18]	Skin-care self-management interventions for people with spinal cord injury	Randomized and nonrandomized trials	1974-2016	17	0	0	12	Fully reported
Bartholdy et al[19]	Exercise for knee osteoarthritis	Any design	1982-2012	133	0	0	12	Completeness of reporting sufficient for replication
Comer et al[20]	Non-pharmacological interventions for non-inflammatory multi-joint pain	Randomized and non-randomized trials, pre- post designs	2006-2015	4	0	0	12	Not stated
Nascimento et al[21]	Interventions for non-specific low back pain	RCTs	1998-2018	18	18	0	12	Completeness of description
Grudniewicz et al[22]	Printed educational materials to improve primary care physicians' knowledge, behavior, and patient outcomes	RCTs, quasi-randomized trials, controlled pre-post studies, interrupted time series	1983-2014	32	0	0	12	Completeness of reporting and replicability
Gspörer and Schrems[23]	Long-term care nursing for elderly people	Any design	2015	0	0	22	12	Completeness of reporting
Hacke et al[24]	Exercise for hypertension	RCTs	1980-2010	0	0	23	12	Completeness of intervention description
Howlett et al[25]	Physical activity interventions for inactive healthy adults	RCTs	1998-2016	26	19	0	12	Not stated
Knols et al[26]	Exercise for lung transplant recipients	RCTs	2003-20017	7	0	0	12	Not stated
Liljeberg et al[27]	Oral nutritional supplements	RCTs	2002-2015	58	17	1	12	Completeness of reporting
Mackie et al[28]	Promotion of family involvement in the care of hospitalized patients	Any design, including qualitative	2003-2014	11	0	0	12	(Sufficiently) reported
Picariello et al[29]	Social-psychological interventions for fatigue in end-stage kidney disease	RCTs and quasi-RCTs	2000-2015	16	0	0	12	Adequately reported

Ross et al[30]	Exercise for posterior tibial tendon dysfunction in adults	RCTs	2008-2015	3	0	0	12	Completely reported
Stevens et al[31]	Advice for low back pain selfcare	RCTs	1987-2015	29	0	0		Completeness of reporting
		Guidelines	2003-2014	14	0	0	9	
Sykes et al[32]	Audit and feedback to improve dementia care	Any longitudinal design	2002-2015	13	0	0	8	Not stated
Zandstra et al[33]	Shared decision making and care for women with heavy menstrual bleeding	Any design	1999-2007	8	0	0	12	Quality of description and reproducibility

no modification in a primary study, the authors ought to say so explicitly in their report, and a code of 1 could be given. (If there is individualization or revision, TIDieR instructs to report the details; see table 1.) A similar decision was made with respect to the code *?*, which occurred in about 10% of the ratings for items 10, 11 (strategies to improve fidelity and fidelity assessment methods), and 12 (actual fidelity). Arguing that a description that is unclear enough to make the rater question what rating to give is an insufficient description, we recoded these too to 0. Blank cells in the lists of published ratings, reflecting that authors had not used a particular TIDieR item, were left blank in the data file we created, and in the Rasch analysis (see below) these did not contribute information to the analysis for the item in question. Supplemental appendix S1 lists for each of the 16 papers what the nature of the published data was and how we recoded information to create a single file containing all 489 sets of up to 12 TIDieR ratings. This data file is available in supplemental appendix S1.

Winsteps[a] software was used to perform Rasch analysis. The TIDieR items had 2 categories: 0 (not [adequately] reported) and 1 (adequately reported). We therefore used the dichotomous Rasch model to evaluate the psychometric characteristics of TIDieR used as a scale applied to studies (rather than the traditional persons) assessing its unidimensionality, reliability, and targeting.[34] To assess structure and unidimensionality, we examined infit and outfit mean squares (MNSQs). Infit and outfit MNSQ values in the range of 0.6-1.4 are considered acceptable for rating scales and surveys.[34] We also assessed unidimensionality by using infit and outfit values, as well as Rasch principal component analysis. Local dependency was assessed using Yen's Q3 statistic, where highly locally dependent items will have correlations exceeding 0.70.[35] We evaluated the internal consistency of person (here: study) and item performance by examining separation reliability estimates and separation ratio. Separation reliability for persons (here, papers) refers to the consistency of a paper's *responses* across items, whereas the separation reliability for items refers to the consistency of TIDieR item performances across papers. Targeting was addressed using visual inspection of the person-item map. A measure with good targeting for the cases to be rated results in a map symmetric along the vertical axis with items and persons (papers) clustered in a similar fashion with a similar range.

