當代教育研究季刊　第二十五卷　第四期，2017 年 12 月，頁 207-231
DOI : 10.6151/CERQ.2017.2504.06

《研究紀要》

從評分者間信度與實驗處理忠實度評析日本與臺灣自閉症單一個案研究品質

Ee Rea Hong　蔡欣玶*　陳佩玉
龔麗媛　Jennifer B. Ganz

摘　要

近年來國內外皆倡導選擇有實證研究支持其介入成效的策略或方案，目前國外已有數篇文獻聚焦於探討自閉症研究品質，然而這些文獻所分析的研究主要發表於英語系期刊中，故結論無法推論至發表於非英語系期刊的研究品質。為了瞭解非英語系期刊中 自閉症研究的品質，本研究採系統性文獻回顧，分析日本與臺灣主要之特殊教育學術研究導向期刊，以探討其刊登之所有自閉症單一個案研究的評分者間信度與實驗處理忠實度之現況與趨勢。本研究依（1）從日本與臺灣共七份主要特殊教育研究期刊中，搜尋文獻；（2）依預先設定的文獻選取與排除準則，進行研究評估；（3）針對文獻中描述之評分者間信度與實驗處理忠實度，進行資料編碼與分析等三步驟，最後進行 194 篇符合選取準則之文獻評析。本研究結果如下：在評分者間信度方面，日本於過去數十年間符合單一個案研究評分者間信度標準的研究數量，整體而言呈現增加的趨勢，臺灣文獻則

Ee Rea Hong，日本筑波大學人間科學院助理教授
電子郵件：irehong@human.tsukuba.ac.jp
* 蔡欣玶（通訊作者），國立臺北教育大學特殊教育學系助理教授
電子郵件：sptsai@tea.ntue.edu.tw
陳佩玉，國立臺北教育大學特殊教育學系副教授
電子郵件：pychen@tea.ntue.edu.tw
龔麗媛，日本筑波大學人間科學院博士生
電子郵件：gongliyuan89@yahoo.co.jp
Jennifer B. Ganz，美國德州農工大學卡城分校教育與人類發展學院教授
電子郵件：jeniganz@tamu.edu
投稿日期：2017 年 6 月 30 日；修正日期：2017 年 10 月 15 日；接受日期：2017 年 12 月 7 日

未呈現同樣的趨勢；在實驗處理忠實度方面，日本尚未有單一個案研究描述研究的實驗處理忠實度，而臺灣則已有數篇研究符合國際間檢視實驗處理忠實度資料的標準。就整體研究品質而論，日本與臺灣的自閉症單一個案研究已逐漸符合評分者信度間的標準，然而對於實驗處理忠實度的敘述仍處於初步發展階段。為了提升自閉症介入方案的研究品質，本研究建議在亞洲地區的相關學術研究論文應依建議的指標與標準說明各研究之評分者間信度與實驗處理忠實度。

關鍵詞：自閉症、單一個案研究、評分者間信度、實驗處理忠實度、日本、臺灣

Contemporary Educational Research Quarterly
December, 2017, Vol. 25, No. 4, pp. 207-231
DOI : 10.6151/CERQ.2017.2504.06

《Research Note》

THE QUALITY OF SINGLE-CASE RESEARCH FOR INDIVIDUALS WITH AUTISM SPECTRUM DISORDERS IN JAPAN AND TAIWAN: AN INVESTIGATION OF INTER-RATER RELIABILITY AND TREATMENT FIDELITY

Ee Rea Hong Shin-Ping Tsai* Pei-Yu Chen
Li-yuan Gong Jennifer B. Ganz

ABSTRACT

Given this era of identification of evidence-based practices, previous reviews have provided a snapshot of the current status of the research quality of autism spectrum disorder (ASD)-focused studies. However, including only studies in English language journals does not represent the quality of the literature published in journals of languages other than English. To evaluate the overall quality of ASD intervention research in non-English-language journals, this systematic review summarizes the current status and trends of inter-rater reliability (IRR) and treatment fidelity in ASD-focused single-case research published in major academically oriented Japanese and Taiwanese special education journals. To conduct this review, the following three steps were taken: (a) literature search of the seven prominent special education journals in Japan and Taiwan, (b) assessment of potential studies against pre-set inclusion and exclusion criteria, and (c) study coding and analysis of descriptive study characteristics and measures of IRR and treatment fidelity. A

Ee Rea Hong, Assistant Professor, University of Tsukuba, Tokyo, Japan.
E-mail: irehong@human.tsukuba.ac.jp

* Shin-Ping Tsai (corresponding author), Assistant Professor, National Taipei University of Education, Taipei, Taiwan.
E-mail: sptsai@tea.ntue.edu.tw

Pei-Yu Chen, Associate Professor, National Taipei University of Education, Taipei, Taiwan.
E-mail: pychen@tea.ntue.edu.tw

Li-yuan Gong, Doctoral Student, University of Tsukuba, Tokyo, Japan.
E-mail: gongliyuan89@yahoo.co.jp

Jennifer B. Ganz, Professor, Texas A&M University, College Station, Texas, USA.
E-mail: jeniganz@tamu.edu

Manuscript received: June 30, 2017; Modified: October 15, 2017; Accepted: December 7, 2017

total of 194 articles met the inclusion criteria for the review. The results show that an overall increasing trend in the number of articles that reported IRR data with acceptable levels were observed over time in the Japanese journals while no such trend was found in the Taiwanese journals. In contrast, it was found that no article published in the Japanese journals had reported treatment fidelity data while a small number of articles that reported treatment fidelity data with acceptable quality degrees were observed in the Taiwanese journals. As to the overall quality of ASD research, researchers in Japan and Taiwan are increasingly attending to quality with regard to collecting and reporting acceptable IRR data for outcome variables in their single-case autism-related articles. Yet, the evaluation of treatment fidelity and its IRR is still at the initial stage. To improve the overall quality of ASD intervention research, efforts should be made to report both IRR and treatment fidelity data based on the suggested standards with acceptable quality degrees in Asian-language journals.