We considered conducting, in addition, separate Rasch analyses of TIDieR ratings of experimental interventions, of comparators, and of treatments and their comparators combined, but decided that the numbers available for the latter 2 categories (37 and 22, respectively) were too small to result in reliable findings as to differences in Rasch results between the 3 (eg, regarding difficulty of individual items).

The *measure* (ie, TIDieR reporting adequacy score) reported by Winsteps for each of the 489 papers was merged back with data extracted from the 16 papers literature and further processed using SPSS[b] to calculate mean scores by secondary study and trial arm.

Results

The 16 secondary studies from which data were used covered a variety of health care research areas, research

```
MEASURE    PERSON - MAP - ITEM
               <more>|<rare>
    4           .##  +
                     |
                     |
                     |
                     |
                     |
                     |
    3            .#  +T
                     |
                     |
                  .  |
                    T|  T10
                     |
               .###  |
    2                +
                     |
                     |
              .####  |
                     |S
                  . S|
                     |
    1       .######  +  T11    T9
                  .  |
                     |
        .##########  |  T12    T5
                     |  T7
                     |  T3
       .########### M|
    0             .  +M T6
                     |  T8
          #########  |
                  .  |
                  .  |
                     |
         .#########  |  T4
   -1               S+
                  .  |
                  .  |
             ######  |S
                     |  T2
                     |
                     |
   -2             .  +
                 ##  |
                    T|
                     |
                     |
                     |
                     |
   -3                +T
                     |
                  .  |
                  .  |
                     |
                     |  T1
                     |
   -4                +
               <less>|<freq>
```

Fig 1 Person-item map of TIDieR items and primary studies. Notes: # and . represent studies. Each # is 7 studies; each . is 1-6 studies. T1-T12 represent the 12 TIDieR items.

designs, years of publication, and number of primary studies evaluated (see table 2).

Rasch analysis infit and outfit MNSQs along with item difficulty (measure) for each item are reported in table 1. Item 1 was the most adequately reported TIDieR item, whereas Item 10 was the least frequently endorsed. Infit and outfit values were generally acceptable (range, 0.70-1.45), suggesting that the items fit the Rasch model expectation of unidimensionality. Examination of the Rasch residual-based principal components showed that the total variance explained by the measures was only 35%. However, the expected variance (the variance that would likely be explained if the data fit the Rasch model exactly) also was 35%. In addition, the first contrast eigenvalue was 1.99 and the disattenuated correlations of person measures on different clusters of items were 1.0. Taken together, this pattern of findings suggests that TIDieR is reasonably unidimensional in its structure.

We found a person (paper) separation ratio of 1.25 with a corresponding reliability of 0.61. The assumption of local item independence did not appear to be violated. None of the correlations between items and standardized residuals exceeded 0.70. The item-person (paper) map is shown in fig 1. The map is symmetric along the vertical axis with items and persons (papers) clustered in a similar fashion with a similar range, indicating good targeting.

Table 3 provides mean TIDieR summary scores (measure, in the Winsteps report) and mean model SE (also from Winsteps) by study and (if applicable) by TIDieR ratings subgroup (experimental treatment only, comparator only, both) in a study. The last line indicates that the average TIDieR summary score of the 489 studies was 0.19, with a standard deviation of 1.30. However, the score of the average study had a model SE of 0.78, indicating that the *confidence interval* for each estimate was wide (running from −0.59 [0.19-0.78] to 0.97 [0.19+0.78]).

There was a great difference in mean TIDieR scores between studies or groups, from 2.08 for Stevens' primary studies,[31] to −1.48 for Stevens' clinical practice guidelines.[31] At the bottom of table 3, mean scores are provided for ratings of treatment, comparator and combined arms. This information suggests that in the average paper, the report of the experimental arm was somewhat better (0.14, on average) than that of the comparator (−0.22).

Discussion

Our analysis confirms the findings of the Yamato et al study,[17] that the TIDieR items satisfy the assumptions of Rasch analysis and can be combined into an interval level summary score. This confirmation occurred despite the fact that we made some major changes in the methodology, including scoring individual arms (rather than a combination of the 2 arms in an RCT) and assigning a score reflecting *no (adequate) report* to all entries of *N/A*, *NM*, or *?* in the 16 studies that contributed data. However, based on the various indicators reported (person [paper] separation ratio of 1.25, paper reliability of 0.61, and model SE value of 0.78, on average), we

Table 3 Mean and standard deviation for measure and model SE, by secondary study and subgroup

Study/Subgroup	Measure Mean ± SD	Model SE Mean ± SD	n
Baron et al exp[18]	0.28 ±0.88	0.70 ±0.07	17
Bartholdy et al exp[19]	−0.56 ±0.77	0.73 ±0.07	133
Comer et al exp[20]	−0.72 ±1.28	0.77 ±0.14	4
Nascimento et al exp[21]	0.14 ±0.64	0.68 ±0.04	18
Nascimento et al comp[21]	−0.23 ±0.93	0.71 ±0.14	18
Grudniewicz et al exp[22]	−0.03 ±1.09	0.72 ±0.13	32
Gspörer and Schrems both[23]	1.88 ±1.33	0.94 ±0.41	22
Hacke et al exp[24]	1.09 ±0.90	0.74 ±0.13	23
Howlett et al exp[25]	0.66 ±0.83	0.70 ±0.06	26
Howlett et al comp[25]	−0.22 ±1.13	0.73 ±0.15	19
Knols et al exp[26]	−1.18 ±0.59	0.79 ±0.09	7
Liljeberg et al exp[27]	0.21 ±1.00	0.71 ±0.10	76
Mack et al exp[28]	0.22 ±1.27	0.73 ±0.09	11
Picariello exp[29]	0.34 ±1.02	0.71 ±0.06	16
Ross et al exp[30]	2.20 ±2.70	1.24 ±0.60	3
Stevens et al exp[31]	2.08 ±1.22	1.26 ±0.47	29
Stevens et al cpg[31]	−1.48 ±1.07	0.93 ±0.18	14
Sykes et al exp[32]	1.19 ±1.43	0.99 ±0.31	13
Zandstra et al exp[33]	1.38 ±0.68	0.74 ±0.08	8
Both experimental and comparator arm are described	1.88 ±1.33	0.94 ±0.41	22
Comparator arm only is described	-0.22 ±1.02	0.72 ±0.14	37
Experimental arm only is described	0.14 ±1.26	0.77 ±0.22	430
Total	0.19 ±1.30	0.78 ±0.23	489

Abbreviations: both, experimental and comparator arms rated together; comp, comparator arm; cpg, clinical practice guideline; exp, experimental arm.

conclude that the Rasch-calculated score is a poor measure of reporting quality and completeness; it basically allows one to distinguish 2 strata of papers only, good and poor ones.

Yamato did not come to this conclusion; they decided that their Rash analysis "supports the use of a summary TIDieR score as an indicator of completeness of reporting."[17(p34)] Because they used a trichotomy for scoring items and we a dichotomy, and the fact that they used 11 items only (item 10 was omitted as being a constant across 200 papers), the numerical results are not directly comparable. However, some of the outfit MNSQs they report, of 0.47 (for item 11) and 0.19 (for item 1) suggest problematic items. So do the disordering of thresholds and a person (paper) reliability index of 0.62. They note the same restriction as we do ("the summary score may only be able to discriminate between the least and most detailed reports"[17(p31)]), but they are much more optimistic that this should not stand in the way of use of TIDieR as a rating scale.

The poor Rasch results we obtained may be the consequence of a number of factors, including the following: (1) the nature of TIDieR, designed as a checklist rather than a measurement tool per se; (2) our decisions on how to handle scores of ?, N/A, and NM; (3) differences in rater severity between the various secondary studies; and (4) some secondary studies not using all 12 TIDieR items, and this pattern possible being associated with rater severity and/or the application of N/A, and so on.