Keywords: *autism spectrum disorder, single-case research, inter-rater reliability, treatment fidelity, Japan, Taiwan*

Introduction

Autism spectrum disorder (ASD) is among the most common neurodevelopmental disorders and is characterized by impairments in social interaction and communication with restricted and repetitive patterns of behaviors and interests (American Psychiatric Association, 2013). ASD appears to affect approximately 1 of every 68 children, aged 8 years in the United States (Centers for Disease Control and Prevention, 2014), and a rise in the prevalence of ASD has been reported world-wide (Elsabbagh et al., 2012), including many Asian countries. For example, in Japan, the estimates of prevalence of ASD range from about 37.5 to 181.1 per 10,000 individuals (see Honda, Shimizu, & Rutter, 2005; Kawamura, Takahashi, & Ishii, 2008). In addition, while the ASD diagnosis is not as prevalent as it is in the United States and Japan, the prevalence of ASD in children, aged 6 to 11 years in Taiwan, has increased from 1 of every 556 children in 2007 to 1 of every 363 children in 2016 (Ministry of Health and Welfare, 2017). Notwithstanding the increasing trends in prevalence, the etiology of ASD remains relatively unknown, so does the cure for the disorder. As a result, parents of and professionals who work with children with ASD are often apt to use treatments that have been widely advertised but considered controversial in regards with empirical evidence (Simpson, 2005). These controversial treatments refer to the invalidated and scientifically unsolid intervention strategies that show little or no effect (Worley, Fodstad, & Neal, 2014). Hence, there has been a concerted effort to examine the strengths of evidence for the existing ASD treatments in order to ensure the quality of such treatments. Given high educational expectations of Asian parents and educators towards children's performances, identifying effective intervention strategies has increasingly become a pressing issue across Asian countries.

Many times, single-case research methodology is utilized in targeting the behaviors of individuals with low prevalence disorders, such as ASD (Horner et al., 2005). Given the nature of flexibility and adaptability to the research designs, single-case research methodology is particularly useful when determining an effective intervention for targeted behaviors of individuals with ASD while controlling for threats to experimental validity (Kratochwill et al., 2010, 2014). In the autism and single-case research literature, to be

considered effective, treatments should be rigorously evaluated against the standards of quality for research experimentation and measurement (Zane, Davis, & Rosswurm, 2008). There is no clear consensus on what quality indicator should be used over the other; however, most researchers agree on several indicators that must be presented in a study for testing treatment efficacy and to be considered of high quality (Horner et al., 2005; Kratochwill, 2013; Kratochwill et al., 2010, 2014).

When evaluating treatment effect of single-case design research, the standards developed by What Works Clearinghouse (see Kratochwill et al., 2010, 2014) and Council for Exceptional Children (see Cook et al., 2014) are among the most frequently cited references. To assess soundness of research methodology and provide researchers and practitioners with guidelines for identifying and selecting evidence-based practices, both groups established conceptual frameworks of quality indicators for single-case research designs that include some common criteria. The criteria for designs that should be met to be considered of high quality include (a) systematic manipulation of the independent variable, (b) repeated measurements of the outcome variables by more than one assessor, (c) three attempts to demonstrate an intervention effect, and (d) a minimum of three data points collected in each condition (e.g., baseline, intervention; Cook et al., 2014; Kratochwill et al., 2010, 2014). In addition to these standards of research quality, the CEC standards address the importance of reporting measurement data on the independent variable, called implementation fidelity and/or treatment fidelity (Cook et al., 2014).

While a well-designed independent variable and a repeated measured dependent variable are essential to effective interventions, researchers and practitioners need to rely on the accuracy of the implementation of independent variables (i.e., treatment fidelity) and of the observation of dependent variables (i.e., reliability) to determine and select high quality interventions. As can be seen in the standards suggested by WWC and CEC, researchers converge on the importance of adherence to accurate and reliable measurements in behavioral research that involves individuals with ASD (Cone, 1982; Hops, Davis, & Longoria, 1995). Thus, this review aims to investigate the reliability and treatment fidelity of single-case research for individuals with ASD published in Asian countries.

Many researchers have argued that human observers are not bias-free in behavioral research, and therefore, observational methodologies can result in invalid data (Hops et al., 1995). To enhance credibility of one's findings in intervention research, reporting inter-rater reliability (IRR) scores on outcome measures is performed as the most common strategy to ensure the accuracy of observational data (Foster, Sclan, Welkowitz, Boksay, & Seeland, 1988; Kratochwill et al., 2010, 2014). Numerous indices have been developed and applied to assess IRR (Baer, 1977), and among those, percent agreement and *kappa* have been utilized prominently in behavioral research (Cohen, 1960). Given the computational simplicity and ease of interpretation (Baer, 1977; Hops et al., 1995), percent agreement is regularly used in single-case design research (Artman, Wolery, & Yoder, 2012). However, percent agreement indices often tend to inflate the degree of observer agreement due to underestimation of chance agreements (Berk, 1979). As an alternative index, the *kappa* coefficient has been suggested to improve the faults of percent agreement indices by taking into account chance agreements (Cohen, 1960; Parker, Vannest, & Davis, 2013). So far as can be observed in the ASD single-case research, it has been considered a common practice to record reliability data on outcome variables (Hartmann, 1977). However, recording measurements of and reliability data on treatment fidelity are not standard practices in behavioral experiments while considerable attention has been given to the importance of reporting those measures (Cook et al., 2014; McIntyre, Gresham, DiGennaro, & Reed, 2007).

Treatment fidelity is defined as the methodological strategies that monitor and ensure if a treatment condition is implemented and systematically manipulated as planned (Kazdin, 1986; Vermilyea, Barlow, & O'Brien, 1984). Given the fact that fidelity data can help researchers determine the factors associated with implementation success or failure of the intervention, collecting and reporting treatment fidelity scores at acceptable levels in intervention research are important (Dusenbury, Brannigan, Falco, & Hansen, 2003). Furthermore, lack of or no treatment fidelity data cannot ascertain if an independent variable was the sole factor responsible for study outcomes (Bellg et al., 2004), and this uncertainty can raise doubts about the efficacy of the intervention. Given the importance of collecting and reporting treatment

fidelity measures in behavioral research, a slight increasing trend in documentation and measurement of treatment fidelity data in the ASD-focused research has been detected over the last 30 years (Gresham, Gansle, Noell, & Cohen, 1993; Neely, H. Davis, J. Davis, & Rispoli, 2015); however, the occurrence rate for studies that meet the minimum quality standards (see Cook et al., 2014; Kratochwill et al., 2010, 2014) still remains low (Neely et al., 2015).

In the development and identification of evidence-based practices in ASD interventions, as more emphasis has been placed on reliability and treatment fidelity data in single-case experiments, there has been an effort to analyze measurements of and trends for those indices in the ASD research (e.g., Billingsley, White, & Munson, 1980; Gresham et al., 1993; Mudford, Taylor, & Martin, 2009; Neely et al., 2015). For example, Neely et al. (2015) reviewed trends in reporting reliability and treatment fidelity measures in ASD-focused single-case research across the years 1992, 2002, and 2012. Overall, 119 studies were evaluated based on the pre-set reliability and treatment fidelity criteria. As a result, a total of 118 studies (99%) reported IRR on outcome variables, and 58 studies (48%) reported treatment integrity data. Of the 58 studies, 20 studies (38%) collected IRR-integrity measures. The results of this review were consistent with the findings from previous reviews, indicating that relatively more recent studies tended to report both reliability and treatment fidelity measures in their studies (Neely et al., 2015). While the previous reviews provide a snapshot view of the degree to which the trends in reporting reliability and treatment fidelity data with an acceptable level increase in the ASD-focused single-case research, assessing only studies published in major English-language journals might not have captured worldwide trends.