This all does not mean that TIDieR as developed originally, a checklist to assist authors of intervention reports, is a poor tool; converting it to a quality scoring measure is, as of now, problematic. As Yamato stated, reviewers who want to assess the intervention reporting quality of the research literature in a particular area, or determine whether a particular intervention (eg, requiring the submission of a reporting checklist along with a research paper)[6] has resulted in improved reporting, may want to have a simple single summary score, rather than working with the 12 TIDieR items separately. However, we doubt that such a summary score ever is sufficient, even if it can be created. Any such analysis will want to focus on the specifics, we think: what particular reporting areas are strong, which ones have improved the least, etc. If a reliable summary score can be created, it presumably will primarily have a supplemental use.

The ease with which Rasch analysis produces scores on an interval measure may seduce one into believing that the TIDieR checklist is reliable as a measure and has utility in differentiating, with a high level of precision, between multiple strata of reporting completeness and quality. As a checklist, TIDieR can be improved by providing explicit guidance to authors for what to do in an instance of something not being planned or observed: they should be told to write something like "there was no tailoring; no modification; no ...." Then readers and report assessors do not need to guess what happened, or read the tea leaves of ambiguous statements.

Many secondary study authors (whether included in our analysis[18,19] or not[36,37]) have distinguished *subitems* under various TIDieR items, to make sure that they determined whether issues that could be distinguished within an item were reported or not. There now is a tool similar to TIDieR, Consensus on Exercise Reporting Template,[38] that has a lot of overlap in terms of items, but uses subitems for a few of them. Distinguishing *smaller* items makes it easier for raters to evaluate an intervention and for rating teams to determine causes of (dis)agreement with an independent second rater. A *finer grained* TIDieR checklist might result in a better Rasch measure, but it would increase the burden on authors to report what was done or not done and how exactly it was done. Consensus on Exercise Reporting Template was designed for the assessment of the completeness of exercise studies, and there now also is a TIDieR version for public health intervention studies.[39] Presumably, additional specialized versions will follow, and authors of studies applying them to published intervention research are advised to take a careful look at subitems and how they should be scored—following the rule that each subitem needs to be satisfied before the parent item is considered satisfied, or otherwise.

A last issue is the satisfactoriness of TIDieR per se in assisting authors in describing what is or are the *active ingredient(s)* in their interventions. For duplicators, it is nice to know what the name of the treatment is they are duplicating (item 1), but that knowledge by itself does not affect the participants in the replication intervention study. Nor do the descriptions of What-procedures and What-materials (items 4 and 3, respectively). The Hawthorne studies should be a reminder of the fact that often we have little idea of what actually is the cause of changes in clients or patients. Various research teams have started investigations of the (classification of) active ingredients in complex interventions (eg, Michie et al[40-42] and Dijkers et al[43-45]), but the authors of TIDieR seem not to have relied on this work. Whyte et al[46-49] most recently published a series of papers on the Rehabilitation Treatment Specification System that make clear how involved and painfully slow is the task of identifying and adequately describing a single intervention, let alone a bundling of treatments into what we traditionally call an intervention (or comparator).[46-49]

Study limitations

We followed Yamato et al[17] in assuming that TIDieR is a reflective measure, and that therefore conducting a Rasch analysis makes sense to begin with, which is not necessarily the case. Research reports may not have an innate trait that can be called *reportiveness* and that determines with great certainty the satisfactoriness of answers to the 12 TIDieR items. If the quality of the text of a paper is determined by a large number of *random* factors (including journal page limitations, standards applied by editor and peer reviewers, authors' urgency to publish or perish) and TIDieR constitutes a measure created with effect indicators, Rasch analysis has no role, and other means of determining a reporting quality total score need to be used.