Compared to the considerable amount of research conducted and published in English-language journals, examining the quality of evidence for ASD interventions is still at an early stage in Asian countries. For example, in the past five years, researchers in Taiwan have conducted several meta-analytic reviews in an effort to evaluate different intervention techniques applied to children with intellectual disabilities and ASD, such as social skill training, social story, and function-based intervention (e.g., Wu & Niew, 2012;

Huang & Niew, 2010; Chen, Tsai, & Lin, 2015). However, only one of those reviews that focused on intellectual disabilities assessed the methodological quality of the studies included in the analyses, which leads to uncertainty of the treatment efficacy of interventions for ASD in Taiwan. In addition, no similar review has yet been published in Japan even though Japan has been a longtime advocate for individuals with disabilities over the past century for special needs education. To date, although various types of ASD treatments have been empirically validated by many researchers and reported in multiple English-language journals, it is still not known if such findings can be equally supported in Asian countries when considering the different research and educational environments.

Therefore, we attempted to replicate and extend previous findings by evaluating single-case studies published in ASD-focused and non-English-language journals, including those published in Japanese and Taiwanese. In this review, Japanese and Taiwanese journals were selected to be evaluated since these two countries had published comparably high numbers of ASD-focused single-case studies among Asian countries. The purpose of this review is to investigate the current status and trends of the quality of the reliability and treatment fidelity measures reported in the ASD-focused single-case studies published in Japanese and Taiwanese special education journals.

Method

A systematic review was applied in this study, which comprehensively synthesized data focusing on inter-rater reliability and treatment fidelity in ASD-focused single-case research in Japan and Taiwan. To conduct this review, the following steps were taken: (a) literature search of the seven prominent journals in ASD and single-case research in Japan and Taiwan, (b) assessment of potential studies against pre-set inclusion and exclusion criteria, and (c) evaluation for measures of reliability and treatment fidelity in the studies that met the inclusion criteria.

Literature Search

While the examination of the quality of evidence for ASD interventions is still at an early stage in Asian countries, this review attempts to explore this

topic by analyzing reliability and treatment fidelity of ASD-focus single-case research in academically oriented, peer-reviewed special education journals. The authors applied the approach used by Gresham et al. (1993), Mudford et al. (2009) and Neely et al. (2015) to identify appropriate studies for this analysis. Specifically, seven most prominent academic-oriented and peer-reviewed special education journals in Japan and Taiwan were reviewed, including three Japanese journals, *Japanese Journal of Special Education*, *Japanese Journal of Behavior Analysis*, and *Japanese Journal of Behavior Therapy*, and four Taiwanese journals, *Journal of Special Education*, *Bulletin of Special Education*, *Bulletin of Special Education and Rehabilitation*, and *Bulletin of Eastern Taiwan Special Education*. All volumes and issues of the seven journals until 2015 of the publication year were reviewed. These journals were selected based on their academic orientation and reputation in the field of special education in Japan and Taiwan and/or their emphasis on the ASD research.

As a result, a total of 5,098 articles were identified from the seven journals: 2,824 from Japanese Journal of Special Education published from 1964 to 2015, 318 from Japanese Association for Behavior Analysis published from 1987 to 2015, 805 from Japanese Journal of Behavior Therapy published from 1976 to 2015, 290 from Journal of Special Education published from 1986 to 2015, 525 from Bulletin of Special Education published from 1985 to 2015, 197 from Bulletin of Special Education and Rehabilitation published from 1991 to 2015, and 139 from Bulletin of Eastern Taiwan Special Education published from 1998 to 2015.

Inclusion and Exclusion Criteria Evaluation

To be included in this review, studies identified in the searches were examined using a two-step process. First, the authors reviewed the title and abstract of each article to evaluate a research methodology and participant characteristics (i.e., diagnosis). From the first evaluation, studies that utilized a group experimental design, survey research, qualitative research methodology (e.g., case study) or editorial commentary or review were excluded from the further analysis. In addition, if there was no indication of participants with ASD either in the title or abstract, those articles were also

excluded from the second evaluation. Following the first evaluation, the authors looked into each of the remaining articles and examined if these articles met the following criteria: (a) included at least one participant who had either a primary or secondary diagnosis of ASD, (b) utilized single-case research methodology, and (c) presented data in a graph and collected the data on outcome behaviors of the participants with ASD. If studies indicated that participants showed autistic features but had no diagnosis of ASD, those studies were excluded from this review.

From the evaluation, a total of 194 (Japanese journals: n = 168, Taiwanese journals: n = 26) articles were identified to meet the inclusion criteria, and therefore, included in this review. Figure 1 shows the literature search leading to selection of the final articles.

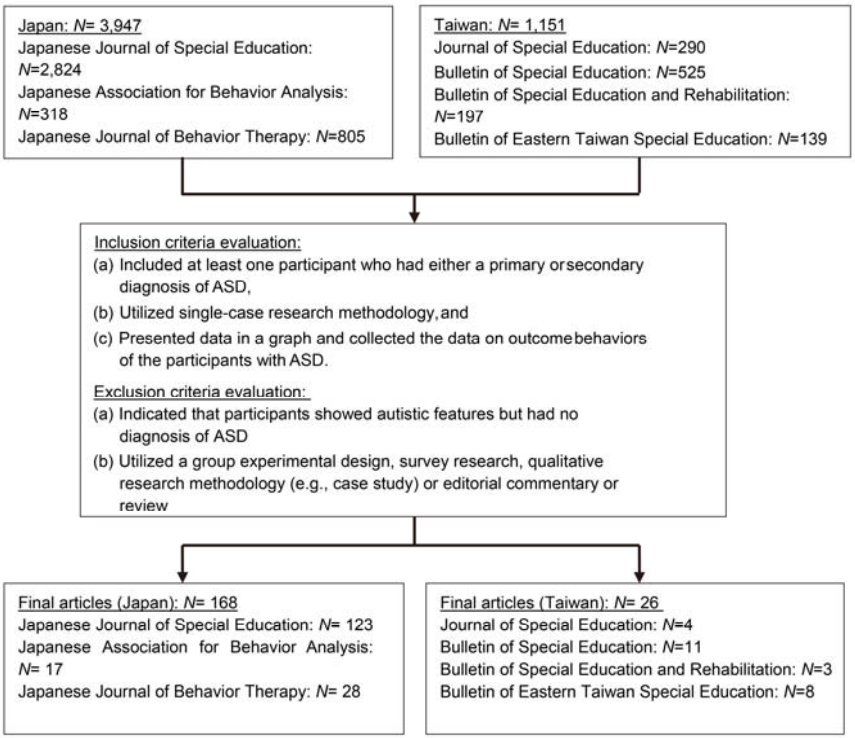


Figure 1. Literature Search

Study Coding: Descriptive Study Characteristics and Measures of Reliability and Treatment Fidelity

Each study was evaluated across the following categories: (a) inter-rater reliability (IRR) for outcome variables and (b) independent variable treatment fidelity and IRR for treatment fidelity. To evaluate measures of IRR and treatment fidelity of each study included in this review, we adapted the coding protocols developed by Kratochwill et al. (2010, 2014), Neely et al. (2015), and What Works Clearinghouse (see Kratochwill, 2013). A total of nine standards were used in this review, three for IRR for outcome variables, three for independent variable treatment fidelity, and three for the IRR for treatment fidelity. Table 1 provides the requirements within the coding protocol for determination of whether or not each study met each of the nine standards developed.

Inter-Rater Reliability of Study Coding

Reliability was calculated for the literature search and study coding. To calculate IRR scores, a percentage of agreement between two raters was used throughout this review. IRR scores were calculated by dividing agreements by agreements plus disagreements and multiplying by 100. To determine whether or not the articles met the initial inclusion criteria, a second and third independent rater reviewed 20% of each of initial group of articles (Wang & Parrila, 2008) published in the Japanese and Taiwanese journals, respectively. Initial IRR scores obtained were 97% (range, 95 ~ 98%) for the Japanese journals and 91% (range, 85 ~ 92%) for the Taiwanese journals.

Results

In this review, a total of 194 articles met the inclusion criteria and were evaluated for the reliability data on dependent variables as well as the treatment fidelity measures on independent variables. Across 168 articles published in the Japanese journals, a total of 283 subjects whose ages ranged from 1 to 62 years old participated in the experiments. As for 26 articles published in the Taiwan journals, a total of 41 subjects whose ages ranged from 2 to 15 years old participated in the experiments. In both groups of the journals evaluated, various types of intervention strategies were utilized, such as video modeling, function-based intervention, augmentative and alternative communication training, and task analysis. In addition, a wide range of skills

were targeted for change, including communication, behavior, academic, functional living, and leisure skills. Data were grouped within 10-year spans and are presented in Tables 1 and 2.

Publication Status and Trends

Consequently, a total of 168 articles published in the Japanese journals were identified to have met the inclusion criteria, and therefore, evaluated in this review for the reliability and treatment fidelity measures. Of the 168 articles assessed in the review, 5 articles were published between 1976 and 1985, 30 articles between 1986 and 1995 (a 500% increase), 64 articles between 1996 and 2005 (a 113% increase), and 69 articles between 2006 and 2015 (a 7% increase). Only slight increases were observed in recent publications in terms of the number of publications of ASD-focused single-case research in the Japanese journals. Overall, there appeared increasing trends in the number of publications over 40 years in the Japanese journals.

On the contrary, such trends were not observed in the Taiwanese journals. A total of 26 articles met the inclusion and exclusion criteria and were evaluated in the review. Of the 26 articles, 15 articles were published between 1996 and 2005 and 11 articles between 2006 and 2015. Overall, the number of publications of ASD-focused single-case research over 20 years in the Taiwanese journals was slightly decreasing.

Inter-Rater Reliability for Dependent Variables

A total of 91 articles (54.2%) published in the Japanese journals were found to have reported IRR data for dependent variables. As a result of the evaluation of the reliability data for dependent variables, overall increasing trends in reporting IRR data with acceptable levels were observed over time in the Japanese language ASD-focused single-case research, except for the Standard 1.2 “IRR was collected in each condition and on at least 20% of the data points in each condition” (see Fig. 2). For Standard 1.1 and 1.3, the 1976 ~ 1985 data were the lowest and the 2006 ~ 2016 were the highest among the four decades, regarding the number and percentage of included studies reporting IRR data for dependent variables as well as reporting IRR data that met the minimum quality thresholds (above an 80% criterion if utilizing percent agreement or 0.6 if utilizing *kappa*). Compared to 1976 ~ 1985, the percentage of the studies reporting IRR for dependent variables were doubled in 1986 ~ 1995 and 1996 ~ 2005, and tripled in 2006 ~ 2015. As to the quality of IRR on dependent variables, more than half of the studies in Japan

Table 1
Coding Protocol and the Number of Studies Evaluated for Inter-rater Reliability and Treatment Fidelity: Japanese Journals

	1. Inter-rater Reliability (IRR)			2. Treatment Fidelity (TF)					
Years	1.1 IRR was measured for outcome variables.	1.2. IRR was collected in each condition and on at least 20% of the data points in each condition (e.g., baseline, intervention).	1.3. Resulting IRR scores were above 80% if calculated by percentage agreement or at least 0.6 if measured by Cohen's kappa for each outcome variable.	2.1 TF was measured for independent variable.	2.2. TF for independent variable was collected in each intervention condition and on at least 20% of the data points in each intervention condition.	2.3. Resulting TF scores were above 80% if calculated by percentage agreement or at least 0.6 if measured by Cohen's kappa.	2.4. IRR was collected for the independent variable TF.	2.5. IRR was collected for the independent variable TF on at least 20% of data points in each intervention condition.	2.6. Resulting reliability scores were above 80% if calculated by percentage agreement or at least 0.6 if measured by Cohen's kappa for the independent variable.
1976- *Total*	5	5	5	5	5	5	5	5	5
1985 *YES*[a]	1 (20%)	1 (20%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
NO[b]	4 (80%)	4 (80%)	5 (100%)	5 (100%)	5 (100%)	5 (100%)	5 (100%)	5 (100%)	5 (100%)
1986- *Total*	30	30	30	30	30	30	30	30	30
1995 *YES*	14 (46.7%)	4 (13.3%)	9 (30%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
NO	16 (53.3%)	26 (86.7%)	21 (70%)	30 (100%)	30 (100%)	30 (100%)	30 (100%)	30 (100%)	30 (100%)
1996- *Total*	64	64	64	64	64	64	64	64	64
2005 *YES*	31 (48.4%)	10 (15.6%)	31 (48.4%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
NO	33 (51.6%)	54 (84.4%)	33 (51.6%)	64 (100%)	64 (100%)	64 (100%)	64 (100%)	64 (100%)	64 (100%)
2006- *Total*	69	69	69	69	69	69	69	69	69
2015 *YES*	45 (65.2%)	10 (14.5%)	41 (59.4%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
NO	24 (34.8%)	59 (85.5%)	28 (40.6%)	69 (100%)	69 (100%)	69 (100%)	69 (100%)	69 (100%)	69 (100%)
Total	168	168	168	168	168	168	168	168	168
YES	91 (54.2%)	25 (14.9%)	81 (48.2%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
NO	77 (45.8%)	143 (85.1%)	87 (51.8%)	168 (100%)	168 (100%)	168 (100%)	168 (100%)	168 (100%)	168 (100%)