We were dependent on sets of TIDieR ratings we found in the literature using a systematic search, but know that there are at least 20 more studies that did not publish their raw TIDieR scores. Some of the available reviews were very small (in terms of the number of primary studies rated), and much of the information as to what was rated (experimental treatment only, or experimental and comparator), and how, was missing or at least ambiguous. If we erred in our decisions that the TIDieR ratings in a particular study applied to the experimental treatment only (and not to the intervention and comparator combined), that does not invalidate the results. Some may disagree with the method we selected to handle scores such as N/A, NM, and ?, and an alternative approach may result in somewhat different results. Last, in our analysis we did not take into account the clustering of primary studies within secondary ones, because the Winsteps program does not offer such an option.

It should be considered that only 2 papers (Nascimento et al[21] and Howlett et al[25]) contributed to the comparator appraisal subgroup, so it is possible that rater severity in these 16 studies was confounded with arm rated. The same holds true for the average report on both intervention and control arm combined, which is based on just 1 secondary study (Gspörer and Schrems[23]).

Conclusions

Our analysis of scores on the 12 TIDieR items reported for 489 primary studies in 16 reviews suggests that the unidimensionality assumption underlying Rasch analysis are satisfied. However, multiple indicators suggest that the TIDieR summary score does not offer a strong instrument for rating the overall quality of reporting on an intervention. Consequently, rehabilitation researchers should be careful in depending on this TIDieR summary score for an analysis of the quality of intervention reporting, over time or comparing between various areas of complex interventions. Improvement in TIDieR itself (subitems, better instructions for reporting absences) may enhance it as a checklist helping authors.

Suppliers

a. Winsteps; Winsteps Software Technologies.
b. SPSS; IBM Corp.

Corresponding author

Marcel P. Dijkers, PhD, FACRM, Department of Physical Medicine and Rehabilitation, Wayne State University, 1450 Wiltshire Road, Berkley, MI 48201. *E-mail address:* marcellinus.p.dijkers@wayne.edu.

Acknowledgments

We thank Justine Barron, PhD, Cecilie Bartholdy, PT, MSc, Claudia Hacke, PhD, Evelina Liljeberg, RD, MSc, Benjamin Mackie, RN, PhD, Federica Picariello, PhD, Megan Ross, PT, MSc, Matthew Stevens, MChiroprac, and Michael Sykes, RN, PhD, for communications clarifying their published papers.