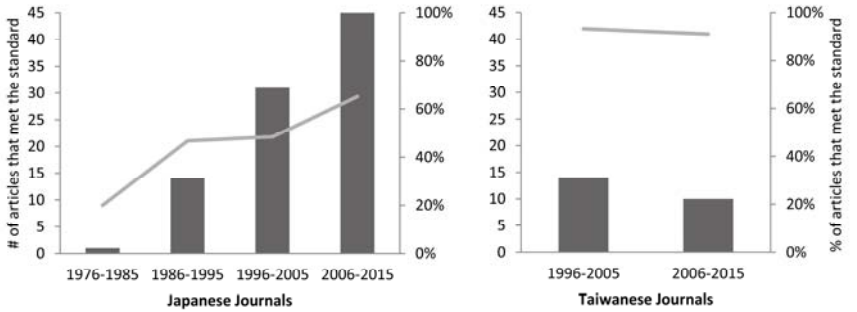
Note. [a]YES=the number of articles that met the corresponding criterion; [b]NO=the number of articles that did not meet the corresponding criterion.

Table 2
Coding Protocol and the Number of Studies Evaluated for Inter-rater Reliability and Treatment Fidelity: Taiwanese Journals

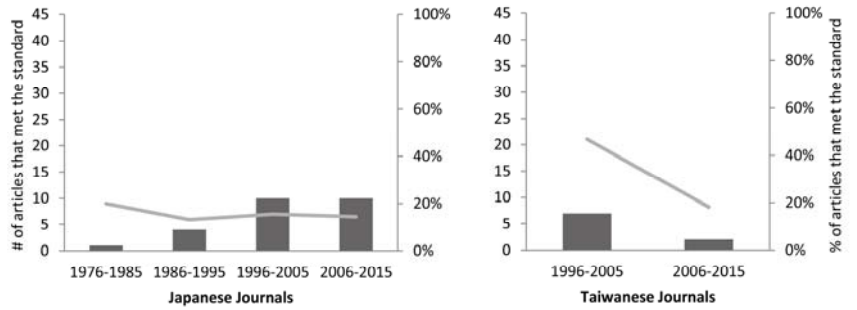
		1. Inter-rater Reliability (IRR)			2. Treatment Fidelity (TF)					
Years		1.1 IRR was measured for outcome variables.	1.2. IRR was collected in each condition and on at least 20% of the data points in each condition (e.g., baseline, intervention).	1.3. Resulting IRR scores were above 80% if calculated by percentage agreement or at least 0.6 if measured by Cohen's kappa for each outcome variable.	2.1 TF was measured for independent variable.	2.2. TF for independent variable was collected in each intervention condition and on at least 20% of the data points in each intervention condition.	2.3. Resulting TF scores were above 80% if calculated by percentage agreement or at least 0.6 if measured by Cohen's kappa.	2.4. IRR was collected for the independent variable TF.	2.5. IRR was collected for the independent variable TF on at least 20% of data points in each intervention condition.	2.6. Resulting reliability scores were above 80% if calculated by percentage agreement or at least 0.6 if measured by Cohen's kappa for the independent variable.
1996-	*Total*	15	15	15	15	15	15	15	15	15
2005	*YES*[a]	14 (93.3%)	7 (46.7%)	12 (80%)	6 (40%)	2 (13.3%)	3 (20%)	3 (20%)	1 (6.7%)	3 (20%)
	NO[b]	1 (6.7%)	8 (53.3%)	3 (20%)	9 (60%)	13 (86.7%)	12 (80%)	12 (80%)	14 93.3%)	12 (80%)
2006-	*Total*	11	11	11	11	11	11	11	11	11
2015	*YES*	10 (91.1%)	2 (18.2%)	7 (63.6%)	4 (36.4%)	1 (9.1%)	3 (27.3%)	2 (18.2%)	0 (0%)	1 (9.1%)
	NO	1 (9,1%)	9 (81.8%)	4 (36.4%)	7 (63.6%)	10 (90.9%	8 (72.7%)	9 (81.8%)	11 (100%)	10 (90.9%)
Total		26	26	26	26	26	26	26	26	26
	YES	24 (92.3%)	9 (34.6%)	19 (73.1%)	10 (38.5%)	3 (11.5%)	6 (23.1%)	5 (19.2%)	1 (3.8%)	4 (15.4%)
	NO	2 (7.7%)	17 (65.4%)	7 (26.9%)	16 (61.5%)	23 (88.5%)	20 (76.9%)	21 (80.8%)	25 (96.2%)	22 (84.6%)

Note. [a]YES=the number of articles that met the corresponding criterion; [b]NO=the number of articles that did not meet the corresponding criterion.

1.1. IRR was collected for dependent variable



1.2. IRR was collected in each condition and on at least 20% of the data points in each condition (e.g., baseline, intervention)



1.3. Resulting IRR scores were above 80% if calculated by percentage agreement or at least 0.6 if measured by Cohen's kappa

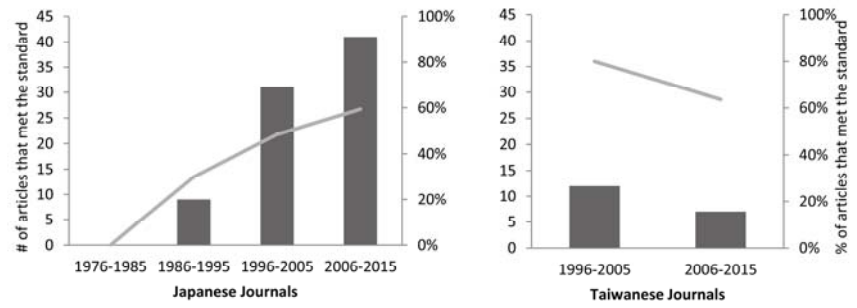


Figure 2. Results: Evaluation of Inter-rater Reliability Standards

Note. The dark bars indicate the number of articles that met the standard; the lines indicate the percentage of articles that met the standard.

published between 2006 and 2015 (59.4%, $n = 41$) reported IRR on dependent variables that met the quality standard. However, such an increasing trend was not observed for the Standard 1.2. A percentage of equal to or under 20% of the studies that collected IRR data for a minimum of 20% of the data points in each condition was consistently found across the four decades.