References

1. Baar J, Tannock I. Analyzing the same data in two ways: a demonstration model to illustrate the reporting and misreporting of clinical trials. J Clin Oncol 1989;7:969-78.
2. Sewerin I. Reporting radiographic methods in dental epidemiologic and experimental studies. Community Dent Oral Epidemiol 1986;14:90-3.
3. Sonis J, Joines J. The quality of clinical trials published in the *Journal of Family Practice*, 1974-1991. J Fam Pract 1994;39:225-35.
4. Zola P, Volpe T, Castelli G, et al. Is the published literature a reliable guide for deciding between alternative treatments for patients with early cervical cancer? Int J Radiat Oncol Biol Phys 1989;16:785-97.
5. Liberati A, Himel HN, Chalmers TC. A quality assessment of randomized control trials of primary treatment of breast cancer. J Clin Oncol 1986;4:942-51.
6. Chan L, Heinemann AW, Roberts J. Elevating the quality of disability and rehabilitation research: mandatory use of the reporting guidelines. Arch Phys Med Rehabil 2014;95:415-7.
7. Freemantle N, Mason JM, Haines A, Eccles MP. CONSORT: an important step toward evidence-based health care. Consolidated Standards of Reporting Trials. Ann Int Med 1997;126:81-3.
8. Vohra S, Shamseer L, Sampson M, et al. CONSORT extension for reporting N-of-1 trials (CENT) 2015 Statement. J Clin Epidemiol 2016;76:9-17.
9. Gagnier JJ, Boon H, Rochon P, Moher D, Barnes J, Bombardier C. Recommendations for reporting randomized controlled trials of herbal interventions: explanation and elaboration. J Clin Epidemiol 2006;59:1134-49.
10. Ioannidis JP, Evans SJ, Gotzsche PC, et al. Better reporting of harms in randomized trials: an extension of the CONSORT statement. Ann Int Med 2004;141:781-8.
11. Boutron I, Moher D, Altman DG, Schulz KF, Ravaud P. Extending the CONSORT statement to randomized trials of nonpharmacologic treatment: explanation and elaboration. Ann Int Med 2008;148:295-309.
12. Boutron I, Altman DG, Moher D, Schulz KF, Ravaud P. CONSORT Statement for randomized trials of nonpharmacologic treatments: a 2017 update and a CONSORT extension for nonpharmacologic trial abstracts. Ann Int Med 2017;167:40-7.
13. Hoffmann TC, Glasziou PP, Boutron I, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. BMJ 2014;348:g1687.
14. Chan AW, Tetzlaff JM, Altman DG, et al. SPIRIT 2013 statement: defining standard protocol items for clinical trials. Ann Int Med 2013;158:200-7.
15. Ah-See KW, Molony NC. A qualitative assessment of randomized controlled trials in otolaryngology. J Laryngol Otol 1998;112:460-3.
16. Yamato TP, Maher CG, Saragiotto BT, Hoffmann TC, Moseley AM. How completely are physiotherapy interventions described in reports of randomised trials? Physiotherapy 2016;102:121-6.
17. Yamato TP, Maher CG, Saragiotto BT, Catley MJ, Moseley AM. Rasch analysis suggested that items from the template for intervention description and replication (TIDieR) checklist can be summed to create a score. J Clin Epidemiol 2018;101:28-34.
18. Baron JS, Sullivan KJ, Swaine JM, et al. Self-management interventions for skin care in people with a spinal cord injury: part 2-a systematic review of use of theory and quality of intervention reporting. Spinal Cord 2018;56:837-46.
19. Bartholdy C, Nielsen SM, Warming S, Hunter DJ, Christensen R, Henriksen M. Poor replicability of recommended exercise interventions for knee osteoarthritis: a descriptive analysis of evidence informing current guidelines and recommendations. Osteoarthritis Cartilage 2019;27:3-22.
20. Comer C, Smith TO, Drew B, Raja R, Kingsbury SR, Conaghan PG. A systematic review assessing non-pharmacological conservative treatment studies for people with non-inflammatory multi-joint pain: clinical outcomes and research design considerations. Rheumatol Int 2018;38:331-41.
21. Nascimento P, Costa LOP, Araujo AC, Poitras S, Bilodeau M. Effectiveness of interventions for non-specific low back pain in older adults. A systematic review and meta-analysis. Physiotherapy 2019;105:147-62.
22. Grudniewicz A, Kealy R, Rodseth RN, Hamid J, Rudoler D, Straus SE. What is the effectiveness of printed educational materials on primary care physician knowledge, behaviour, and patient outcomes: a systematic review and meta-analyses. Implement Sci 2015;10:164.
23. Gspörer I, Schrems BM. [Transparency and replicability of nursing intervention studies in long-term care: A selective literature review] [German]. Z Evid Fortbild Qual Gesundhwes 2018;133:1-8.
24. Hacke C, Nunan D, Weisser B. Do exercise trials for hypertension adequately report interventions? A reporting quality study. Int J Sports Med 2018;39:902-8.
25. Howlett N, Trivedi D, Troop NA, Chater AM. Are physical activity interventions for healthy inactive adults effective in promoting behavior change and maintenance, and which behavior change techniques are effective? A systematic review and meta-analysis. Transl Behav Med 2019;9:147-57.
26. Knols RH, Fischer N, Kohlbrenner D, Manettas A, de Bruin ED. Replicability of physical exercise interventions in lung transplant recipients: a systematic review. Front Physiol 2018;9:946.
27. Liljeberg E, Andersson A, Lovestam E, Nydahl M. Incomplete descriptions of oral nutritional supplement interventions in reports of randomised controlled trials. Clin Nutr 2018;37:61-71.
28. Mack BR, Mitchell M, Marshall PA. The impact of interventions that promote family involvement in care on adult acute-care wards: an integrative review. Collegian 2018;25:131-40.
29. Picariello F, Hudson JL, Moss-Morris R, Macdougall IC, Chilcot J. Examining the efficacy of social-psychological interventions for the management of fatigue in end-stage kidney disease (ESKD): a systematic review with meta-analysis. Health Psychol Rev 2017;11:197-216.
30. Ross MH, Smith MD, Mellor R, Vicenzino B. Exercise for posterior tibial tendon dysfunction: a systematic review of randomised clinical trials and clinical guidelines. BMJ Open Sport Exerc Med 2018;4:e000430.
31. Stevens ML, Lin CC, de Carvalho FA, Phan K, Koes B, Maher CG. Advice for acute low back pain: a comparison of what research supports and what guidelines recommend. Spine J 2017;17:1537-46.
32. Sykes MJ, McAnuff J, Kolehmainen N. When is audit and feedback effective in dementia care? A systematic review. Int J Nurs Stud 2018;79:27-35.
33. Zandstra D, Busser JAS, Aarts JWM, Nieboer TE. Interventions to support shared decision-making for women with heavy menstrual bleeding: a systematic review. Eur J Obstet Gynecol Reprod Biol 2017;211:156-63.
34. Bond TG, Fox CM. Applying the Rasch model: Fundamental measurement in the human sciences. 3rd ed. New York: Routledge/Taylor & Francis Group; 2015.
35. Yen WM. Scaling performance assessments: strategies for managing local item dependence. J Educ Meas 1993;30:187-213.