On the contrary, there existed no increasing trends in the number and percentage of articles that met the quality standards in the Taiwanese language ASD-focused single-case research over time (see Fig. 2). For the Standard 1.1, an overall high percentage of studies collecting IRR for dependent variable was stably found in 1996 ~ 2005 (93.3%, $n = 14$) and 2006 ~ 2015 (91.1%, $n = 10$). For the Standard 1.2, a decrease in percentage was found over time from 46.7% in 1996 ~ 2005 ($n = 7$) to 18.2% in 2006 ~ 2015 ($n = 2$) of studies collecting IRR for at least 20% of the data points in each condition. Similarly, for the Standard 1.3, there was also a decreasing trend from 80% in 1996 ~ 2005 ($n = 12$) to 63.6% in 2006 ~ 2015 ($n = 7$) of studies meeting the minimum quality thresholds of IRR (the reliability coefficient was above an 80% criterion if utilizing percent agreement or 0.6 if utilizing *kappa*). Although the overall quality of ASD-focused single-case research in Taiwanese journals was higher than that in Japanese journals, caution is needed in the interpretation since the number of Taiwanese studies that met the inclusion criteria was small ($N = 26$).

Treatment Fidelity for Independent Variable(s) Reliability Data on Treatment Fidelity

None of the articles published in the Japanese journals included in this review was found to have reported treatment fidelity data for independent variable(s) while 10 articles (38.5%) published in the Taiwanese journals reported treatment fidelity data for independent variable(s). Only a small number of articles that reported treatment fidelity data with acceptable quality degrees were observed (see Fig. 3). Of the 15 articles published in between 1996 and 2005, 6 articles (40%) collected treatment fidelity for independent variable(s). Among these articles, only 2 articles (13.3%) collected treatment fidelity for independent variable(s) for each least 20% of the data points in each condition, and 3 articles (20%) met the minimum quality thresholds of

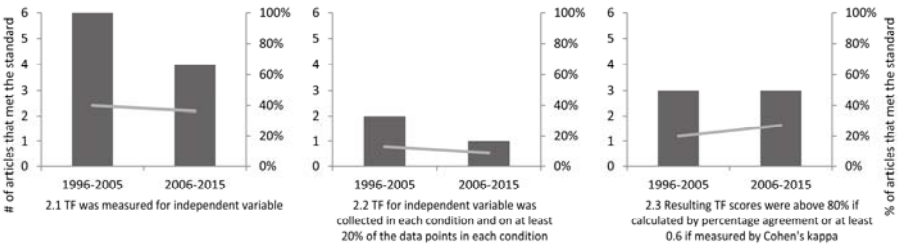


Figure 3. Results: Evaluation of Treatment Fidelity of Articles Published in the Taiwanese Journals

Note. The dark bars indicate the number of articles that met the standard; the lines indicate the percentage of articles that met the standard.

treatment fidelity data (above an 80% criterion if utilizing percent agreement or 0.6 if utilizing *kappa*). Of the 11 articles published between 2006 and 2015, 4 articles (36.4%) collected treatment fidelity for independent variable(s). Among these, 1 article (9.1%) collected treatment fidelity data for at least 20% of the data points in each intervention condition, and 3 articles (27.3%) met the minimum quality thresholds.

Of the 15 articles published between 1996 and 2005, 3 articles (20%) collected IRR on treatment fidelity data, and all 3 articles (20%) met the minimum quality thresholds of the measures. Among these, only 1 article (6.7%) was found to have collected IRR on treatment fidelity data for at least 20% of the data points in each intervention condition. Among the 11 articles published between 2006 and 2015, 2 articles (18.2%) collected IRR on treatment fidelity data, and 1 (9.1%) of these met the minimum quality thresholds. None of the articles was identified to have collected IRR on treatment fidelity for at least 20% of the data points in each session. Figure 4 presents the results on the collection of IRR on treatment fidelity data. Overall, the relatively low percentage of reporting treatment fidelity and IRR on treatment fidelity may influence the interpretation of the intervention effects on the dependent variable.

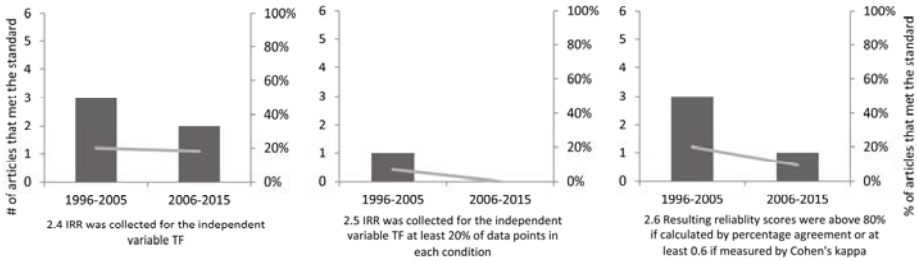


Figure 4. Results: Evaluation of Reliability Data on Treatment Fidelity of Articles Published in the Taiwanese Journals

Note. The dark bars indicate the number of articles that met the standard; the lines indicate the percentage of articles that met the standard.

Discussion

In summary, more articles were published in the Japanese journals that met the standards for IRR than in the Taiwanese journals. That follows from the fact that more than three times the number of Japanese articles met the inclusion criteria than Taiwanese articles. The discrepancy is not surprising given the difference in the number of potential articles identified for each country in the initial literature search and given the fact that Japanese journals began publishing single-case studies in autism approximately 20 years prior to Taiwanese journals. Furthermore, the field of Applied Behavior Analysis, which uses single-case designs as its major research methodology, has been developed in Japan since 1980s, whereas ABA was formally introduced to Taiwan in 2007. Therefore, the single-case research method might be used by more researchers in Japan (about 32 single case design articles per year) than those in Taiwan (about 11 articles per year).

There was an increasing trend of numbers and percentages of Japanese articles that met most of the IRR standards and a level trend for the reporting of IRR for meeting the 80% agreement or .6 Cohen's kappa. While the increasing trend is not found for Taiwanese articles, more than 90% of the articles published in Taiwan had reported IRR for outcome variables in the past twenty years. Overall, it does appear that Asian authors are increasingly attending to quality with regard to collecting and reporting IRR data for outcome variables in their single-case autism-related articles. Nevertheless,

among the three IRR standards, a relative low percentage of articles in Japan and Taiwan reported of IRR for at least 20% of data points within all conditions. The low percentage might result from the different interpretation of the "collecting IRR for at least 20% data points" standard by researchers in Asian countries. For instance, while coding Taiwanese articles, it was found that the researchers tended to report IRR collected from 20% of data points of the entire study, and thus most of the articles did not meet the standard. This finding indicates that most researchers in Asian countries need to be more explicit about whether IRR for 20% of data points is collected within each condition.

Interestingly, opposite trends resulted from investigations of data collected on treatment fidelity. That is, none of the Japanese articles reported treatment fidelity while about 40% of the Taiwanese articles did report treatment fidelity data and a slightly increasing trend of numbers and percentages of Taiwanese articles met the 80% or 0.6 treatment fidelity threshold. However, very few Taiwanese researchers reported the percentage of sessions where they collected treatment fidelity data and thus did not meet the "collecting TF for 20% of data point within each intervention condition" standard. The overall low percentage of Taiwanese articles that collected IRR for independent variable TF shows that researchers in Asian countries may not be familiar with the procedure of collecting TF data.

In comparison to English-language journals, Japanese journals appear to report IRR for outcome variables and meet standards for the amount of IRR data collected and reported treatment fidelity at lower rates, although this should be interpreted with caution given the discrepancy in the numbers of articles published by both sets of journals. Given the low numbers of Taiwanese journals reporting either IRR or treatment integrity, it is not possible to make strict comparisons. It is promising, however, that Japanese articles have paid increasing attention to quality of IRR data reporting and that Taiwanese articles have begun reporting treatment fidelity data.

Both IRR and treatment fidelity data collection are issues of questioning whether or not the authors did what they reported to do and are reporting accurately. That is, IRR measures whether or not independent observers agree whether or not a particular outcome behavior occurred (Hops et al., 1995).

Lack of high rates of agreement from independent observers could indicate that the outcome variable was not well-defined or the observers not well-trained; in either case, unless independent observers agree on outcome behavior occurrences, readers cannot be certain that the effects of the study are accurate. Treatment fidelity data report accuracy of implementation of the intervention according to a pre-determined protocol (Kazdin, 1986). Thus, if treatment fidelity data are not collected, the reader cannot be certain that the intervention reported was accurately implemented, calling the results into question also. For example, interventionists may have delivered additional reinforcement, resulting in positive results from the reinforcement rather than the stated intervention. It is encouraging that both English- language and Asian journals are more frequently reporting results of both of these measures for single-case experiments on interventions for individuals with autism.

This research does have some limitations and implications for future research. First, this review only explores the topic of inter-rater reliability and treatment fidelity in academically oriented journals in Japan and Taiwan. Such a method of literature search may restrict the interpretation of the results since not all of the Japanese and Taiwanese ASD-focused single-case studies were published in these selected journals. Further reviews should consider sampling from databases and practically oriented journals to examine whether the identified trends are unique to the reviewed journals or prevalent throughout ASD-focused single-case research in both countries. Second, given the discrepancy in the numbers of articles published in these Japanese and Taiwanese journals, interpretation with caution is needed since meaningful comparisons may not be made with such a small number of articles in Taiwanese journals. Third, the criteria used in the analysis for inter-rater reliability only involve three requirements (including collecting IRR data, at least 20% of data points in each condition, and above the 80% or 0.6 threshold). Other requirements, such as the number of repeated measurements and the appropriateness of the IRR formulas used for the dependent variables in each study, should be considered in further reviews. Fourth, the literature searched for this article included papers published in only two languages. Future research on the quality of single-case design could include articles published in a wider range of language, albeit there is difficulty acquiring

literature translated across several languages. Such a report would allow for worldwide comparisons in quality of research design. Finally, research on interventions for individuals with autism is increasing in quality of research design, including reporting of IRR and treatment fidelity; however, it is apparent that there is room for improvement, both in Asian-language and in English-language journals. Future research could provide guidance to researchers regarding key components of designing single-case experiments on interventions for people with autism. Given this era of identification of evidence-based practices, it is imperative that researchers produce high-quality, reliable research as low-quality research may not be appropriate for inclusion in systematic literature reviews or meta-analyses.

Conclusion

This systematic review aims to investigate the current status and trends of quality of ASD-focused single-case research in reporting IRR for dependent variables and treatment fidelity for independent variables and the IRR for treatment fidelity in Japanese and Taiwanese academic-oriented special education journals. The results suggest an increasing trend in the collection and quality of IRR data for dependent variables in Japan. Although such a trend was not observed in Taiwanese journals, the overall quality in IRR for dependent variables in these journals was relatively high. Furthermore, the treatment fidelity data and the quality of treatment fidelity data were only found in Taiwan, although remained at a low level. In sum, researchers in Japan and Taiwan are increasingly attending to quality with regard to collecting and reporting IRR data for outcome variables in their single-case autism-related articles. Yet, the evaluation of treatment fidelity and its IRR is still at the initial stage. To improve the overall quality of ASD intervention research, efforts should be made to report both IRR and treatment fidelity data with acceptable quality degrees in Asian-language journals.

References

Note. A list of the articles in this review is available upon request.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: Author.

Artman, K., Wolery, M., & Yoder, P. (2012). Embracing our visual inspection and analysis tradition: Graphing interobserver agreement data. *Remedial and Special Education, 33*(2), 71-77. doi: 10.1177/0741932510381653

Baer, D. M. (1977). Perhaps it would be better not to know everything. *Journal of Applied Behavior Analysis, 10,* 167-172. doi:10.1901/jaba.1977.10-167

Bellg, A. J., Borrelli, B., Resnick, B., Hecht, J., Minicucci, D. S., Ory, M., ... Czajkowski, S. (2004). Enhancing treatment fidelity in health behavior change studies: Best practices and recommendations from the NIH Behavior Change Consortium. *Health Psychology, 23*(5), 443-451. doi:10.1037/0278-6133.23.5.443

Berk, R. A. (1979). Generalizability of behavioral observations: A clarification of interobserver agreement and interobserver reliability. *American Journal of Mental Disability, 83*(5), 460-472.

Billingsley, F. F., White, O. R., & Munson, R. (1980). Procedural reliability: A rationale and an example. *Behavioral Assessment, 2,* 229-241.

Centers for Disease Control and Prevention. (2014). *Prevalence of autism spectrum disorders among children aged 8 years—Autism and developmental disabilities monitoring network, 11 sites, United States, 2010.* Retrieved from Centers for Disease Control and Prevention website: https://www.cdc.gov/mmwr/pdf/ss/ss6302.pdf

Chen, P., Tsai, S., & Lin, P. (2015). The effects of function-based behavioral interventions for students with disabilities: A meta-analysis. *Bulletin of Special Education, 40,* 1-30 (in Chinese).