36. Tew GA, Brabyn S, Cook L, Peckham E. The completeness of intervention descriptions in randomised trials of supervised exercise training in peripheral arterial disease. PLoS One 2016; 11:e0150869.
37. Yu AM, Balasubramanaiam B, Offringa M, Kelly LE. Reporting of interventions and "standard of care" control arms in pediatric clinical trials: a quantitative analysis. Pediatr Res 2018;84:393-8.
38. Slade SC, Dionne CE, Underwood M, Buchbinder R. Consensus on exercise reporting template (CERT): explanation and elaboration statement. Br J Sports Med 2016;50:1428-37.
39. Campbell M, Katikireddi SV, Hoffmann T, Armstrong R, Waters E, Craig P. TIDieR-PHP: a reporting guideline for population health and policy interventions. BMJ 2018;361:k1079.
40. Johnston M, Johnston D, Wood CE, Hardeman W, Francis J, Michie S. Communication of behaviour change interventions: can they be recognised from written descriptions? Psychol Health 2018;33:713-23.
41. Michie S, Carey RN, Johnston M, et al. From theory-inspired to theory-based interventions: a protocol for developing and testing a methodology for linking behaviour change techniques to theoretical mechanisms of action. Ann Behav Med 2018;52: 501-12.
42. Wood CE, Hardeman W, Johnston M, Francis J, Abraham C, Michie S. Reporting behaviour change interventions: do the behaviour change technique taxonomy v1, and training in its use, improve the quality of intervention descriptions? Implement Sci 2016;11:84.
43. Dijkers M, Hart T, Whyte J, MZ J, Packel A, Tsaousides T. Rehabilitation treatment taxonomy: implications and continuations. Arch Phys Med Rehabil 2014;95. S45-54.e42.
44. Hart T, Tsaousides T, Zanca JM, et al. Toward a theory-driven classification of rehabilitation treatments. Arch Phys Med Rehabil 2014;95. S33-44.e32.
45. Whyte J, Dijkers MP, Hart T, et al. Development of a theory-driven rehabilitation treatment taxonomy: conceptual issues. Arch Phys Med Rehabil 2014;95. S24-32.e22.
46. Hart T, Dijkers MP, Whyte J, et al. A theory-driven system for the specification of rehabilitation treatments. Arch Phys Med Rehabil 2019;100:172-80.
47. Zanca JM, Turkstra LS, Chen C, et al. Advancing rehabilitation practice through improved specification of interventions. Arch Phys Med Rehabil 2019;100:164-71.
48. Whyte J, Dijkers MP, Hart T, et al. The importance of voluntary behavior in rehabilitation treatment and outcomes. Arch Phys Med Rehabil 2019;100:156-63.
49. Van Stan JH, Dijkers MP, Whyte J, et al. The rehabilitation treatment specification system: implications for improvements in research design, reporting, replication, and synthesis. Arch Phys Med Rehabil 2019;100:146-55.