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37-46. doi:10.1177/001316446002000104

Cone, J. D. (1982). Validity of direct observation procedures. *New Directions for Methodology of Social and Behavioral Science, 12,* 67-69.

Cook, B., Buysse, V., Klingner, J., Landrum, T., McWilliam, R., Tankersley, M., & Test, D. (2014). Council for Exceptional Children: Standards for evidence-based practices in special education. *Teaching Exceptional Children, 46,* 206-212. doi:10.1177/0040059914531389

Dusenbury, L., Brannigan, R., Falco, M., & Hansen, W. B. (2003). A review of research on fidelity of implementation: Implications for drug abuse prevention in school settings. *Health Education Research, 18,* 237-256. doi:10.1093/her/18.2.237

Elsabbagh, M., Divan, G., Koh, Y. J., Kim, Y. S., Kauchali, S., Marcín, C., ... Yasamy, M. T. (2012). Global prevalence of autism and other pervasive developmental disorders. *Autism Research, 5,* 160-179. doi:10.1002/aur.239

Foster, J. R., Sclan, S., Welkowitz, J., Boksay, I., & Seeland, I. (1988). Psychiatric assessment in medical long-term care facilities: Reliability of commonly used rating scales. *International Journal of Geriatric Psychiatry, 3,* 229-233. doi:10.1002/gps.930030310

Gresham, F. M., Gansle, K. A., Noell, G. H., & Cohen, S. (1993). Treatment integrity of school-based behavioral intervention studies: 1980-1990. *School Psychology Review, 22,* 254-272.

Hartmann, D. P. (1977). Considerations in the choice of reliability estimates. *Journal of Applied Behavior Analysis, 10,* 103-116. doi: 10.1901/jaba.1977.10-103

Honda, H., Shimizu, Y., & Rutter, M. (2005). No effect of MMR withdrawal on the incidence of autism: A total population study. *Journal of Child Psychology and Psychiatry, 46,* 572-579. doi:10.1111/j.1469-7610.2005.01425.x

Hops, H., Davis, B., & Longoria, N. (1995). Methodological issues in direct observation: Illustrations with the living in familial environments (LIFE) coding system. *Journal of Clinical Child Psychology, 24,* 193-203. doi:10.1207/s15374424jccp2402_7

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children, 71,* 165-179. doi:10.1177/001440290507100203

Huang, M., & Niew, W. (2010). A research synthesis of social story interventions for students with autism spectrum disorders. *Bulletin of Special Education and Rehabilitation, 22,* 1-23 (in Chinese).

Kawamura, Y., Takahashi, O., & Ishii, T. (2008). Reevaluating the incidence of pervasive developmental disorders: Impact of elevated rates of detection through implementation of an integrated system of screening in Toyota, Japan. *Psychiatry and Clinical Neurosciences, 62,* 152-159. doi:10.1111/j.1440-1819.2008.01748.x

Kazdin, A. E. (1986). The evaluation of psychotherapy: Research designs and methodology. In S. L. Garfield & A. E. Bergin (Eds.), *Handbook of psychotherapy and behavior change* (pp. 23-68). New York, NY: Wiley.

Kratochwill, T. R. (Ed.). (2013). *Single subject research: Strategies for evaluating change*. New York, NY: Academic Press.

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M, & Shadish, W. R. (2010). Single case designs technical documentation. *What works clearinghouse: Procedures and standards handbook (version 2.0).* Retrieved from What Works Clearinghouse website: http://ies.ed.gov/ncee/wwc/pdf/wwc_procedures_v2_standards_handbook.pdf

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M, & Shadish, W. R. (2014). Single case designs technical documentation. *What works clearinghouse: Procedures and standards handbook (version 3.0).* Retrieved from What Works Clearinghouse website: http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf

McIntyre, L. L., Gresham, F. M., DiGennaro, F. D., & Reed, D. D. (2007). Treatment integrity of school-based interventions with children in the journal of applied behavior analysis 1991-2005. *Journal of Applied Behavior Analysis*, *40*, 659-672. doi:10.1901/jaba.2007.659672

Ministry of Health and Welfare. (2017). *Population of individuals with disabilities categorized by city and age.* Retrieved from http://dep.mohw.gov.tw/DOS/cp-1745-3328-113.html (in Chinese).

Mudford, O. C., Taylor, S. A., & Martin, N. T. (2009). Continuous recording and interobserver agreement algorithms reported in the Journal of Applied Behavior Analysis (1995-2005). *Journal of Applied Behavior Analysis*, *42*, 165-169. doi:10.1901/jaba.2009.42-165

Neely, L., Davis, H., Davis, J., & Rispoli, M. (2015). Review of reliability and treatment integrity trends in autism-focused research. *Research in Autism Spectrum Disorders*, *9*, 1-12. doi:10.1016/j.rasd.2014.09.011

Parker, R. I., Vannest, K. J., & Davis, J. L. (2013). Reliability of multi-category rating scales. *Journal of School Psychology*, *51*, 217-229. doi:10.1016/j.jsp.2012.12.003

Simpson, R. L. (2005). Evidence-based practices and students with autism spectrum disorders. *Focus on Autism and Other Developmental Disabilities*, *20*, 140-149. doi:10.1177/10883576050200030201

Vermilyea, B. B., Barlow, D. H., & O'Brien, G. T. (1984). The importance of assessing treatment integrity: An example in the anxiety disorders. *Journal of Psychopathology and Behavioral Assessment*, *6*, 1-11. doi:10.1007/bf01321456

Wang, S. Y., & Parrila, R. (2008). Quality indicators for single-case research on social skill interventions for children with autistic spectrum disorder. *Developmental Disabilities Bulletin*, *36*(1), 81-105.

Worley, J. A., Fodstad, J. C., & Neal, D. (2014). Controversial treatments for autism spectrum disorders. In J. Tarbox, D. R. Dixon, P. Sturmey, & J. L. Matson (Eds.), *Handbook of early intervention for autism spectrum disorders. Research, policy and practice* (pp. 617-646). New York, NY: Springer.

Wu, C., & Niew, W. (2012). A meta-analysis of social skills intervention for individuals with high-functioning autism and asperger syndrome. *Bulletin of Special Education, 37,* 29-57 (in Chinese).

Zane, T., Davis, C., & Rosswurm, M. (2008). The cost of fad treatments in autism. *Journal of Early and Intensive Behavior Intervention, 5*(2), 44-51. doi:10.1037/h0100418